Secondary structure model for the complete simian virus 40 late precursor mRNA

Ruth Nussinov and Ignacio Tinoco, Jr.

Department of Chemistry and Laboratory of Chemical Biodynamics, University of California at Berkeley, Berkeley, CA 94720, and

Ann B. Jacobson

Department of Microbiology, State University of New York, Stony Brook, NY 11794, USA

Received 22 September 1981

ABSTRACT

Structures for all sequences containing less than 1790 nucleotides in the 2600 nucleotide late region of the SV40 virus have been computed and saved on magnetic tape. Previously the longest sequence whose secondary structure was calculated in a single computer run contained 950 nucleotides. In the past, analysis of long molecules required numerous repeated, partially overlapping computations on much shorter segments. The structure obtained for the late half of the SV40 is Y-shaped with two unequal arms. It has 52 short hairpins. Two long range interactions between nucleotides near 650 and 1350 and between 1450 and 2450 appear to play an important role. The first is within the 16S intron; the second is in the 3' exon. The 5' and 3' ends of the molecule are close to each other and are found in the major elongated stem in the vicinity of the fork.

INTRODUCTION

The small SV40 virus containing 5224 base pairs has been studied during the last few years. It is considered a model system for eukaryotic gene organization and expression, and for viral transformation of animal cells.

Before viral DNA replication, the early half of the virus is transcribed in a counter clockwise direction (see physical map of SV40, Fig. 1). The early messenger RNA codes for two antigenic proteins, T and t. After the onset of viral DNA replication the late half of the virus is transcribed clockwise into an mRNA which codes for three structural proteins: VP1 (major), VP2 and VP3 (minor) with molecular weights 43, 500-48,000, 39,000 and 27,000 respectively (1). These proteins are synthesized after the precursor RNA has undergone splicing.

SV40 DNA was sequenced three years ago by Reddy et al. (2) and by Fiers and his colleagues (3). The sequence of the viral mRNAs have been identified (4-7) as well as some of the splice sites (Fig. 1). The major leader starts at nucleotide 243 (we follow the numbering system of

ref. 2) and the minor ones at 110 and 182. The RNAs extend to about nucleotide 2585. The VP1 protein is coded by nucleotides 1423-2508, the VP2 by 480-1535, and the VP3 by nucleotides 834-1535. The 19S RNA splice sites are at positions 293 and 477, and the 16S splice sites are at 445 and 1385 (see Fig. 1). The 16S mRNA codes for VP1 protein. The 19S mRNA carries the message for the VP2 and VP3 proteins.

Although intensive investigation of the viral genome and its precursors and processed RNAs has been carried out (8), not much is known about the secondary structure of its RNA. As with structural studies of many messenger RNAs, a major difficulty lies in the size of these molecules. The length of the SV40 precursor RNA is roughly half the viral size, i.e. about 2600 nucleotides.

Using a recently developed efficient computer code (9,10) we present here a calculated secondary structure for the late half of the SV40 i.e. for nucleotides 1-2600 (Fig. 2). Secondary structure calculations on such a scale could not have been attempted before. It is only the availability of this algorithmic tool (10), use of truncated energy values (11), and further computational devices implemented into the code that allowed us to carry out this calculation. Owing to the nature of this

XBL 818-1727

Figure 1. A schematic physical map of the SV40 genome. From the origin of replication transcription (and replication) proceeds clockwise in the late region and anticlockwise in the early. The single EcoRI cleavage site is taken as point 0. Reddy et al. (2) numbering system starts near the origin of replication (at nucleotide 66 of Fiers et al. (3) late region) and continues, clockwise, throughout the circle. Wavy lines indicate spliced out sections.

Figure 2. A schematic structure for the whole late precursor RNA. The loops, mRNA starts and stops and the splice sites are marked. This 2600 nucleotide long structure is the composite of two, partially overlapping calculated structures, 1-1790 and 810-2600. Numbering system followed is of Reddy et al. (2).

algorithm we also have stored on magnetic tape the results of the secondary structure calculations and the corresponding free energies for every subsection (≤ 1790 nucleotides) from the molecule (see also ref. 12).

This information, as well as the algorithm used, is available upon request from the first author.

THE PRINCIPLES OF THE METHOD

Secondary structure calculations use free energy assignments of their various components (13-18). In addition, data such as that derived from the analysis of electron micrographs, enzymatic cleavage studies, crosslinking, chemical modification, and phylogenetic data can be utilized to identify likely folding patterns. For long chains, the number of possible structures is very large and efficient computer codes are also required for thorough and optimal searches.

Despite the limited amount of biochemical data available on the structure of the late SV40 precursor mRNA, we have chosen to attempt to calculate the folding of this sequence by a single computer run as a first effort in analyzing the structure of one very large RNA sequence.

Using this approach we eliminate problems of interpretation associated with structures derived from short overlapping segments, and allow for the formation of long range interactions.

The run was performed on a VAX 11/780 computer with a VMS operating system using the algorithm written in Fortran we have described previously (9). The run was set to examine the full 2600 nucleotides of the mRNA. It was terminated after three months when sections ≤ 1790 nucleotides had been examined. The elapsed running time (CPU) at this point was less than 11 days, however, the time utilized for paging (the time required for relocation of data stored on disc into the physical memory of the computer) was close to two months, and computer time to complete the run was not available. As discussed previously (9), the search procedure employed by the algorithm proceeds in a stepwise fashion such that an optimal solution for all nucleotide segments of a given length is obtained at each step. Thus at the termination of the run, information was available on the optimal structure of all subsections of the sequence which were less than or equal to 1790 nucleotides long.

The structure we present for the 2600 nucleotide sequence was obtained by overlapping the segments from 1-1790 with that from 810 to 2600. The two structures overlap for 980 nucleotides. As might be expected, the major differences in the overlapped sections were found in the long range interactions. In the 1-1790 run the long range pairings in the overlapped section are partly with bases preceding nucleotide 810; in the second section (810-2600) they are partly with nucleotides following 1790. Comparing the close range interactions, most of the hairpin loops, i.e. loops R,U,V,W,Y,Z and part of P (Fig. 2) are present in both structures. Loops Q,S,T and X are present in 1-1790 only. In the total structure for the sequence from 1-2600, the section from 1 to 1400 (loop Z) was taken from the 1-1790 structure. As seen in Fig. 2 it constitutes a closed structure with its end paired with the 5' end of the sequence. From nucleotide 1400 to 1790 the structure formed a separate branch with 4 hairpins growing on it. In the 810-2600 structure the section 1400-1790 was paired with the 3' end of the sequence (Fig. 2). The section 810 to near 1400 is a closed branched structure. Most of its hairpins are identical with those found in the 1-1790 structure. Thus, in the final structure the section 1-1400 was taken from 1-1790, and the rest from 810-2600. The long range interactions and the 5' and 3' ends of the sequence are drawn in detail in Fig. 3.

THE CALCULATION

The secondary structure calculations are based on the SV40 nucleotide sequence (2,3) and the free energy rules as determined by Tinoco and his colleagues (13-15), Delisi and Crothers (16) and Gralla and Crothers (17, 18), and as systematized by Salser (19). The sequence and free energy assignments were used in the secondary structure algorithm which has already been described in detail (9,10). In order to be able to run the code on a molecule this long, two modifications were introduced. The detailed three digit free energy values proposed by Salser were truncated (11). Extensive calculations and analysis showed that the results were not affected by this simplification. In addition, since these calculations were long and since power failure or computer crashes can be expected, a device was put in that allows stopping and restarting the program at will. At fixed intervals the program writes onto a tape the state of the calculations and the structures obtained. Owing to this device we also have now the structures and the corresponding energies of all subsections. Through a monitoring file that the program creates we could easily follow the position, energy and progress of the calculations while the program was running.

THE MODEL

(a). The Structure.

The secondary structure of the late precursor RNA is elongated, Y-shaped (Fig. 2) and is extensively base paired (Fig. 2,3). Whereas out of one of the forked (Y) arms only a few hairpins grow, the second arm is larger, bulkier and more branched. Most of the hairpin structures starting from the 5' end through nucleotide 1400 were already observed in previous studies of the folding process of the growing RNA chain (12). Out of a total of 25 hairpins in this region, 23 were formed by the sequentially folding RNA. Loops G and K (Fig. 2,3) were not observed previously. The nucleotides forming the G loop were previously involved either in a variant hairpin, or in long range pairings (294-310 with 672-694, see ref. 12). The nucleotides forming the K loop were previously involved in a variant hairpin (nucleotides 587-608). The previous structural studies encompassed only a maximum of 950 nucleotides in a single run and were extended to the 3' 16S splice site (at nucleotide 1380).

The 5' and 3' ends of the RNA are close in the planar structure. The major leader starts at position 243 (4-7), which is at the base of

-- XBL 818-1725 --

Figure 3. The detailed secondary structure of a portion of Figure 2. It includes the right hand side arm of the Y and part of the stem. The left hand side arm (loops L-Y) and the bottom part of the stem (loops AA-AX) are not shown.

the forking Y. The minor leaders start at positions 110 and 182, i.e. between loops A and B and before loop D respectively (Fig. 2).

The 5' 16S splice is found in the stem of loop I, at the tip of one of the Y-arms. In all the substructures generated at present or previously (12), if the nucleotides constituting this loop were present, this quite stable loop (Fig. 3) is formed. The 3', acceptor splice is found in the major stem of the precursor RNA. Thus, though the 16S splice sites are distant in the planar secondary structure, their locations suggest that they can easily be drawn together in a three dimensional folding by curving of the right hand side arm. The structure of the intron is compact, and has a lower energy than the exon. The free energy of the exon, on the 5' side of the splice is -0.319 Kcal/nucleotide. The energy of the 940 nucleotide intron (445-1385) is -0.342 Kcal/nucleotide and the energy of the exon on the 3' side (nucleotides 1381-2600) is -0.310 Kcal/nucleotide. Partial results from the sequentially folding RNA had already indicated this trend. Detailed comparison of the structure of the whole precursor RNA with that obtained for the intron only (12), shows that the structures are nearly identical, except, as expected, at the ends of the intron.

The 19S splicings of the primary transcript RNA occur less often than the 16S ones. The 19S 5' splice (at nucleotide 293) is found in the stem of loop G and the 3' splice site (at nucleotide 477) is between loops I and J (Fig. 3). These two sites can be brought together by the arm folding upon itself.

Three structural genes are known to be coded by the late precursor RNA. The VP1 initiator code is in a bulge present in loop Z (Figure 2,3). The VP2 initiator is in the basepaired stem of loop J and the VP3 initiator starts in the head of loop P. The VP2 and VP3 terminator is in the stem of loops AD-AE (Fig.2). The VP1 terminator is in an integral loop in the major stem of the Y.

(b) Some Evidence for the Structure.

Although large structures predicted solely on the basis of thermodynamic rules may have limited validity, it is of interest that the predicted structure is consistent with other data. Shen and Hearst (20) have studied the secondary structure of single stranded SV40 DNA by photochemical crosslinking followed by electron microscopy. In the late region their results indicate the presence of loops at 0.84 ± 0.02 and 0.94 ± 0.01 map units (corresponding to nucleotides 900 ± 100 , 1400 ± 50).

The first of these fits loops P-Q-R and the second which may be the one observed by Hsu and Jelinik (21) at $0.90 - 0.96 (\pm 0.04)$ (1200-1500, ± 200) may correspond to loops X-Y present in the 16S intron (Fig. 2). Further studies by Shen and Hearst (22) discerned several additional loop structures. A long range loop is often observed at position 0.77 ± 0.05 and 0.94 ± 0.03 (nucleotides 500 ± 250 and 1400 ± 150). This loop may coincide with the loops seen in Fig. 2 (and partly Fig. 3) in position 592-1370. An additional loop was located at 0.72 ± 0.03 0.78 ± 0.04 , which coincides with Hsu and Jelinik's data ($0.70 - 0.76$) (nucleotides 250 ± 150 and 550 ± 200) and may correspond with the right hand arm of the forking Y (nucleotides 257-590).

(c) Some Implications of the Model

(i) Protein synthesis regulation.

It has been suggested (23) that high levels of gene expression are obtained when the AUG initiator codon is present at the head of a hairpin loop. In our structure only the VP3 initiator is so located.

(ii) Stability.

The percentage of calculated base pairing (64%) in the SV40 RNA is not extra large. Fig. 2 shows, however, that the 5' and 3' ends of the molecule are close to each other. Had we finished the computer run they might have formed a base paired stem stabilizing the messenger and inhibiting exonuclease activity.

(iii) The Splicing Event.

In any model for splicing, several criteria should be met (24). Spatial proximity and proper orientation of the sites to be spliced and ligated are crucial. It is also essential that the structure be stable and hold together long enough in this favorable conformation for splicing and ligation to take place. As seen in Figs. 2 and 3 the 16S splice sites which are about 940 nucleotides apart in the primary chain can easily be brought together in three dimensional folding by bending of the right hand arm of the Y. The positions of the 16S spliced sites in the proposed late structure (Figs. 2,3), i.e. one site at the end of one arm and the other on the major stem, suggest that steric hindrance may not arise. This would satisfy the first two criteria.

Comparison of the late 16S partial intronic and partial exonic structures shows that in all cases the intron has a lower energy than the exon (12). Although in the late SV40 precursor RNA the lowest energy structure yields a very compact intron, we do not propose that this is a

general phenomenon. The intron sequences may be long, comprising hundreds or thousands of nucleotides. It is unlikely that in each of these cases the intron has lower energy than the exon, and that it folds very tightly and brings into proper juxtaposition the splicing sites. Moreover, intron portions can be deleted or inserted, exons can be reshuffled (8), and splicing at the same sites may still take place. We suggest, however, that in order for efficient splicing and ligation to take place, the intron has to have a low energy structure (not always the lowest) and the exons have to bring into proper alignment the sequences to be spliced.

(iv) The Sequential Folding Process.

Studies on the folding process of the RNA as it is being synthesized showed that no major reshuffling of the structure of the molecule occurred (12). For the most part the close range hairpins recurred and usually only the long range stems rearranged. These studies were carried out on portions of the SV40 late RNA molecule which had the same 5' end and different lengths i.e. 1-100, 1-200.....1-900 and 1-950 which was then the largest section we could run. Additional sections, with different 5' ends (like the intron itself) were run too. In the present study we extended the previous analysis, i.e. 1-1000, 1-1100.....1-1790. The same sequential nature of RNA folding was observed. Comparing the final structure with the close-range hairpin loops seen in the previously analyzed regions (12) yields 23/25 (92%) loop recurrence.

(v) Long Range Interactions.

In the calculated structure two sets of long range interactions are important, namely, between nucleotides 650 and 1350 and between 1450 and 2450. The first encompasses most of the intron, forms the left arm of the Y and thus allows a possible spatial proximity between the 16S splice sites. The second brings the 3' end of the molecule close to its 5' end.

THE THERMODYNAMIC APPROACH TO RNA FOLDING

The present approach is based solely on the current thermodynamic data (13-18). We did not have additional biological information, such as enzyme cleavage sites, chemical modification or detailed results of cross linkage studies. If available, such information can be incorporated into (and will considerably accelerate) the calculation. The program neglects spatial hindrance, tertiary structure and in vivo interactions with proteins. We do not have the complete physical-chemical (or combinatorial) tools necessary for such an undertaking. Even in solution, we neglect the

stabilizing effects of Mg^{+2} ions and of free dangling ends (25,26). Nonetheless, the calculation was exactly carried out within a well defined framework, i.e. the "folding rules" abstracted from oligonucleotide studies. We then incorporate base stacking/pairing stabilization and hairpin, bulge and internal loop destabilization. The whole structure (or all segments of length ≤ 1790 in the present case) was computed in a completely objective, step by step procedure.

In addition to the complete secondary structure, the program created the energy and secondary structure half-matrices. These form a single 2600×2600 matrix available on tape. From this matrix one can immediately read the stability of any of the $(2600)^2/2$ subsections (strictly only $(1/2) [(2600)^2 - (810)^2]$ subsections of length ≤ 1790). The most stable secondary structure can be reconstructed in milliseconds using a simple backtracking procedure for any subsection. This allows us to answer any specific "local" question pertaining to a given area (such as, does a particular loop exist between nucleotides i,j ?). Thus, within a given set of folding rules all structural information pertinent to the RNA in question can be obtained in a single run and stored on tape.

One can also systematically address the question of common structures of biologically related mRNAs by comparing the energy/structure matrices. The automatic plotting program for secondary structure recently devised by R. Feldmann, NIH (personal communication) and interfaced with the algorithm may be extremely helpful in such a venture.

CONCLUSIONS

The largest RNA whose structure has been analyzed to date is that of 16S ribosomal RNA (28,29,30). Both biochemical and phylogenetic data have been used in conjunction with computer searches to generate a probable secondary structure for this molecule. Although the models proposed in different laboratories differ slightly from one another, two are rather close (28,29). This work represents an important step towards understanding ribosome structure. Nonetheless it is clear that additional tools will be required both in refining the structure of 16S ribosomal RNA and in the analysis of other large structures. A very important capability is the ability to handle large sequences in a single computer run. The computer studies with ribosomal RNA were performed by overlapping short segments. Aside from being laborious in nature, this approach has several problems associated with it. Most important, the choice between different

structures in the overlapping regions is arbitrary. In our hands, the minimal energy structure found by computer for large segments is often quite different from the structures we have generated from two overlapping segments. While correct short range interactions may be obtained by this approach, long range interactions between the overlapping segments are certainly missed. Also, end effects, such as improper hydrogen bonding of the 5' and 3' ends of the molecule, may occur. Thus, it is clearly preferable to run an entire sequence in one run.

In the present work we have demonstrated the feasibility of analyzing very large nucleotide sequences by computer. The structure we have obtained may have some validity since it is consistent with major splice sites which are involved in processing the RNA and compatible with certain loops seen by electron microscopy in the structure of psoralencrosslinked SV40 DNA (20,22). Nonetheless, it should be recognized that these results represent only a first approximation to the structure of the molecule. Limitations to its validity derive both from our lack of the appropriate thermodynamic data and the lack of combinatorial tools to predict the full three dimensional structure of the molecule, and also from limitations in the search procedure of the algorithm as we have employed it in this study. Recent studies have shown that the search procedure is not as exhaustive as we previously believed. Small hairpins are not always correctly identified, and in special instances certain long range interactions can be missed; these problems are not intrinsic to the combinatorial method employed. They do not seem to exist in the program described by Zuker and Stiegler (27). This program is, however, slower than ours, and could not have been used in the current study.

Our current efforts are being devoted to improving both the speed and search procedure of the algorithm. Owing to a suggestion by D. Kleitman from the Department of Mathematics at M.I.T. (personal communication), it should be possible to drastically reduce both paging and CPU time, and possibly enable an order of magnitude increase in the scope of the calculation. Although the time requirement for the calculation we have described here may seem prohibitive for routine use, we are confident that with the improvements outlined above, runs of this size will become routine in the near future and folding of larger sequences can begin to be contemplated.

ACKNOWLEDGEMENTS

We wish to thank D. Canaani for discussions and suggestions. We also wish to thank K. Wiley and S. Wong from the Laboratory of Chemical Biodynamics Computer Center for their time, advice and help in creating the monitoring files, and in carrying out the test runs. Especially, we are indebted to J. L. Dilworth from the Lawrence Berkeley Laboratory Computer Center who enabled the execution of this project. Without his help, time and patience, this project could not have been carried out.

One of us (R.N.) also wishes to thank Dr. H. A. Scheraga, J. Dunn and N. Hansen for their warm hospitality during the time in which the manuscript was written.

This work was supported in part by National Institutes of Health Grant GM10840 (IT), AI 15273(AJ) and by the Division of Biomedical and Environmental Research of the U.S. Department of Energy under Contract No. W-7405-Eng. 48.

REFERENCES

1. Salzman, N.P. and Khoury, G. (1974) in *Comprehensive Virology*, Fraenkel-Conrat, H. and Wagner, R.R. Editors, pp. 63-141, Plenum, NY.
2. Reddy, V.B., Thimmappaya, B., Dhar, R., Subramanian, K. N., Zain, B.S., Pan, J., Ghosh, P.K., Celma, M.L. and Weissman, S.M. (1978) *Science* 200, 494-502.
3. Fiers, W., Contreras, R., Haegman, G., Rogiers, R., Van de Voorde, A., Van Heuverswyn, H., Van Herreweghe, J., Volckaert, G. and Ysebaert, M. (1978) *Nature* 273, 113-120.
4. Lai, C.-J., Dhar, R. and Khoury, G. (1978) *Cell* 14, 971-982.
5. Villarreal, L.P., White, R.T. and Berg, P. (1979) *J. Virol.* 29, 209-219.
6. Ghosh, P.K., Reddy, V.B., Swinscoe, J., Lebowitz, P. and Weissman, S.M. (1978) *J. Mol. Biol.* 126, 813-846.
7. Bina, M., Stein, A., Thoren, M., Salzman, N. and Thompson, J.A. (1979) *Proc. Natl. Acad. Sci. USA* 76, 731-735.
8. Hamer, D.H. and Leder, P. (1979) *Cell* 18, 1299-1302.
9. Nussinov, R. and Jacobson, A. (1980) *Proc. Natl. Acad. Sci. USA* 77, 6309-6313.
10. Nussinov, R., Pieczenik, G., Griggs, J. and Kleitman, D. (1978) *Soc. Ind. Appl. Math (C) J. Appl. Math* 35, 68-82.
11. Nussinov, R., Tinoco, I., Jr. and Jacobson, A.B. (1981) *Nucleic Acids Res.*, submitted.
12. Nussinov, R. and Tinoco, I. Jr. (1981) *J. Mol. Biol.* in press.
13. Borer, P.N., Dengler, B., Tinoco, I., Jr. and Uhlenbeck, O.C. (1974) *J. Mol. Biol.* 86, 843-853.
14. Tinoco, I., Jr., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbeck, O.C., Crothers, D.M. and Gralla, J. (1973) *Nature New Biol.* 246, 40-41.
15. Tinoco, I., Jr., Uhlenbeck, O.C. and Levine, M.D. (1971) *Nature* 230, 362-367.

16. Delisi, C. and Crothers, D.M. (1971) *Proc. Natl. Acad. Sci. USA* 68, 2682-2685.
17. Gralla, J. and Crothers, D.M. (1973a) *J. Mol. Biol.* 73, 497-511.
18. Gralla, J. and Crothers, D.M. (1973b) *J. Mol. Biol.* 301-319.
19. Salser, W. (1977) *Cold Spring Harbor Symp. on Quant. Biol.* 42, 985-1002.
20. Shen, C.-K. J. and Hearst, J.E. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 1363-1367.
21. Hsu, M.-T. and Jelinik, W.R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* 74, 1631-1634.
22. Shen, C.-K. J. and Hearst, J.E. (1979) *Analytical Biochem.* 95, 108-116.
23. Iserentant, D. and Fiers, W. (1980) *Gene* 9, 1-12.
24. Nussinov, R. (1980) *J. Theor. Biol.* 83, 647-662.
25. Martin, F.H., Uhlenbeck, O.C. and Doty, P. (1971) *J. Mol. Biol.* 57, 201-215.
26. Romaniuk, P.J., Hughes, D.W., Gregoire, R.J., Neilson, T. and Bell, R.A. (1978) *J. Am. Chem. Soc.* 100, 3971-3972.
27. Zuker, M. and Stiegler, P. (1981) *Nucl. Acid Res.* 9, 133-148.
28. Woese, C.R., Magrum, L.J., Gupta, R., Siegel, R.B., Stahl, D.L., Kop, J., Crawford, N., Brosius, Jr., and Noller, H.F. (1980) *Nucl. Acids Res.* 8, 2275-2294.
29. Stiegler, P., Carbon, P., Zuker, M., Ebel, M.-P. and Ehresmann, C. (1981) *Nucl. ACids Res.* 9, 2153-2172.
30. Glotz, C. and Brimacombe, R. (1980) *Nucl. Acids Res.* 8, 2377-2395.