A Technique for Relating Long-Range Base Pairing on Single-Stranded DNA and Eukaryotic RNA Processing

CHE-KUN JAMES SHEN* AND JOHN E. HEARST†

*Division of Biology, California Institute of Technology, Pasadena, California 91125 and
†Department of Chemistry and Laboratory of Chemical Biodynamics
University of California, Berkeley, California 94720

Received October 23, 1978

Our previous results have shown that after photochemical crosslinking with psoralens in solutions of different concentrations of NaCl, short (100-300 base pairs) hairpins can be stabilized on denatured Simian virus 40 (SV40) DNA for electron microscope visualization. Six major hairpins were mapped. This study has been extended by crosslinking single-stranded SV40 linear molecules, EcoRI-SV40 and Bgl I-SV40, in the presence of magnesium ions. After the reaction, in addition to the hairpins detected before, seven DNA hairpins, and hairpin loops with sizes from 300 to nearly 2000 nucleotides are observed. These loops are located at specific regions on the SV40 genome; at the EcoRI map positions (0.16 ± 0.02)/(0.26 ± 0.03), (0.33 ± 0.03)/(0.39 ± 0.03), (0.51 ± 0.03)/(0.57 ± 0.03), (0.72 ± 0.03)/(0.78 ± 0.04), (0.74 ± 0.03)/(0.85 ± 0.03), (0.77 ± 0.05)/(0.94 ± 0.03), and at (0.74 ± 0.04)/(0.09 ± 0.07). At least three of these loops are found in regions thought to be involved in splicing of the early and late SV40 transcripts: (0.51 ± 0.03)/(0.57 ± 0.03), (0.72 ± 0.03)/(0.78 ± 0.04), (0.77 ± 0.05)/(0.94 ± 0.03). These loop structures are most likely to arise from base pairing between distant regions of the single-stranded DNA. RNA transcribed from the double-stranded DNA template is expected to behave in a similar way, possibly providing recognition sites for processing (splicing) enzymes.

A novel "splicing" phenomenon associated with gene organization and expression has recently been discovered in many eukaryotic systems, including the genes of adenovirus (1-3), SV40 virus (4-8), and specific eucaryotic genes coding for ribosomal RNA (9-11), tRNA (12-14), ovalbumin (15-17), globins (18,19), and immunoglobulin (20). In these systems, the corresponding cytoplasmic messenger RNAs (mRNA) are derived from more than two noncontiguous segments of their DNA templates. The intervening sequences of DNA between the flanking RNA-coding regions have been called either "inserts" (18) or "intervening sequences" (19).

The most basic and important question that arises immediately is how the coding information of the noncontiguous regions is joined to generate the mature cytoplasmic RNA molecules. In general, four mechanisms have been proposed (3,4,14). First, RNA pieces could be transcribed from the noncontiguous RNA-coding regions and then rejoined by cellular enzymes. Second, RNA polymerase could transcribe the noncontiguous coding regions continuously without incorporation into the RNA of the information within the intervening sequences. This could be accomplished either by transcription on a DNA template in which the intervening DNA sequences are looped out so that the polymerase moves easily from one coding region to the next one, or by a con-
tinuous movement of the polymerase along the DNA template with no RNA biosynthesis in the intervening sequence regions due to unknown “start–stop–start” signals. The third mechanism incorporates DNA recombination which removes the intervening sequences from the DNA template before it is transcribed. In the fourth mechanism, a long transcript containing both the coding and intervening sequences is synthesized. A specific cellular enzyme then breaks and rejoins this precursor at specific sites so that the intervening sequences are excised.

Biochemical evidence presently available seems to favor the last model. A 15 S precursor of mouse β-globin mRNA has been isolated and shown by electron microscopy to contain the intervening sequence of the genomic globin gene (21). In addition, an enzymatic RNA “splicing” activity has been detected recently in yeast which converts at least three tRNA precursors to their matured form (14).

An extension of the fourth model predicts that the regions surrounding the two ends of each intervening sequence are brought together by the base pairing of inverted repeats. The splicing enzyme then recognizes this specific base-paired region, cleaves at specific sites in this base-paired region, and then ligates the two coding regions originally separated by the intervening sequence. Similar schemes have been proposed for the processing of T7 early RNA species (22) as well as that of the viral or eucaryotic mRNAs mentioned above.

Since most of the intervening sequences observed in viral genomes or in eucaryotic DNA are from a few hundred to over 1000 base pairs long, it is possible to prove the existence of inverted repeats near the two ends of each intervening sequence by direct observation of specific hairpin loops in an appropriate sequence of single-stranded DNA utilizing electron microscopy. Previously, we stabilized and mapped six short hairpins on denatured SV40 DNA in NaCl solutions using the photochemical crosslinking reagent, 4,5',8-trimethyl psoralen (tri-oxsalen) (23). In this communication, we describe the results of crosslinking single-stranded SV40 DNA molecules with tri-oxsalen in the presence of magnesium ions. Seven additional DNA hairpin and hairpin loop sizes from 300 to nearly 2000 nucleotides are observed with high frequency in the microscope. Furthermore, three, and possibly four, of these structures coincide with the regions where the intervening sequences of SV40 have been mapped.

MATERIAL AND METHODS

Materials. Supercoiled DNA was kindly provided by Richard Mulligan, Dr. Steven Goff, and Dr. Tao-Shih Hsieh at Stanford University. Trioxsalen powder was purchased from Paul B. Elder Company and was dissolved in 100% EtOH to 1 mg/ml to form a stock solution. Restriction enzymes EcoRI and Bgl I were gifts from Dr. Paul Modrich and Dr. Tom Maniatis, respectively.

Photochemical crosslinking of denatured SV40 DNA. The irradiation of denatured restriction fragments of SV40 DNA with long-wavelength uv light in the presence of tri-oxsalen was performed as described in Ref. (23). The only difference was that the samples were dialyzed against 1 mM Tris·HCl (pH 7.5), 0.1 mM NaEDTA, and then brought to the desired ionic concentration by addition of a buffer containing 3 mM MgCl₂, 150 mM KCl, and 50 mM Tris·HCl (pH 7.5) (buffer A).

Electron microscopy. The DNA samples were spread in 50% formamide using the protein monolayer technique (24,25) and studied by electron microscopy as described before (23).

RESULTS

Increase of Loops on Single-Stranded SV40 DNA in the Presence of MgCl₂ and KCl

Specific linear single-stranded SV40 DNA molecules of genome length were generated
and subjected to psoralen crosslinking at 15°C in the presence of a series of dilutions of buffer A. In 1 mM Tris·HCl (pH 7.5) and 0.1 mM NaEDTA there is no observable secondary structure on the molecules; but, just as we observed in NaCl solutions, (23) the average number of DNA hairpins per molecule increases with increasing ionic strength. The DNA molecules eventually collapse to bush-like forms at high ionic strength. In this study, the optimum range of ionic concentrations for investigating the helix-coil transition of single-stranded SV40 DNA at 15°C is from \(1/70\) buffer A to \(1/30\) buffer A. The secondary structures were clearly discernible and molecules were not shortened greatly under these conditions.

Linear molecules generated from either EcoRI or Bgl I cleavage and crosslinked in \(1/50\) buffer A have been used to map secondary structure. We found that, in addition to the six major unlooped hairpins observed before (23) (hairpins a–f in Fig. 4), specific hairpins containing approximately 300 nucleotides and large hairpin loops containing 500 to 2000 nucleotides were detected on many of the molecules (Fig. 1). Since noncrosslinked DNA can also fold back to form loops during the monolayer spreading, we have carried out an analysis of the percentage of molecules with discernible loops (looped molecules). The data are listed in Table 1. As can be seen, the presence of MgCl\(_2\) and KCl dramatically increases the number of looped molecules after crosslinking. Under conditions where the single-stranded SV40 DNA molecules are shortened by the crosslinking reaction to equivalent extents, DNA samples treated in MgCl\(_2\) and KCl have two-fold more looped molecules than those crosslinked in NaCl solutions. Elimination of the irradiation step reduces the percentage of looped molecules by more than another factor of 2 (Table 1), indicating that the formation of these large DNA loops results primarily from the covalent DNA-psoralen-DNA photoreaction.

Mapping of Three New Small Hairpins on Denatured SV40 DNA

As a first step in mapping the large DNA loops, we have investigated whether the same short hairpins previously found on denatured SV40 DNA in NaCl (23) are stabilized by trioxsalen and uv light in MgCl\(_2\) and KCl. It was found that most of the six hairpins formed in NaCl (hairpins a–f, Fig. 4) appear with high frequencies (approximately 30 to 60%) in \(1/50\) buffer A. The exceptions arise with hairpins a and f, which are often not present on the looped molecules. The reason of this will become clear with the mapping of the large DNA loops.

Besides hairpins a–f, new small hairpin structures were also detected with high frequencies and were mapped at specific regions on the SV40 genome (Fig. 2) using the six hairpins a–f and the cleavage sites for the EcoRI and Bgl I enzymes as reference points. Because these hairpins didn't show up with high yield in NaCl (less than 20%), we think they probably contain a large number of mismatched bases or unpaired regions which are stabilized only by the presence of Mg\(^{2+}\) ion. They are designated in Fig. 4 as A, B, and C, respectively, in the order of their positions on the EcoRI–SV40 map. From Fig. 2, the frequencies of these three hairpins are 41% for A and B, and 22% for C on the looped molecules. On unlooped molecules (data not shown), the frequencies are 29, 71, and 82% for A, B, and C. As with hairpins a and f, the formation of hairpin C is inhibited by the appearance of the large loops. Hairpins A, B, and C are formed in the regions of the EcoRI–SV40 map at (0.33 \(\pm\) 0.03)/(0.39 \(\pm\) 0.03), (0.51 \(\pm\) 0.03)/(0.57 \(\pm\) 0.03), and (0.72 \(\pm\) 0.03)/(0.78 \(\pm\) 0.04) map units.

Mapping of the Large DNA Loops

As we have mentioned earlier in Table 1, 40 to 50% of the denatured SV40 DNA molecules form large loops in \(1/50\) buffer A at 15°C. On either denatured EcoRI–SV40
FIG. 1. Electron micrographs of examples of denatured EcoRI–SV40 molecules (d-EcoRI–SV40) or denatured Bgl I–SV40 molecules (d-Bgl I–SV40) crosslinked by trioxsalen at 15°C in % buffer A. (I) Unlooped d-Bgl I–SV40; (II) loop D on d-EcoRI–SV40; (III) loop E on d-EcoRI–SV40; (IV) loop E on d-EcoRI–SV40; (V) loop F on d-Bgl I–SV40; (VI) loop F on d-EcoRI–SV40; (VII) loop F on d-Bgl I–SV40; (VIII) loop G on d-Bgl I–SV40. The bars represent the length of 1000 nucleotides. For other details, see text.

fragments or denatured Bgl I–SV40 fragments, most of these loops (approximately 70%) were found to locate near one end of the molecule. After a careful mapping analysis, using the positions of the hairpins a through f, A, B, and C, and the restriction enzyme cleavage sites, we have identified at least two major classes of specific loops.
TABLE 1

STATISTICAL MEASUREMENT OF DENATURED SV40 MOLECULES IN THE ELECTRON MICROSCOPE

<table>
<thead>
<tr>
<th>Sample number</th>
<th>Salt conditions during crosslinking or dark control</th>
<th>Percentage of molecules containing at least one looped hairpin</th>
<th>Shortening factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncrosslinked samples</td>
<td>NaCl</td>
<td>7 ± 3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>MgCl₂</td>
<td>12 ± 2</td>
<td>0.95 ± 0.03</td>
</tr>
<tr>
<td>2</td>
<td>KCl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crosslinked samples</td>
<td>NaCl</td>
<td>20 ± 4</td>
<td>0.70 ± 0.05</td>
</tr>
<tr>
<td>3</td>
<td>NaCl</td>
<td>27 ± 5</td>
<td>0.75 ± 0.03</td>
</tr>
<tr>
<td>4</td>
<td>NaCl</td>
<td>51 ± 5</td>
<td>0.87 ± 0.09</td>
</tr>
<tr>
<td>5</td>
<td>KCl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MgCl₂</td>
<td>40 ± 5</td>
<td>0.75 ± 0.05</td>
</tr>
<tr>
<td>7</td>
<td>KCl</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* The different samples are: (1) uncrosslinked d-EcoRI-SV40; (2) uncrosslinked d-Bgl I-SV40 equilibrated with 1/50 buffer A at 15°C for 10 min, then equilibrated with 10 µg/ml trioxsalen in the dark for another 10 min before spreading; (3) d-EcoRI-SV40 crosslinked at 15°C in 20 mM NaCl; (4) d-EcoRI-SV40 crosslinked in 30 mM NaCl; (5) d-EcoRI-SV40 crosslinked in 1/50 buffer A; (6) d-Bgl I-SV40 crosslinked in 1/50 buffer A.

* Each percentage is the average value of 200-500 molecules obtained in two or three spreadings of the same sample.

* Each shortening factor is the average value of 200-500 molecules obtained in two or three spreadings of the same sample.

One of them (designated as loop F in Figs. 1 and 4) is formed between (0.77 ± 0.05) and (0.94 ± 0.03) units (Fig. 1, V-VII). Approximately 40% of the loops belong to this class. Often the hairpin e can be found within this big loop (for example, see Fig. 1, V and VI). The other major class of loop (designated as loop G in Figs. 1 and 4) is formed by connecting the regions (0.74 ± 0.04) and (0.09 ± 0.07) (Fig. 1, VIII). Loop G only appears on denatured Bgl I-SV40 fragments, with a frequency of 12% per molecule. Similar to the situation in loop F, hairpins (e, f, or both) can form within loop G. A schematic figure indicating the map positions of loops F and G is shown for an array of molecules in Fig. 3.

Two minor classes of specific loops (designated as D and E) are both approximately 550 nucleotides long and appear in the regions (0.74 ± 0.03)/(0.85 ± 0.03) and (0.16 ± 0.02)/(0.26 ± 0.03) on the EcoRI-SV40 map (examples are shown in Fig. 1, II-IV). Because of their relatively low frequencies of formation (less than 15% of the total loops), sometimes it is not easy to differen-

---

**Fig. 2.** Array of 37 looped molecules displaying the hairpins A, B, and C. Twenty-one d-EcoRI-SV40 and sixteen d-Bgl I-SV40 molecules crosslinked at 15°C in 1/50 buffer A are measured and aligned on the EcoRI-SV40 map. For simplicity, the big loops and other hairpins are not shown. The black bars on each molecule represent the DNA regions bounded by the two stems of each hairpin.
FIG. 3. Array of 33 molecules displaying loop F or loop G. Twenty-one d-Bgl I–SV40 and twelve d-EcoRI–SV40 molecules crosslinked at 15°C in ½o buffer A are measured and aligned on the EcoRI–SV40 map. The black bars on each molecule represent the DNA region bounded by the stems of each large loop.

The formation of the loops E, F, or G inhibits the formation of the other two loops. Furthermore, whenever loop E, F, or G forms, the hairpin C [(0.72 ± 0.03)/(0.78 ± 0.04)] is missing from that particular molecule. Hairpins b and c, e, or f cannot be detected on molecules possessing loops D, E, or F, respectively.

DISCUSSION

In summary, a total of 13 secondary structures have been detected and mapped to specific regions on the SV40 viral genome in the presence of 3 mM KCl, 2 mM Tris·HCl (pH 7.5), and 0.06 mM MgCl₂ at 15°C. The locations, sizes, and frequencies of formation of the seven newly detected hairpins and hairpin loops are shown in Fig. 4 and Table 2. Among the seven new secondary structures, at least two hairpins, B and C, have been observed before by spreading single-stranded SV40 DNA in a mixed solution of formamide and ammonium acetate (26). In thermal denaturation studies (27,28), it has been shown that the magnesium ion can stabilize the DNA double helix much more efficiently than sodium ion or other monovalent cations. Furthermore, at low ionic strength (10⁻³ M Na⁺), the binding of Mg⁺² ion to the DNA helix dominates that of the Na⁺ ion binding. The enhancement of hairpin and loop formation on single-stranded SV40 by the substitution of 30 mM NaCl with ½o buffer A is probably due to the helix stabilization effect of the magnesium ions. This effect has also been
TABLE 2  
STATISTICAL DATA OF SECONDARY STRUCTURES OF DENATURED SV40 DNA IN 1/50 BUFFER A AT 15°C  

<table>
<thead>
<tr>
<th>Secondary structure</th>
<th>Positions on EcoRI-SV40</th>
<th>Average loop size a (%)</th>
<th>Frequency on looped molecules b (%)</th>
<th>Frequency on unlooped molecules (%)</th>
<th>Frequency on total molecules (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(0.33 ± 0.03)/(0.39 ± 0.03)</td>
<td>160</td>
<td>41</td>
<td>29</td>
<td>34</td>
</tr>
<tr>
<td>B</td>
<td>(0.51 ± 0.03)/(0.57 ± 0.03)</td>
<td>160</td>
<td>41</td>
<td>71</td>
<td>58</td>
</tr>
<tr>
<td>C</td>
<td>(0.72 ± 0.03)/(0.78 ± 0.04)</td>
<td>160</td>
<td>22</td>
<td>82</td>
<td>55</td>
</tr>
<tr>
<td>D</td>
<td>(0.16 ± 0.02)/(0.26 ± 0.03)</td>
<td>520</td>
<td>6</td>
<td>—</td>
<td>3</td>
</tr>
<tr>
<td>E</td>
<td>(0.74 ± 0.03)/(0.85 ± 0.03)</td>
<td>570</td>
<td>13</td>
<td>—</td>
<td>6</td>
</tr>
<tr>
<td>F</td>
<td>(0.77 ± 0.05)/(0.94 ± 0.03)</td>
<td>890</td>
<td>39</td>
<td>—</td>
<td>17</td>
</tr>
<tr>
<td>G</td>
<td>(0.74 ± 0.04)/(0.09 ± 0.07)</td>
<td>1830</td>
<td>31</td>
<td>—</td>
<td>12</td>
</tr>
</tbody>
</table>

a The average loop size of A, B, or C is the length of the hairpin in base pairs while that of the large loops D, E, F, or G is the length of the loop in nucleotides.

b The frequencies of observation of loop G refer only to those of crosslinked d-Bgl I-SV40 molecules.

seen and used before in the study of secondary structures on MS2 viral RNA molecules and in Escherichia coli ribosomal RNA (29,30). As we have mentioned, hairpins A, B, and C may very well contain more mismatched base pairs or unpaired regions than hairpins a–f. The base paired stems of most of the big loops (D, E, F, and G) appear only as crossover points. This implies that the two distal portions that are responsible for the formation of each loop contain short complementary regions, probably less than 100 nucleotides. Long single-stranded DNA segments are interspersed between these distal complementary regions, and only in the presence of magnesium ions are the loops stable enough to be observed on a significant number of molecules.

The most striking feature of these relatively unstable secondary structures is the coincidence of their locations with the known SV40 intervening sequences. At least four intervening sequences have been discovered on the SV40 genome. Two of them occur in the SV40 early region which codes for two SV40-specific polypeptides, the T antigen and t antigen [for recent reviews on SV40, see Refs. (31,32)]. They have been mapped biochemically and genetically to regions 0.54–0.55 and 0.54–0.59 map units, respectively (8,33). This is where hairpin B has been mapped. In the late or postreplication stage of SV40 infection, two major mRNA species have been detected in the cell cytoplasm, the 16 S and 19 S mRNAs. They code for the viral coat proteins VP1 and VP2, respectively. Each of them is also composed of at least two noncontiguous segments of the SV40 genome (4–7). The 16 S mRNA is derived from the SV40 genomic regions 0.72–0.75 and 0.94–0.17, while the 19 S mRNA contains sequences of 0.72–0.75 and 0.76–0.17. Indeed, hairpin C [(0.72 ± 0.03)/(0.78 ± 0.04)] and loop F [(0.77 ± 0.05)/(0.94 ± 0.03)] are formed in the regions where the “splicing” occurs for the late SV40 messengers. Although strong evidence has not yet been presented for the relationship between the other secondary structures and splicing of SV40 transcripts, loop E [(0.74 ± 0.03)/(0.85 ± 0.03)] is an interesting one to speculate on. It is known that the viral protein VP3 of polyoma is made on an 18 S mRNA (34). Fiers et al. (31) have proposed that a similar VP3 coding 18 S mRNA exists for SV40 and contains the sequences 0.72–0.75 and 0.84–0.17. If this is the case, then loop E [(0.74 ± 0.03)/(0.85 ± 0.03)] would correspond to a fifth SV40 intervening sequence. Hairpin A and loops D and G are located in regions with unknown biological functions.
In view of the fact that all four intervening sequences of SV40 are bounded by the stems of specific hairpins or hairpin loops, we propose here that the higher order structures on the long nuclear transcripts of eucaryotes are important for bringing the two ends of the intervening sequences into proximity. They could form under appropriate temperature and salt conditions on single-stranded RNA transcripts and be recognized by cellular "splicing" enzymes. The RNA segments containing the intervening sequences are then excised after the "breaking and rejoining" events have occurred within the hairpins on the stems of the large loops. This proposal predicts that the relative abundance of different mRNAs is partly controlled by the probability of formation of hairpin or hairpin loop structures. From Table 2 and Fig. 4, we can see that the members in each of the following sets of secondary structures are probably formed from the same or nearly the same sequences and thus exclude each other on the same DNA or RNA molecule: hairpins b and c, loop D; hairpins e, C, loop E; hairpins f, C, loop F; hairpins a, C, loop G; hairpins C, loops E, F, G. Heterogeneity has been noticed in the positions of all the hairpins and hairpin loops. It is not possible at present to assess whether this heterogeneity originates from multiple positions of crosslinked secondary structure or if it arises from the limitations of the electron microscope measurements.

While the nucleotide sequence of the total SV40 genome is now known (31,32) and is useful in predicting stable short-range secondary structures of single-stranded SV40 DNA or RNA (35,36), the psoralen crosslinking technique coupled with electron microscopy is able to detect imperfect short hairpins as well as long-range secondary structures of biological significance. It will be interesting and important to see whether the "intervening sequences–secondary structures" relationship holds true for most, if not all, of the eucaryotic genes containing intervening sequences.

ACKNOWLEDGMENTS

We thank Richard Mulligan, Drs. Tao-Shih Hsieh, Steven Goff, Paul Berg, and Tom Maniatis for providing us with valuable materials. C.-K. J. Shen appreciates Tom Maniatis's continuous interest and encouragement in this study. This work has been supported by National Institutes of Health Grant GM11180 and by American Cancer Society Grant NP185. C.-K. J. Shen is now the recipient of NIH Postdoctoral Fellowship GM 07401.

REFERENCES