Statistical significance of partial base-pairing potential: implications for recombination of SV40 DNA in eukaryotic cells

M.A.Savageau, R.Metter and W.W.Brockman

Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA

Received 16 May 1983; Revised and Accepted 30 August 1983

ABSTRACT

The frequent detection of "patchy homology" between recombining DNAs in eukaryotic systems suggests that partial sequence homology might facilitate formation of partially base-paired hybrid structures and thus define a specific target for recombination. Indeed, the extent of such "patchy homology" initially appears impressive. The question of whether such homology is statistically significant, however, has not been addressed. In this paper we compare the extent of "patchy homology" within the sequences of SV40 recombination sites and within randomly-generated sequences with the same average GC content. We have found no statistically significant differences favoring the existence of "patchy homology" within the recombining regions of SV40. On the average 50% of the bases in randomly-generated sequences with the same GC content can be paired in a pattern of "patchy homology". We have also assessed the ability of sequences near recombination junctions in SV40 to form partially base-paired intra-strand secondary structures. Again, the ability of these sequences to form such configurations was unremarkable when compared with random sequences. Thus, the notion that partial base-pairing provides specificity for sites of recombination must be considered with caution.

INTRODUCTION

A number of studies have suggested that discontinuous or "patchy homology" defines a site of interaction between DNA's undergoing recombination in eukaryotic cells (1-6). Thus, the presence of such noncontiguous tracts of matching primary sequences in two otherwise heterologous regions of duplex DNA has been postulated to lead to partial base-pairing between parental molecules (1). Once such a partially base-paired hybrid intermediate is formed, nicking, excision, and repair presumably could generate the recombinant molecule. However, there have been relatively few examples of this type of interaction and the statistical significance of such "patchy homology" has not been clearly established. To pursue this question further we have focused upon a system well suited for assessing the significance of partial base pairing potential in recombination; namely, recombination of simian virus 40 (SV40) DNA in eukaryotic cells.

One of the difficulties in studying recombination in eukaryotes has been the difficulty of recovering the parental sequences involved in a given recombinational event. This is, of course, a consequence of the the size and complexity of the eukaryotic genome. An experimental approach to this problem has been provided by the selection of natural variants of SV40 that arise by recombination within and between viral genomes and between viral and cellular DNA (reviewed by Brockman, ref. 7). Because of the small size of the SV40 genome and its known DNA sequence (8) one is able to reconstruct the parental sequences involved in a given recombinational event by sequencing the region of recombination.

Gutai and Nathans (1, 2) and Singer and coworkers (9, 10) have taken this approach. By sequencing a number of recombination regions in natural SV40 variants, and comparing these with the reconstructed parental sequences, they found that recombination was not limited to a specific nucleotide sequence signal and did not require extensive sequence homology. The analysis of certain recombinant junctions led Gutai and Nathans (1, 2) to speculate that "AT richness" and "patchy homology" may be important features of some forms of SV40 recombination. Others also have found evidence of "patchy homology" in regions of recombination in other systems (4-6), although in most of these cases it has not been possible to reconstruct both the parental sequences. There have been enough reports of such "patchy homology" to conclude that it might be a widespread feature of recombination in eukaryotic organisms. However, to date it appears that no one has asked whether such "patchy homology" is specific to regions involved in recombination or whether a significant amount of such homology might be expected in random sequences. We have addressed this question by comparing six viral-viral recombinant joint sequences (Table 1) determined by Woodworth-Gutai (2) with computer-generated random sequences. We report that "patchy homology" appears to be as prevalent and extensive in random sequences as it is among the recombining sequences of SV40. In addition to examining the ability of sequences near recombination junctions to form partially base-paired hybrid DNA structures, we have assessed their ability to form partially paired intra-strand secondary structures. Again, the ability of these sequences to form such configurations is unremarkable when compared with random sequences. While these studies neither establish nor eliminate a role for "patchy homology" or intrastrand base pairing in recombination, they demonstrate that such partial base pairing potential does not define specific targets for recombination. Moreover they provide a measure by which to guage the significance of partial base-pairing

Table 1. Recombining sequences of SV40 (2).

Junction Number	Parental Sequences	%GC
ev-1114 CA	<u>ATACTTCTGCCTGCTGGGGAGCCTGGGGACTTTCACACCCTAACTGACA</u>	56
	<u>TATTTTCCATAATTTTCTTGATAGCAGTGCAGCTTTTTCTTTGTGCT</u>	32
ev-1114 CG	<u>GAGCTTTTGCTGCAATTTTGTGAAGGGGAAGATACTGTTGACGGGAAACG</u>	46
	<u>CTGTTTACCACCACTTCTGGAACACAGCAGTGAAGGGACTTCCCAGATA</u>	50
ev-1117 AC	<u>ATTTTAAAGTGTATAATGTGTTAACTACTGATTCTAATTGTTTGTGTATT</u>	22
	<u>TTAGGGTGTGGAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGC</u>	56
ev-1117 AB	<u>CAGAAGAAATGCCATCTAGTGATGATGAGGCTACTGCTGACTCTCAACAT</u>	44
	<u>CTGAACCTGAAACATAAAATGAATGCAATTGTTGTTTAACTTGTTTAT</u>	28
ev-1119 AD	<u>CTCTGATGAGAAAGGCATATTTAAAAAATGCAAGGAGTTTCATCCTGAT</u>	34
	<u>CTTCAATTGCAGCAGCGGCTCTCCAGCAGCAATTCAGCTACTGAAAAT</u>	48
ev-1119 AB	<u>CTTCGATTGCTTTAGAATGTGTTTGGACTTGATCTTTGTGAAGGAACCT</u>	40
	<u>GAATAATTCAAAGTGGCATTGCTTTGCTTCTTATGTTAATTGGTACAGA</u>	32

potential for heterologous DNA's of various G and C contents.

METHODS

Several nucleic acid folding algorithms were examined for their ability to determine the lowest energy state for molecules with weak pairing. The best algorithm under these conditions on our computer was that of Zuker and Stiegler (11). The states with lowest energy determined by another popular algorithm, that of Nissinov and Jacobson (12), were higher than the corresponding determination by the algorithm of Zuker and Stiegler (11). The reasons for this are not known, but they may be due to problems adapting the Nissinov-Jacobson algorithm to our computer and not within the basic algorithm itself (Nissinov, personal communication). Thus, for the work reported in this paper we have used the algorithm of Zuker and Stiegler (11) run on an AMDAHL 5860 computer.

The average GC content for the set of recombining sequences in SV40 (Table 1) is $40 \pm 11\%$. Random sequences of the same length (50 nucleotides) and average GC content ($40 \pm 8\%$) were generated by computer. The GC content

Figure 1. Types of interaction between single strands of DNA. Type I, interaction through homologous sequences, is the basis for the conventional model of recombination. A modification of this interaction, Type II, allows intra-strand pairing. Type III, interaction through secondary structure, is a less orthodox mechanism.

for individual random sequences ranged from 30% to 58%.

RESULTS

Types of Partial Base-Pairing that Might Be Involved in Recombination

To assess the role of partial base pairing in DNA recombination we have considered three types of DNA-DNA interaction (Figure 1):

Type I - Sequence homology was sought between those regions of the parental DNAs in which recombination occurred. This type of interaction provides the basis for the conventional model of homologous recombination (13) and is the type of interaction envisioned when the presence of "patchy homology" at recombination joints had been noted previously (see above).

Type II - This is a modification of the Type I interaction in which pairing within strands is permitted. Such pairing might permit the formation of more stable configurations.

Type III - Here we consider the possibility that individual strands of DNA might form relatively stable secondary structures that are detected by the mechanism of recombination, and that recombination occurs when two such structures of the appropriate form are juxtaposed. In such a mechanism no base pairing between recombining sequences would be required.

Utilizing each of these types of partial base-pairing configuration, the folding pattern of recombining sequences from SV40 was compared with that of computer-generated random sequences having the same average GC content (40%).

Figure 2. Patterns of "partial base-pairing" in sample sequences from the recombining SV40 sequences and randomly generated sequences. (a) Sequences from junction number ev-1119 AD and the pattern of "patchy homology" reported by Woodworth-Gutai (2). (b), (d) and (f) The same sequences folded according to Type I, Type II and Type III interactions. (c), (e) and (g) A sample random sequence folded according to Type I, Type II and Type III interactions.

The following criteria were used for these comparisons.

1. Minimum free-energy for the folded structure.
2. Total number of base pairs between (and within) strands.
3. Total GC and total AT base pairs.
4. Largest block of contiguous base pairs between (and within) strands.
5. Average size block of contiguous base pairs between (and within) strands.
6. Total number of mismatched regions.

Table 2. Comparison of the recombining sequences of SV40 with randomly-generated sequences assuming formation of partially base-paired hybrid intermediates and no pairing within strands (Type I).

Criterion	Mean		t value	p value
	SV40 (n=6)	Random (n=10)		
<u>Minimum Energies</u>	-12.10 $\pm$ 6.30	-18.99 $\pm$ 5.30	2.35	p > 0.02
<u>Homologies</u>				
Total no. of BP	20.00 $\pm$ 6.16	24.60 $\pm$ 4.67	1.68	p > 0.10
No. of GC pairs	10.17 $\pm$ 3.76	12.50 $\pm$ 3.37	1.31	p > 0.10
No. of AT pairs	8.33 $\pm$ 5.16	12.10 $\pm$ 3.90	1.85	p > 0.05
Longest block of homology	4.33 $\pm$ 0.82	5.40 $\pm$ 2.53	2.23	p > 0.02
Avg. sized block of homology	2.43 $\pm$ 0.54	2.48 $\pm$ 0.39	0.22	p > 0.10
<u>Mismatches</u>				
No. of mismatches	11.83 $\pm$ 3.60	14.40 $\pm$ 4.10	1.25	p > 0.10
Size of largest mismatch	8.17 $\pm$ 4.00	7.60 $\pm$ 2.50	0.25	p > 0.10

7. Size of the largest mismatch region.
8. Total number of stem-loop structures.
9. Size of the largest stem structure.
10. Size of the largest loop structure.

Analysis of Potential for Base-Pairing Between Parental DNAs in the Recombining Regions (Type I Interaction)

A sample sequence and the regions of "patchy homology" identified by Woodworth-Gutai are shown in Figure 2a. We have taken these same sequences and used the Zuker-Stiegler algorithm to determine the minimum energy configuration involving base pairs between strands (Figure 2b shows the configuration for the sample sequence in Figure 2a). Clearly this algorithm reveals more "patchy homology" than Woodworth-Gutai originally presented, but is such homology important in determining targets for recombination?

To address this question, computer-generated random sequences were examined by means of the Zuker-Stiegler algorithm. A typical base-pairing configuration is shown in Figure 2c. These structures have been compared with those for the recombining regions of SV40, and by a number of different criteria there are no statistically significant differences favoring the formation of "patchy homology" within the recombining regions of SV40 (see Table 2).

Figure 3. Percentage of nucleotides found in a pattern of "patchy homology" between randomly-generated parental sequences as a function of the average GC content of the sequences. Each point is represented by the mean and standard deviation for 10 pairs of parental sequences.

As a further assessment of the prevalence of "patchy homology" in random DNA sequences, we have examined the effect of GC composition on the extent of homology and the size distribution of blocks of homology for random sequences with a GC content of 40%. As noted in Table 2, 50% of the bases in random sequences are homologous when the GC content is 40%. As one might expect, the extent of homology increases as the GC content becomes extreme (Figure 3). In addition, we find that the "U"-shaped curve is not symmetrical but shifted to lower GC values because of the thermodynamic instability of short regions of

Figure 4. Size distribution of blocks of homology for random sequences with an average GC content of 40%.

Table 3. Comparison of the recombining sequences of SV40 with randomly-generated sequences assuming formation of partially base-pairing hybrid intermediates and pairing within strands (Type II).

Criterion	Mean		t value	p value
	SV40 (n=6)	Random (n=10)		
<u>Minimum Energies</u>	-18.73 $\pm$ 4.59	-20.97 $\pm$ 4.56	0.88	p > 0.1
<u>Homologies</u>				
Total no. of BP	25.17 $\pm$ 4.40	25.20 $\pm$ 4.39	0.01	p > 0.1
No. of GC pairs	12.17 $\pm$ 1.72	12.50 $\pm$ 2.64	0.22	p > 0.1
No. of AT pairs	13.00 $\pm$ 3.46	12.50 $\pm$ 4.22	0.24	p > 0.1
Longest block of homology	5.00 $\pm$ 1.26	5.10 $\pm$ 0.74	0.17	p > 0.1
Avg. sized block of homology	2.95 $\pm$ 0.37	2.70 $\pm$ 0.65	0.51	p > 0.1
<u>Mismatches</u>				
No. of mismatches	13.67 $\pm$ 2.88	14.20 $\pm$ 3.55	1.27	p > 0.1
Size of largest mismatch	7.83 $\pm$ 1.47	6.80 $\pm$ 4.02	0.60	p > 0.1
<u>Hairpins</u>				
No. of hairpins	1.83 $\pm$ 0.40	1.40 $\pm$ 0.70	1.36	p > 0.1
Size of loop	7.83 $\pm$ 3.76	10.00 $\pm$ 6.13	0.78	p > 0.1
Size of stem	3.17 $\pm$ 0.75	2.99 $\pm$ 1.73	0.24	p > 0.1

AT homology. This also is evident in the size distributions of blocks of homology for random sequences with GC contents of 10% and 90% (data not shown). Small block sizes are much less frequent in the low (10%) GC than in the high (90%) GC sequences; larger block sizes, however, occur at comparable frequencies in both low and high GC sequences.

The size distribution of blocks of homology for random sequences with a GC content of 40%, corresponding to the average GC content of the recombining sequences of SV40, is shown in Figure 4. Approximately 21% of the blocks of homology have a size of four or greater.

Analysis of Potential for Base-Pairing Between Parental DNAs in the Recombining Regions When Pairing Within Strands is Allowed (Type II Interaction)

In some sequences it was noted that more stable configurations were possible if base pairing within strands was permitted. Under these conditions the minimal energy configurations were determined for both the SV40 and random sequences. The sample configurations for each type are shown in Figures 2d

Table 4. Comparison of the recombining sequences of SV40 with randomly-generated sequences assuming interaction through secondary structure (Type III).

Criterion	Mean		t value	p value
	SV40 (n=6)	Random (n=10)		
<u>Minimum Energies</u>	-6.06 $\pm$ 2.60	-6.62 $\pm$ 1.90	0.45	p > 0.10
<u>Homologies</u>				
Total no. of BP	8.75 $\pm$ 2.04	10.65 $\pm$ 1.86	1.90	p > 0.05
No. of GC pairs	4.75 $\pm$ 1.33	5.80 $\pm$ 1.46	1.43	p > 0.10
No. of AT pairs	4.00 $\pm$ 1.90	4.80 $\pm$ 1.60	1.74	p > 0.10
Longest block of homology	4.00 $\pm$ 0.89	3.95 $\pm$ 1.00	0.10	p > 0.10
Avg. sized block of homology	3.01 $\pm$ 0.52	2.61 $\pm$ 0.56	0.28	p > 0.10
<u>Mismatches</u>				
No. of mismatches	5.67 $\pm$ 1.25	7.25 $\pm$ 1.43	2.23	p > 0.02
Size of largest mismatch	5.50 $\pm$ 1.26	5.45 $\pm$ 2.19	0.05	p > 0.10
<u>Hairpins</u>				
No. of hairpins	1.00 $\pm$ 0.00	1.20 $\pm$ 0.42	1.15	p > 0.10
Size of loop	8.33 $\pm$ 3.27	9.00 $\pm$ 2.67	0.45	p > 0.10
Size of stem	3.67 $\pm$ 0.82	2.20 $\pm$ 1.03	2.96	p > 0.01

and 2e, respectively. As can be seen from Table 3, we have detected no statistically significant differences favoring the formation of "patchy homology" within the recombining sequences of SV40 over that within random sequences.

Analysis of Secondary Structure Within Individual Strands (Type III Interaction)

In order to test the possibility that this type of interaction might facilitate recombination we have examined the minimum energy configurations under conditions in which base-pairing occurs only within single strands of the recombining sequences. The sample structures for the recombining sequences of SV40 and for random sequences are shown in Figures 2f and 2g, respectively. As can be seen in Table 4, no statistically significant differences favoring the formation of these configurations over those obtained for random sequences were found.

DISCUSSION

In the previous section we considered three different hypotheses

concerning the way in which single stranded regions might interact during the recombination process. The aim was to determine whether partial base-pairing potential provides specific targets for recombination that could be recognized among known recombination sites in SV40. According to all of the hypotheses tested, and according to all of the criteria used, there are no statistically significant differences favoring the formation of partially base-paired hybrid DNA structures and/or partially base-paired intrastrand secondary structures within the recombining sequences of SV40 over that within random sequences.

When we considered interactions based on the conventional model of recombination (Type I), we found that on the average 25 of 50 bases in each of two randomly generated sequences with an average GC content of 40% can be paired in a pattern of "patchy homology" (Tables 2 and 3). An even larger number of bases become involved in such a pattern if the nucleotide sequences become more GC or (significantly more) AT rich (Figure 3). The size distribution of the blocks of contiguous base pairs found in these patterns of "patchy homology" (Figure 4) shows that 21% of the blocks have a size of four or more.

Since no statistically significant differences were found between the recombining sequences of SV40 and randomly-generated sequences when a conventional (Type I) interaction was examined, we then considered more unorthodox types of interaction between the parental strands forming the recombinant junctions (Types II, and III). Again, no statistically significant differences were found between the recombining sequences of SV40 and the randomly-generated sequences.

There still might be specific targets for recombination involving partial base-pairing in SV40, but the features would have to be more subtle than those examined above. Several possibilities might be suggested: (a) The patchy homology observed in a β -globin recombinational event (5) suggests a progressive loss of register between the regions of "patchy homology". The significance of this pattern is difficult to determine, since it represents an isolated example and since the folding algorithm finds even more "patchy homology" with a free energy that is not much different from that of the mean for random sequences of the same size and average GC content (data not shown). Furthermore, under these conditions, the progressive pattern noted above is lost. (b) A transient or metastable (local minimum free energy) structure could provide the target for recombination. This is a difficult hypothesis to examine because of the large number of possibilities to test. To approach this problem, we have used an algorithm that first finds the largest block of

homology, then after "fixing" that region finds the next largest block of homology, etc. (unpublished). To a first approximation such algorithms tend to find the more stable transient structures, and although they often produce the minimum free-energy structures, they also produce metastable structures. In our comparisons of recombining SV40 sequences and random sequences by means of this and a number of similar algorithms we were unable to detect any statistically significant differences (data not shown). (c) "Patchy homology" might facilitate pairing when one parental DNA is in the inverse orientation to that depicted in Figure 1 (i.e., A-M aligned with Z-N rather than with N-Z). It is not readily apparent how a structure displaying this type of pairing would be resolved into the final recombinants. Nevertheless, it is possible that such pairing might be involved in the initiation of recombination. We, therefore analyzed the complementary strands from the recombining regions of SV40 using the same criteria used for the Type I, II and III configurations (data not shown). Again, there were no statistically significant differences favoring the formation of such partially base-paired hybrid DNA's over those obtained for random sequences. (d) The minimum free-energy structures involving naked DNA might differ significantly from those formed in the presence of recombinational or other DNA-binding proteins. However, the implications of this hypothesis are difficult to determine. At any rate, in contrast to known targets for recombination (14), or other DNA binding proteins (15), we must conclude that putative targets for SV40 recombination must be more subtle and that substantially new ideas will be required to pursue the problem further.

On the other hand, one might conclude that no specific target is required for recombination in SV40. In regard to the evolution of variant SV40 DNA molecules this conclusion seems likely based upon the results presented in this paper. It also is supported by recent evidence suggesting a nonspecific mechanism of "blunt-end" recombination in SV40 (16). While it is possible that insertion of SV40 DNA into cellular chromosomes may proceed by a mechanism distinct from the type of recombination analyzed in this paper, the analysis of integrated SV40 DNAs suggests that homology also plays little role in the integration process; when cells with defective SV40 DNA previously integrated into the cellular genome are superinfected with SV40, the integration of the superinfecting DNA does not occur within the homologous resident SV40 sequences (17-19). Examination of a host-viral junction at the site of SV40 integration in chromosomal DNA, however, has revealed a perfect match between the parental DNAs of five base pairs precisely at the crossover

point (20). In contrast, only one (evl119AB) of the six viral-viral junctions we analyzed has more than a single nucleotide common to both parental DNAs at the recombination junction. At any rate, the suggestion that regions of partial base pairing, of the types examined here, define specific targets for recombination should be avoided unless it can be demonstrated that the potential for such pairing differs considerably from that found in random sequences with the same average GC content. This is not to say that partially base-paired hybrid DNA structures or partially base-paired intra-strand secondary structures are unimportant in recombination; rather, that such types of patchy pairing do not appear to provide a basis for specificity since they occur frequently when any two random DNA sequences are compared with each other.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the National Institutes of Health to WWB (CA-19816) and MAS (GM-30054).

REFERENCES

1. Gutai, M.W. and Nathans, D. (1978) J. Mol. Biol. 126:275-288.
2. Woodworth-Gutai, M. (1981a) Virology 109:344-352.
3. Woodworth-Gutai, M. (1981b) Virology 109:353-365.
4. Botchan, M., Stringer, J., Mitchison, T., and Sambrook, J. (1980) Cell 20:143-152.
5. Jahn, C.L., Hutchison, C.A., III, Phillips, S.J., Weaver, S., Haigwood, N.L., Voliva, C.F., and Edgell, M.H. (1980) Cell 21:159-168.
6. Stringer, J.R. (1981) J. Virology 38:671-679.
7. Brockman, W.W. (1977) Prog. Med. Virol. 23:69-95.
8. Buchman, A.R., Burnett, L., and Berg, P. (1980) in DNA Tumor Viruses, Tooze, J., ed., pp. 799-817, Cold Spring Harbor Laboratory, N.Y.
9. Wakamiya, T., McCutchan, T., Rosenberg, M., and Singer, M. (1979) J. Biol. Chem. 254:3584-3591.
10. McCutchan, T., Singer, M., and Rosenberg, M. (1979) J. Biol. Chem. 254:3592-3597.
11. Zuker, M. and Stiegler, P. (1981) Nucleic Acids Research 9:133-149.
12. Nussinov, R. and Jacobson, A.B. (1980) Proc. Natl. Acad. Sci. USA 77:6309-6313.
13. Holliday, R. (1964) Genet. Res. 5:282-304.
14. Nash, H. (1981) Ann. Rev. Genetics 15:143-167.
15. Johnson, A.D., Poteete, A.R., Lauer, G., Sauer, R.T., Ackers, G.K., and Ptashne, M. (1981) Nature (London) 294:217-223.
16. Wilson, J.H., Berget, P.B., and Pipas, J.M. (1982) Mol. Cellular Biol. 2:1258-1269.
17. Graessmann, A., Graessman, M., Topp, W.C. and Botchan, M. (1979). J. Virol. 32:989-994.
18. Botchan, M.R., Topp, W.C., and Sambrook, J. (1979) Cold Spring Harbor Symp. Quant. Biol. 43:709-719.
19. Bender, M.A., Christensen, J., and Brockman, W.W. (1983) J. Virol. 47:115-124.
20. Stringer, J.R. (1982) Nature (London) 296:363-366.