CABIOS

Vol.7. no.1. 1991
Pages 57 - 60

A simulation program to display specific digestion products of predicted RNA foldings

Arieh Zaritsky and Edan Forester

Abstract

A parameterizable program in Pascal was developed for VAX/VMS computers to simulate the autoradiograms of gel-separated RNA fragments generated by partial cleavage of a folded RNA molecule using five specific RNases. Each screen displays the results of cleavage by either one enzyme or all five, with the RNA molecule labeled at either of its ends (5' or 3'); each run is performed with three different lengths and against a ladder containing alkaline hydrolysis products of the same RNA molecule as size markers. The program should be useful for comparing actual results with predicted functional foldings of RNA molecules.

Introduction

Secondary and tertiary structures of ribonucleic acid molecules are involved in various biological activities (e.g. Zaritsky *et al.*, 1988; Dahlberg and Abelson, 1989, 1990; Doudna *et al.*, 1989). Procedures to predict such structures from an RNA primary base sequence have been extensively described (e.g. Zuker and Stiegler, 1981; Williams and Tinoco, 1986; Martinez, 1988; Zuker, 1989). The 'correct' structure is chosen from alternative structures predicted by a computer program according to energetic considerations (Freier *et al.*, 1986) and experimental tests (e.g. Pieler *et al.*, 1986). Possible interactions between unpaired bases of a folded RNA, and between the RNA and other regulatory molecules in a living cell, complicate the achievement of meaningful conclusions (Garrett *et al.*, 1981; Goringer and Wagner, 1988). Strong evidence supporting the existence of a particular structure is usually obtained by analyzing products generated by ribonucleases with known structure-specific activities (Knapp, 1989). The molecule under scrutiny is labeled at either its 5′ or 3′ terminal nucleotide with ^{32}P, and then submitted to partial digestion by each of a battery of specific RNases, under conditions that are assumed not to interfere with the native RNA conformation. Lengths of the labeled fragments thus generated are determined by gel electrophoresis and autoradiography, with a ladder containing all alkaline hydrolysis products of the RNA as size markers. These lengths identify the positions of enzymatic cleavage, and hence the local secondary structure (Gerhart *et al.*, 1986; Pieler *et al.*, 1986).

Department of Biology, Ben Gurion University of the Negev, PO Box 653, Beer Sheva 84105, Israel

System and methods

We have developed a parameterizable simulation program (GEL), implemented in the Pascal language for the DEC (Digital Equipment Corporation) VAX/VMS computer. GEL aids in the analysis of the partial RNase digestion data by displaying ideal autoradiograms, analogous to those obtained for DNA restriction map analyses (e.g. Gross, 1986; Zehtner and Lehrach, 1986). Display is achieved using the VAX GKS (Graphic Kernel Standard) graphics software package, with devices that are supported by the package itself. GEL was tested

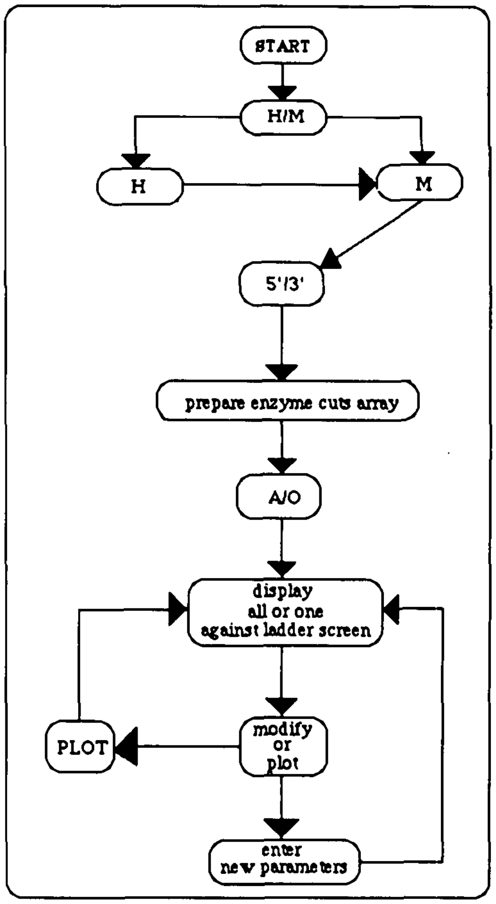

Fig. 1. Flow diagram for operating GEL. Enzyme cutting array is prepared on either 5′- or 3′-end-labeled RNA molecule in a marked base pair sequence (*M*) of an ASCII input file (Figure 2). If a hairpin structure (*H*) format is given (Figure 2), it is converted automatically to an equivalent *M* format. Display can be obtained with GKS software for either all enzymes (*A*) or one enzyme (*O*) cuts against ladder. If the latter, one should call for the selected enzyme.

on a series of benchmarks, and results were displayed successfully on DEC's VT100, VT240, Tektronix 4010 and Tektronix 4014 terminals and plotted on an HP7550 plotter. GEL was designed to be user-friendly and menu-driven; its operation algorithm is shown in Figure 1.

Algorithm

Upon invocation of GEL, the user is prompted to choose between two predefined formats of an ASCII input file, *H* for hairpin structure or *M* for marked (paired) base sequence (Figure 2), and then to call for the file. If *H* is chosen, the file is transformed into the *M* format. The further analysis of the sequence and prediction of labeled fragments depend only on whether a given base is paired (*M* format). Thus, any degree of complexity of looping can be handled in this format.

The transformation is written in a separate module and executed through an 8-connectivity search scheme. All elements in an 8-connectivity structure can be reached, one from the other, by a sequence of any of the eight one-element moves: up, down, left, right and each of the diagonals. The search is performed in a clockwise direction, starting in each step from a direction which depends on the previous path. For example, starting from the first base (5′) of Figure 2, it would be up, up-right diagonal, right, etc. For a right-side hairpin loop, the search would be right, right-down diagonal, down, etc., and so on.

The user is then asked to choose between 5′ and 3′ formats, depending on which end of the RNA molecule has been labeled. An intermediate process of 'enzyme cutting' simulation is performed, in which a list of all possible labeled fragment lengths *n* is prepared, based on the distances between the labeled end and possible cleavage points. Its results are recorded in a log file for reference, and in a local array.

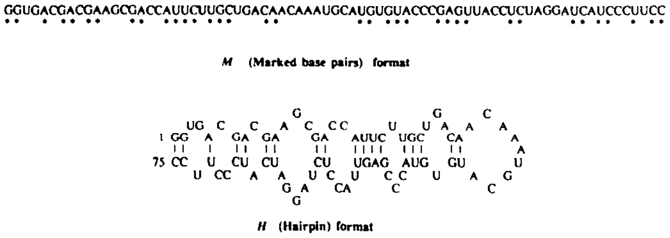
GGUGACGACGAAGCGACCAUUCUUGCUGACAACAAAUGCAUGUGUACCCGAGUUACCUCUAGGAUCAUCCCUUCC
M (Marked base pairs) format
1
75
H (Hairpin) format

Fig. 2. The predicted structure of the 5′-end 75 bases of murine J-chain mRNA (Zaritsky *et al.*, 1988) in its hairpin (*H*) format and transformed to (5′ to 3′ from the left) marked base pairs sequence (*M*) format.

The distance of a fragment with *n* nucleotides (D_n) from the origin (analogous to the slot where a sample is loaded) is generated by the equation $D_n = b^{-n}/P$, where *b* is chosen to suit the experimental system (see below) and can be modified at will, and *P* is a normalization factor that includes all parameters of the experimental procedure (e.g. composition and dimensions of the gel, electrical conductivity of the solvent, run time and voltage or current; Lehrach *et al.*, 1977) as well as of the simulation system (length of a screen page).

Implementation

As an example, we consider the 5′-terminal 75 bases of murine J-chain mRNA (Cann *et al.*, 1982; Matsuuchi *et al.*, 1986; Zaritsky *et al.*, 1988), which has a high potential for forming secondary structures with presumed biological activities (Ben-Alon, 1989). The compact hairpin structure of Figure 2 (Zaritsky *et al.*, 1988), with the lowest predicted free energy of formation ($\Delta G = -23$ kcal/mol), serves to demonstrate the applicability of the program. Five RNases with different cleavage specificities (Table I) are used for the analysis. The program simulates the pattern of migration on an electrophoretic gel of a set of denatured RNA fragments obtained by a partial cleavage of the native structure by each of the enzymes (Rushizky *et al.*, 1970; Donis-Keller *et al.*, 1977; Donis-Keller, 1980; Boguski *et al.*, 1980; Lowman and Draper, 1986).

To initiate graphics, the user is asked to select either *A* for *all* enzymes against ladder or *O* for *one* enzyme against ladder. If *O* is chosen, the selected enzyme is called for. Upon graphics invocation, GKS software is opened, and the appropriate workstation is initialized including a terminal and viewport definitions.

If option *A* is selected, a screen built of three frames, each divided to six lanes (one for each enzyme and one for the ladder) is displayed (Figure 3). The figure simulates the patterns generated by cutting the hairpin structure of Figure 2, labeled at its 5′-end, with each of five enzymes after run to $n_1 = 1$ in frame I, $n_2 = 25$ in frame II and $n_3 = 50$ in frame III. These default values and the exponent base $b = 1.1$ were chosen because they yield reasonable results with a molecule of 75 bases and with the available degree of resolution of the display system.

Table I. Specific ribonucleases, used for determination of RNA secondary structures

RNase	Biological origin	Cleavage specificity[a]	Reference	Abbreviation
T1	*Aspergillus orizae*	$Gp^{\downarrow}N$, single-stranded	Rushizky *et al.* (1970), Donis-Keller *et al.* (1977)	T
PhyM	*Physarum polycephalum*	$A/Up^{\downarrow}N$, single-stranded	Donis-Keller (1980)	M
CL3	chicken liver	$Cp^{\downarrow}N$, single-stranded	Boguski *et al.* (1980)	C
U2	*Ustilago sphaerogina*	$Ap^{\downarrow}N$, single-stranded	Rushizky *et al.* (1970), Donis-Keller (1977)	U
V1	cobra venom	double-stranded RNA	Lowman and Draper (1986)	V

[a] The site of cleavage in ribonucleotide sequences is marked by an arrow.

For other possibilities, the user is asked to choose a frame and the length *n* of a fragment which should run to its bottom. The screen is then updated for another choice. Each screen thus simulates in parallel three run times (corresponding to the frames), to allow identification of either small (small *n*) or large (large *n*) fragments, covering together the whole molecule. Each run contains an alkaline ladder (L) for absolute sizing purposes. The current GEL program is arbitrarily limited to RNA molecules not longer than 120 bases, but can easily be modified to deal with longer RNAs.

The *O* selection screen (Figure 4) displays three run times in each frame, each with products of digestion of the molecule, labeled at its 3′-end, by the same enzyme with an accompanying ladder as a length standard. In all cases, a send to plot is possible upon choice.

All displays appear on the screen in real time (defined to be < 1 s), when running on a DEC VAX 8300. They all contain enzyme abbreviations and screen names.

Discussion

The *H* format input of the present form of GEL (Figure 1) is limited to simple RNA foldings containing single hairpin loop structures, and cannot deal with multi-branched loops and pseudoknots (Pleij and Bosch, 1989) such as in tRNA-like structures, nor with knotted structures (Studnicka *et al.*, 1978). However, if such a complex structure is proposed, one can introduce it directly through its implied *M* format (Figure 1).

If at least one of GEL's predictions is completely discordant with the corresponding experimental results (e.g. Knapp, 1989), the tested structure should be abandoned and another one introduced as a new file; the procedure can be repeated until a particular predicted structure is confirmed. However, scientists trying to decipher RNA structures frequently encounter results with varying degrees of ambiguity. These stem from minor bands reflecting secondary recognition sites for enzyme cleavages. For example, base number 2 (G) in Figure 2 is paired and yet it might be recognized by RNase T1 because its 3′ neighbor U (3) is not paired. Such cases may appear as weak bands which can be ignored by the scientist. Alternatively, one could modify GEL so that predefined secondary cleavage sites would result in weaker bands on the screen (and in the plot; Figures 3 and 4).

GEL can be considered as a prototype of programs to handle such problems. One obvious and easy addition could exploit other RNases with different specificities (e.g. Knapp, 1989). It should be possible to prepare a much more sophisticated program that would scan an experimental autoradiogram and

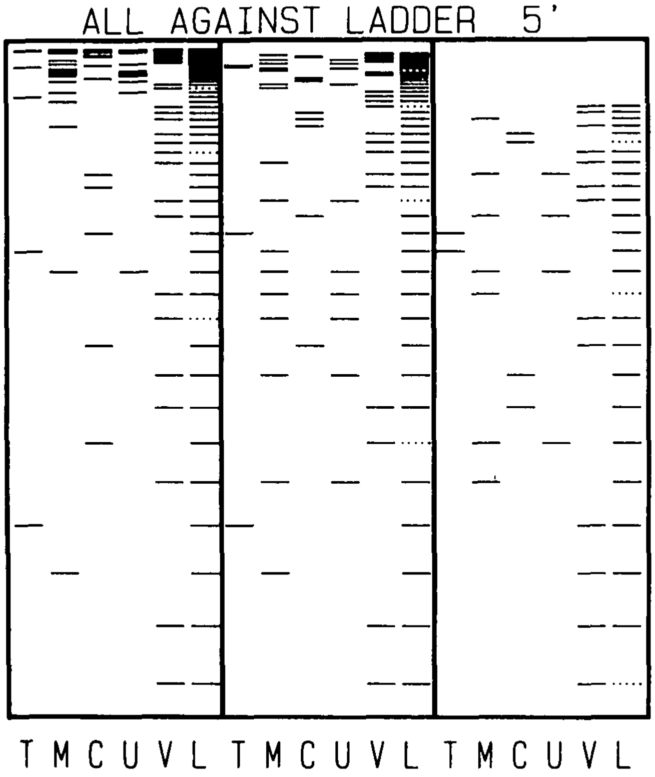

Fig. 3. The fragments generated by cutting the structure of Figure 2, labeled at its 5′-end, after a run to n_1 = 1 (leftmost frame), n_2 = 25 (middle frame) and n_3 = 50 (rightmost frame), each with five enzymes and with an appropriate ladder. Every 10th band on the ladder is dashed.

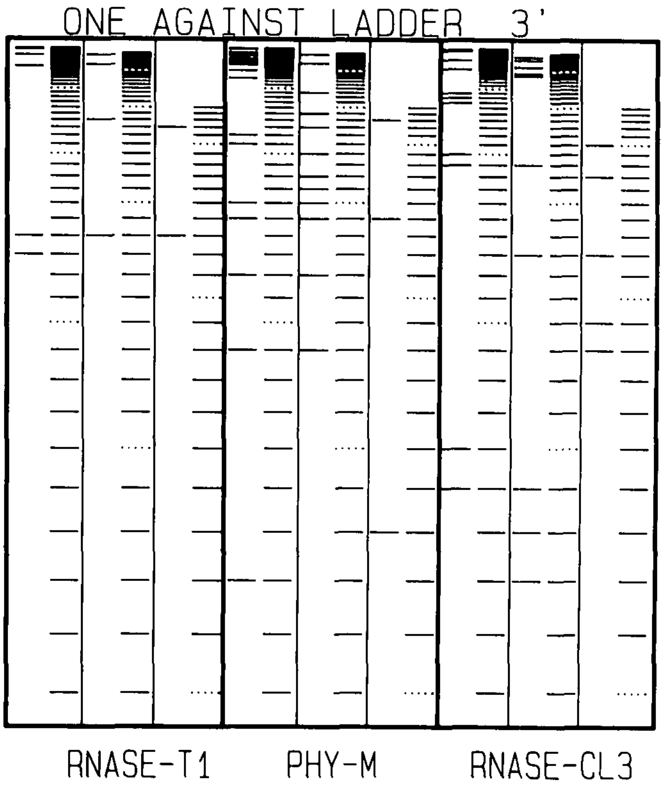

Fig. 4. The fragments generated by cutting the structure of Figure 2, labeled at its 3′-end, by three enzymes as indicated, each after runs as in Figure 3 and with their appropriate ladders. Every 10th band on the ladder is dashed.

process the data to deduce the most likely structure, rather than predict the autoradiogram from a hypothetical structure as does GEL. A simpler program, which can also serve as a preliminary stage to achieve this goal, could sequence an RNA molecule by scanning an autoradiogram of a gel, separating fragments generated by base-specific single-stranded RNases in denaturing conditions. The resultant sequence would resemble the *M* format of GEL (Figure 2), but without marked (paired) bases.

The program GEL is available upon request.

Acknowledgements

Thanks are due for the kind help provided by the staff of the computer center in the Electrical Engineering Department, Dr D.M.Chipman for constructive criticism, attractive ideas and shaping the manuscript, and to Drs D.Bar-Zvi and J.Tal for helpful comments. Shirley and Nitzan Daube initiated this project; they are gratefully acknowledged for thorough discussions. This research received financial support from the Canadian Associates of Ben-Gurion University of the Negev (Biotechnology Laboratory).

References

Ben-Alon,D. (1989) M.Sc. thesis, Ben-Gurion University of the Negev.

Boguski,M.S., Hieter,P.A. and Levy,C.C. (1980) Identification of a cytidine-specific ribonuclease from chicken liver. *J. Biol. Chem.*, **255**, 2160–2163.

Cann,G.M., Zaritsky,A. and Koshland,M.E. (1982) Primary structure of the immunoglobulin J chain from the mouse. *Proc. Natl. Acad. Sci. USA*, **79**, 6656–6660.

Dahlberg,J.E. and Abelson,J.N. (1989) RNA processing (part A: general methods). *Methods Enzymol.*, **180**.

Dahlberg,J.E. and Abelson,J.N. (1990) RNA processing (part B: specific methods). *Methods Enzymol.*, **181**.

Donis-Keller,H., Maxam,M. and Gilbert,W. (1977) Mapping adenines, guanines, and pyrimidines in RNA. *Nucleic Acids Res.*, **4**, 2527–2538.

Donis-Keller,H. (1980) Phy-M: an RNase activity specific for U and A residues useful in RNA sequence analysis. *Nucleic Acids Res.*, **8**, 3133–3142.

Doudna,J.A., Cormack,B.P. and Szostak,J.W. (1989) RNA structure, not sequence, determines the 5' splice-site specificity of a group I intron. *Proc. Natl. Acad. Sci. USA*, **86**, 7402–7406.

Freier,S.M., Kierzek,R., Jaeger,J.A., Sugimoto,N., Caruthers,M.H., Neilson,T. and Turner,D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability *Proc. Natl. Acad. Sci. USA*, **83**, 9373–9377.

Garrett,R.A., Douthwaite,S. and Noller,H.F. (1981) Structure and role of 5S RNA-protein complexes in protein synthesis. *Trends Biochem. Sci.*, **6**, 137–139.

Gehart,E., Wagner,H. and Nordstrom,K. (1986) Structural analysis of an RNA molecule involved in replication control of plasmid R1. *Nucleic Acids Res.*, **14**, 2523–2538.

Goringer,H.U. and Wagner,R. (1988) 5S RNA structure and function. *Methods Enzymol.*, **164**, 721–747.

Gross,R.H. (1986) A DNA sequence analysis program for the Apple Macintosh. *Nucleic Acids Res.*, **14**, 591–596.

Knapp,G. (1989) Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol.*, **180**, 192–212.

Lehrach,H., Diamond,D., Wozney,J.M. and Boedtker,H. (1977) RNA molecular weight determinations by gel electrophoresis under denaturing conditions, a critical reexamination. *Biochemistry*, **16**, 4743–4751.

Lowman,H.B. and Draper,D.E. (1986) On the recognition of helical RNA by cobra venom V1 nuclease. *J. Biol. Chem.*, **261**, 5396–5403.

Martinez,H.M. (1988) An RNA secondary structure workbench. *Nucleic Acids Res.*, **16**, 1789–1798.

Matsuuchi,L., Cann,G.M. and Koshland,M.E. (1986) Immunoglobulin J-chain gene from the mouse. *Proc. Natl. Acad. Sci. USA*, **83**, 456–460.

Pieler,T., Guddat,U., Oei,S.L. and Erdmann,V.A. (1986) Analysis of the RNA structural elements involved in the binding of the transcription factor IIIA from *Xenopus laevis*. *Nucleic Acids Res.*, **14**, 6313–6326.

Pleij,C.W.A. and Bosch,L. (1989) RNA pseudoknots: structure, detection, and prediction. *Methods Enzymol.*, **180**, 289–303.

Rushizky,G.W., Mozeiko,J.H., Rogerson,D.L.,Jr and Sober,H.A. (1970) Characterization of enzymatic specificity of a ribonuclease from *Ustilago spaerogena*. *Biochemistry*, **9**, 4966–4971.

Studnicka,G.M., Rahn,G.M., Cummings,I.W. and Salser,W.A. (1978) Computer method for predicting the secondary structure of single-stranded RNA. *Nucleic Acids Res.*, **5**, 3365–3387.

Williams,A.L.,Jr and Tinoco,I.,Jr (1986) A dynamic programming algorithm for finding alternative RNA secondary structures. *Nucleic Acids Res.*, **14**, 299–315.

Zaritsky,A., Gollop,R. and Cann,G.M. (1988) Tertiary structure of the mRNA coding for J-chain might be involved in differentiation of mature B-lymophocytes to $(IgM)_5$-secretory cells. *Speculat. Sci. Technol.*, **11**, 205–213.

Zehtner,G. and Lehrach,H. (1986) A computer program package for restriction map analysis and manipulation. *Nucleic Acids Res.*, **14**, 335–349.

Zuker,M. (1989) Computer prediction of RNA structure. *Methods Enzymol.*, **180**, 262–288.

Zuker,M. and Stiegler,P. (1981) Optimal folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.

Received on May 28, 1990; accepted on August 28, 1990