PROGRESS IN

Nucleic Acid Research and Molecular Biology

Volume 25

edited by

WALDO E. COHN

PROGRESS IN
Nucleic Acid Research and Molecular Biology

edited by

WALDO E. COHN

Biology Division
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Volume 25

1981

ACADEMIC PRESS

A Subsidiary of Harcourt Brace Jovanovich, Publishers

New York London Toronto Sydney San Francisco

ACADEMIC PRESS, INC.
111 Fifth Avenue, New York, New York 10003

United Kingdom Edition published by
ACADEMIC PRESS, INC. (LONDON) LTD.
24/28 Oval Road, London NW1 7DX

LIBRARY OF CONGRESS CATALOG CARD NUMBER: 63–15847

ISBN 0–12–540025–X

PRINTED IN THE UNITED STATES OF AMERICA

81 82 83 84 9 8 7 6 5 4 3 2 1

List of Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

YOSEF ALONI (1), *Department of Genetics, Weizmann Institute of Science, Rehovot, Israel*

W. FRENCH ANDERSON (127), *Laboratory of Molecular Hematology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20205*

GOKUL C. DAS* (187), *The University of Tennessee–Oak Ridge Graduate School of Biomedical Sciences and Biology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830*

JOSEPH FRIEDMAN† (33), *Department of Cellular Biochemistry, The Hebrew University Hadassah Medical School, Jerusalem, Israel*

JENNIFER D. HALL (53), *Department of Cellular and Developmental Biology, College of Liberal Arts, University of Arizona, Tucson, Arizona 85721*

ROSEMARY JAGUS (127), *Laboratory of Molecular Hematology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20205*

DAVID W. MOUNT (53), *Department of Microbiology, College of Medicine, University of Arizona, Tucson, Arizona 85721*

SALIL K. NIYOGI (187), *The University of Tennessee–Oak Ridge Graduate School of Biomedical Sciences and Biology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830*

AHARON RAZIN (33), *Department of Cellular Biochemistry, The Hebrew University Hadassah Medical School, Jerusalem, Israel*

BRIAN SAFER (127), *Laboratory of Molecular Hematology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland 20205*

* Present address: Laboratory of Biology of Viruses, NIAID, National Institutes of Health, Bethesda, Maryland 20014.

† Present address: Biology Division, Oak Ridge National Laboratory, Box Y, Oak Ridge, Tennessee 37830.

Abbreviations and Symbols

All contributors to this Series are asked to use the terminology (abbreviations and symbols) recommended by the IUPAC-IUB Commission on Biochemical Nomenclature (CBN) and approved by IUPAC and IUB, and the Editor endeavors to assure conformity. These Recommendations have been published in many journals (*1*, *2*) and compendia (*3*) in four languages and are available in reprint form from the Office of Biochemical Nomenclature (OBN), as stated in each publication, and are therefore considered to be generally known. Those used in nucleic acid work, originally set out in section 5 of the first Recommendations (*1*) and subsequently revised and expanded (*2*, *3*), are given in condensed form (I–V) below for the convenience of the reader. Authors may use them without definition, when necessary.

I. Bases, Nucleosides, Mononucleotides

1. *Bases* (in tables, figures, equations, or chromatograms) are symbolized by Ade, Gua, Hyp, Xan, Cyt, Thy, Oro, Ura; Pur = any purine, Pyr = any pyrimidine, Base = any base. The prefixes S–, H_2, F–, Br, Me, etc., may be used for modifications of these.

2. *Ribonucleosides* (in tables, figures, equations, or chromatograms) are symbolized, in the same order, by Ado, Guo, Ino, Xao, Cyd, Thd, Ord, Urd (Ψrd), Puo, Pyd, Nuc. Modifications may be expressed as indicated in (1) above. Sugar residues may be specified by the prefixes r (optional), d (=deoxyribo), a, x, l, etc., to these, or by two three-letter symbols, as in Ara-Cyt (for aCyd) or dRib-Ade (for dAdo).

3. *Mono-, di-, and triphosphates of nucleosides* (5′) are designated by NMP, NDP, NTP. The N (for "nucleoside") may be replaced by any one of the nucleoside symbols given in II-1 below. 2′-, 3′-, and 5′- are used as prefixes when necessary. The prefix d signifies "deoxy." [Alternatively, nucleotides may be expressed by attaching P to the symbols in (2) above. Thus: P-Ado = AMP; Ado-P = 3′-AMP] cNMP = cyclic 3′:5′-NMP; Bt_2cAMP = dibutyryl cAMP, etc.

II. Oligonucleotides and Polynucleotides

1. Ribonucleoside Residues

(a) Common: A, G, I, X, C, T, O, U, Ψ, R, Y, N (in the order of I-2 above).

(b) Base-modified: sI or M for thioinosine = 6-mercaptopurine ribonucleoside; sU or S for thiouridine; brU or B for 5-bromouridine; hU or D for 5,6-dihydrouridine; i for isopentenyl; f for formyl. Other modifications are similarly indicated by appropriate *lower-case* prefixes (in contrast to I-1 above) (*2*, *3*).

(c) Sugar-modified: prefixes are d, a, x, or l as in I-2 above; alternatively, by ***italics*** or boldface type (with definition) unless the entire chain is specified by an appropriate prefix. The 2′-*O*-methyl group is indicated by *suffix* m (e.g., -Am- for 2′-*O*-methyladenosine, but -mA- for 6-methyladenosine).

(d) Locants and multipliers, when necessary, are indicated by superscripts and subscripts, respectively, e.g., -m_2^6A- = 6-dimethyladenosine; -s^4U- or -^{4}S- = 4-thiouridine; -ac^{4}Cm- = 2′-*O*-methyl-4-acetylcytidine.

(e) When space is limited, as in two-dimensional arrays or in aligning homologous sequences, the prefixes may be placed *over the capital letter*, the suffixes *over the phosphodiester symbol*.

2. Phosphoric Residues [left side = 5′, right side = 3′ (or 2′)]

(a) Terminal: p; e.g., pppN . . . is a polynucleotide with a 5′-triphosphate at one end; Ap is adenosine 3′-phosphate; C > p is cytidine 2′:3′-cyclic phosphate (*1*, *2*, *3*); p < A is adenosine 3′:5′-cyclic phosphate.

(b) Internal: hyphen (for known sequence), comma (for unknown sequence); unknown sequences are enclosed in parentheses. E.g., pA-G-A-C(C_2,A,U)A-U-G-C > p is a sequence with a (5′) phosphate at one end, a 2′:3′-cyclic phosphate at the other, and a tetranucleotide of unknown sequence in the middle. **(Only codon triplets should be written without some punctuation separating the residues.)**

3. Polarity, or Direction of Chain

The symbol for the phosphodiester group (whether hyphen or comma or parentheses, as in 2b) represents a 3′-5′ link (i.e., a 5′ . . . 3′ chain) unless otherwise indicated by appropriate numbers. "Reverse polarity" (a chain proceeding from a 3′ terminus at left to a 5′ terminus at right) may be shown by numerals or by right-to-left arrows. Polarity in any direction, as in a two-dimensional array, may be shown by appropriate rotation of the (capital) letters so that 5′ is at left, 3′ at right when the letter is viewed right-side-up.

4. Synthetic Polymers

The complete name or the appropriate group of symbols (see II-1 above) of the repeating unit, **enclosed in parentheses if complex or a symbol,** is either (a) preceded by "poly," or (b) followed by a subscript "n" or appropriate number. **No space follows "poly"** (*2, 5*).

The conventions of II-2b are used to specify known or unknown (random) sequence, e.g.,

polyadenylate = poly(A) or A_n, a simple homopolymer;

poly(3 adenylate, 2 cytidylate) = poly(A_3C_2) or $(A_3,C_2)_n$, an *irregular* copolymer of A and C in 3 : 2 proportions;

poly(deoxyadenylate-deoxythymidylate) = poly[d(A-T)] or poly(dA-dT) or $(dA\text{-}dT)_n$ or $d(A\text{-}T)_n$, an *alternating* copolymer of dA and dT;

poly(adenylate,guanylate,cytidylate,uridylate) = poly(A,G,C,U) or $(A,G,C,U)_n$, a random assortment of A, G, C, and U residues, proportions unspecified.

The prefix copoly or oligo may replace poly, if desired. The subscript "n" may be replaced by numerals indicating actual size, e.g., $A_n \cdot dT_{12\text{-}18}$.

III. Association of Polynucleotide Chains

1. *Associated* (e.g., H-bonded) chains, or bases within chains, are indicated by a *center dot* (not a hyphen or a plus sign) separating the *complete* names or symbols, e.g.:

poly(A) · poly(U) or $A_n \cdot U_m$
poly(A) · 2 poly(U) or $A_n \cdot 2U_m$
poly(dA-dC) · poly(dG-dT) or $(dA\text{-}dC)_n \cdot (dG\text{-}dT)_m$.

2. *Nonassociated* chains are separated by the plus sign, e.g.:

$2[\text{poly(A)} \cdot \text{poly(U)}] \rightarrow \text{poly(A)} \cdot 2\,\text{poly(U)} + \text{poly(A)}$
or $2[A_n \cdot U_m] \rightarrow A_n \cdot 2U_m + A_n$.

3. Unspecified or unknown association is expressed by a comma (again meaning "unknown") between the completely specified chains.

Note: In all cases, each chain is completely specified in one or the other of the two systems described in II-4 above.

IV. Natural Nucleic Acids

RNA	ribonucleic acid or ribonucleate
DNA	deoxyribonucleic acid or deoxyribonucleate
mRNA; rRNA; nRNA	messenger RNA; ribosomal RNA; nuclear RNA
hnRNA	heterogeneous nuclear RNA
D-RNA; cRNA	"DNA-like" RNA; complementary RNA

mtDNA	mitochondrial DNA
tRNA	transfer (or acceptor or amino-acid-accepting) RNA; replaces sRNA, which is not to be used for any purpose
aminoacyl-tRNA	"charged" tRNA (i.e., tRNA's carrying aminoacyl residues); may be abbreviated to AA-tRNA
alanine tRNA or $tRNA^{Ala}$, etc.	tRNA normally capable of accepting alanine, to form alanyl-tRNA, etc.
alanyl-tRNA or alanyl-$tRNA^{Ala}$	The same, with alanyl residue covalently attached. [*Note:* fMet = formylmethionyl; hence $tRNA^{fMet}$, identical with $tRNA_f^{Met}$]

Isoacceptors are indicated by appropriate subscripts, i.e., $tRNA_1^{Ala}$, $tRNA_2^{Ala}$, etc.

V. Miscellaneous Abbreviations

P_i, PP_i	inorganic orthophosphate, pyrophosphate
RNase, DNase	ribonuclease, deoxyribonuclease
t_m (not T_m)	melting temperature (°C)

Others listed in Table II of Reference 1 may also be used without definition. No others, with or without definition, are used unless, in the opinion of the editor, they increase the ease of reading.

Enzymes

In naming enzymes, the 1978 recommendations of the IUB Commission on Biochemical Nomenclature (*4*) are followed as far as possible. At first mention, each enzyme is described *either* by its systematic name *or* by the equation for the reaction catalyzed *or* by the recommended trivial name, followed by its EC number in parentheses. Thereafter, a trivial name may be used. Enzyme names are not to be abbreviated except when the substrate has an approved abbreviation (e.g., ATPase, but not LDH, is acceptable).

References*

1. *JBC* **241,** 527 (1966); *Bchem* **5,** 1445 (1966); *BJ* **101,** 1 (1966); *ABB* **115,** 1 (1966), **129,** 1 (1969); and elsewhere.†
2. *EJB* **15,** 203 (1970); *JBC* **245,** 5171 (1970); *JMB* **55,** 299 (1971); and elsewhere.†
3. "Handbook of Biochemistry" (G. Fasman, ed.), 3rd ed. Chemical Rubber Co., Cleveland, Ohio, 1970, 1975, Nucleic Acids, Vols. I and II, pp. 3–59.
4. "Enzyme Nomenclature" [Recommendations (1978) of the Nomenclature Committee of the IUB]. Academic Press, New York, 1979.
5. "Nomenclature of Synthetic Polypeptides," *JBC* **247,** 323 (1972); *Biopolymers* **11,** 321 (1972); and elsewhere.†

Abbreviations of Journal Titles

Journals	*Abbreviations used*
Annu. Rev. Biochem.	ARB
Arch. Biochem. Biophys.	ABB
Biochem. Biophys. Res. Commun.	BBRC

*Contractions for names of journals follow.

†Reprints of all CBN Recommendations are available from the Office of Biochemical Nomenclature (W. E. Cohn, Director), Biology Division, Oak Ridge National Laboratory, Box Y, Oak Ridge, Tennessee 37830, USA.

Biochemistry	Bchem
Biochem. J.	BJ
Biochim. Biophys. Acta	BBA
Cold Spring Harbor Symp. Quant. Biol.	CSHSQB
Eur. J. Biochem.	EJB
Fed. Proc.	FP
Hoppe-Seyler's Z. physiol. Chem.	ZpChem
J. Amer. Chem. Soc.	JACS
J. Bacteriol.	J. Bact.
J. Biol. Chem.	JBC
J. Chem. Soc.	JCS
J. Mol. Biol.	JMB
Nature, New Biology	Nature NB
Nucleic Acid Research	NARes
Proc. Nat. Acad. Sci. U.S.	PNAS
Proc. Soc. Exp. Biol. Med.	PSEBM
Progr. Nucl. Acid Res. Mol. Biol.	This Series

Some Articles Planned for Future Volumes

tRNA Splicing in Lower Eukaryotes
J. ABELSON AND G. KNAPP

Ribosomal RNA: Structure and Interactions with Proteins
R. BRIMACOMBE

Metabolism and Function of Cyclic Nucleotides
W. Y. CHEUNG

Accuracy of Protein Synthesis: A Reexamination of Specificity in Codon–Anticodon Interaction
H. GROSJEAN AND R. BUCKINGHAM

Mechanism of Interferon Action
G. SEN

The Regulatory Function of the 3'-Region of mRNA and Viral RNA Translation
U. LITTAUER AND H. SOREQ

Participation of Aminoacyl-tRNA Synthetases and tRNAs in Regulatory Processes
G. NASS

Queuine
S. NISHIMURA

Viral Inhibition of Host Protein Synthesis
A. SHATKIN

Ribosomal Proteins: Structure and Function
A. R. SUBRAMANIAN

RNA-Helix Destabilizing Proteins
W. SZER AND J. O. THOMAS

Contents

Splicing of Viral mRNAs

Yosef Aloni

DNA Methylation and Its Possible Biological Roles

Aharon Razin and Joseph Friedman

Mechanisms of DNA Replication and Mutagenesis in Ultraviolet-Irradiated Bacteria and Mammalian Cells

Jennifer D. Hall and David W. Mount

The Regulation of Initiation of Mammalian Protein Synthesis

Rosemary Jagus, W. French Anderson, and Brian Safer

Structure, Replication, and Transcription of the SV40 Genome

Gokul C. Das and Salil K. Niyogi

Splicing of Viral mRNAs

YOSEF ALONI

Department of Genetics,
Weizmann Institute of Science,
Rehovot, Israel

I. Introduction

In the prokaryotic cell, translation of the mRNA by ribosomes starts before transcription of the message from the DNA has been completed. In a eukaryotic cell, by contrast, the process of transcription occurs in the nucleus while translation takes place mainly, if not entirely, in the cytoplasm. The two operations are separated by the nuclear membrane, providing an obvious opportunity for additional processing and maturation of the primary RNA product. The various aspects of RNA processing are covered in comprehensive reviews (*1–12*).

In the last three years it has become apparent that understanding the process of RNA transcript maturation is one of the great frontiers in the field of eukaryotic cell physiology. The basis for this belief has been the discovery of a novel maturation mechanism

Progress in Nucleic Acid Research and Molecular Biology, Vol. 25

ISBN 0-12-540025-X

termed "RNA splicing," identified first in the DNA and RNA tumor viruses (*13–28*) and then extended to many eukaryotic organisms (*29–32*). The principle of RNA splicing is that, during processing of a primary RNA transcript internal sequences in the molecule are excised, and the remaining pieces are ligated to give rise to a continuous, translatable RNA molecule. This mechanism implies that the gene itself is split, in the sense that it contains sequences that are ultimately translated, and sequences that are removed and not translated. Gilbert (*33*) has referred to the intervening sequences, those sequences on the DNA that do not end up in the translatable mRNA, as "introns"; the expressed sequences, as "exons."

The purpose of this review is mainly to describe studies on the identification of RNA splicing in the biological systems of DNA and RNA viruses, and the first attempts to determine the mechanism of this process. Reviews on split genes and RNA splicing have been published elsewhere (*12, 33–35*).

II. SV40 as a Model System[1]

The use of animal viruses as model systems for probing the complexities of molecular control mechanisms has been particularly fruitful. It is generally felt that an understanding of genetic regulation in viruses will provide insight into similar regulatory processes in eukaryotic cells. The molecular biology of SV40 has been under intensive investigation for a number of years, and these studies have provided considerable information regarding the regulation of gene expression, in particular, transcriptional and posttranscriptional processing of mRNA (*36–39*).

SV40 provides several unique advantages as a model system for such studies. They include the following.

1. The viral genome is a small circular molecule (molecular weight 3.4×10^6) that contains genetic information for only five or six proteins. The DNA can be obtained in large quantities, which is imperative for many experiments in molecular biology.
2. The same RNA polymerase (polymerase II) transcribes both viral and cellular RNA.
3. The viral and cellular RNAs undergo similar posttranscriptional modifications (e.g., polyadenylylation at the 3′ terminus, capping at the 5′ terminus, and internal methylation).
4. A number of mutants and hybrid viruses are available for study.

[1] See also Das and Niyogi in this volume. [Ed.]

5. Transcriptional complexes are easy to obtain.
6. The entire nucleotide sequence of this virus has been determined.

The SV40 genome is comprised of early and late genes that are localized in symmetrical halves of the viral DNA (*40*). The segment between 0.67 and 0.17 on the map is transcribed in a counterclockwise direction prior to the onset of viral DNA replication and codes for the "early" viral proteins. The second segment (from 0.67 to 0.17) is transcribed in abundance after initiation of viral DNA replication in a clockwise direction. It encodes the information for the late proteins: VP_1, VP_2, and VP_3. The capsid proteins have been mapped approximately between 0.95 and 0.16, 0.76 and 0.97, and

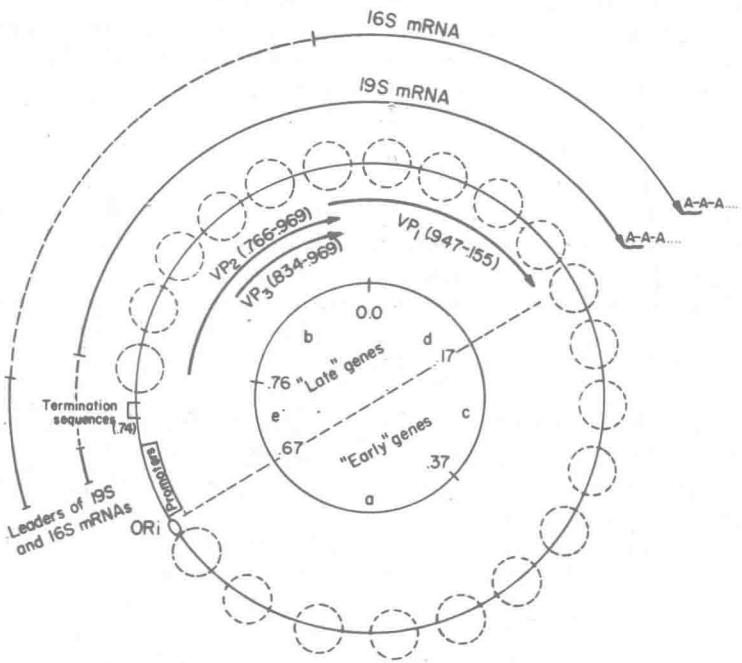

FIG. 1. Transcriptional map of SV40 indicating the five *Eco*RI *Hpa*I *Bgl*I restriction fragments of SV40 genome in the central circle. Arrows on RNAs indicate 3′ termini, poly(A) tails, and direction of transcription. Decimal numbers represent map units of the SV40 genome. Dashed lines (- - -) indicate sequences spliced out of 16 S and 19 S mRNAs. The location of the coding regions are indicated by heavy lines. The small dashed circles denote the distribution of nucleosomes about the SV40 minichromosome with an "exposed" region that contains the origin of replication (ORi) promoters for late transcription and transcription termination sequences (*38*). Only the major leaders of 19 S and 16 S mRNAs are represented; for more details, see ref. 38.

0.83 and 0.97, respectively (see Fig. 1) (*36–38*). The late viral mRNAs are known as "16 S" and "19 S." The 16 S RNA codes for VP_1 while the 19 S RNAs code for VP_2 and VP_3 (*38*). The viral DNA segment between 0.67 and 0.76 is transcribed late in infection, but the genetic information encoded in this region is unknown. Our studies have been concerned with the expression of the "late" genes. Reviews concerning the "early" genes have been published (*36–39*).

III. The Initiation of Transcription of SV40 DNA Late after Infection

The excitement that arose after the observation of RNA splicing delayed the investigations of other mechanisms involved in the regulation of SV40 gene expression. Among them are (*a*) the controls that operate at the initial steps of transcription and determine the specificity of transcriptional initiation; (*b*) the frequency of completion of primary transcripts; and (*c*) the mechanism of strand selection. I shall describe first our approaches to determine the localization of the initiation sites of SV40 late transcription, and the possible mechanisms involved in determining this specificity.

The best approach to determine the initiation site for transcription is by determining which nucleotides are at the 5′ end of the newly synthesized RNA. However, we and others (*41*) have failed to detect any labeled 5′ termini of SV40-specific RNAs. We therefore used three independent approaches, which were undertaken in order to localize the initiation site for transcription of SV40 DNA at late time after infection (*42*). Two of these were based on the Dintzis principles (*43*), while in the third we measured nascent RNA chains attached to transcriptional complexes under the electron microscope.

The rationale of localization of the initiation site(s) based on the Dintzis principles is that, after short pulses with radioactive precursors, RNA molecules would contain some labeled sequences complementary to each region of the DNA, but the labeled RNA complementary to a fragment of DNA that includes the initiation site for transcription would be in the shortest chains, while labeled RNA complementary to a DNA fragment far from the initiation site would be in successively longer chains.

RNA labeled *in vivo* was isolated from productively infected cells, and RNA labeled *in vitro* was isolated from transcriptional complexes of SV40. The purified RNAs were denatured and fractionated by sedimentation through sucrose gradients. Labeled RNAs

of various lengths were hybridized with restriction fragments of SV40 DNA of a known order. In both cases, the shortest RNAs hybridized with a fragment that spans between 0.67 and 0.76 on the map. The hybridization with this fragment decreased with successively longer RNAs, indicating that transcription initiates within this fragment or very close to it. Similar enrichment for this fragment was obtained using nascent RNA chains labeled *in vitro* with a short pulse. Electron microscope observation of transcriptional complexes of SV40 has revealed a substantial fraction with one short nascent RNA chain. The initiation site of the nascent chains was placed at coordinate 0.67 ± 0.02.

Based on these and other results in which we mapped the 5′ end of the nuclear viral RNA (*24*), we have concluded that late transcription initiates at alternative sites in a fragment of the genome that spans 0.67 to 0.76. This conclusion is supported by the localization on the map of the "caps" of the viral RNAs (*44*, *45*) and by the localization of the 5′ ends of poly(A)-containing viral RNA, using the primer extension technique (*38*).

The next question was: What determines the specificity of initiation? Is it only sequence-specific, or does the structure of the template contribute also to this specificity?

IV. The SV40 Minichromosome

SV40 DNA is found within infected cells in the form of a minichromosome (*46*), and a variety of methods have been developed for the extraction of viral chromatin from the infected cells (*47–49*). The SV40 minichromosome possesses a beaded structure composed of cellular histones and supercoiled viral DNA in a molecular complex very similar to that of cellular chromatin (*50–52*). The similarity between these structures, together with the fact that the major viral functions take place within cell nuclei via cellular machinery, have made SV40 an attractive system in which to study the organization and expression of the more complex eukaryotic chromatin. Early studies on SV40 chromatin, attempting to determine the precise distribution of the nucleosomes along the DNA, indicated that the nucleosomes were randomly distributed relative to the viral DNA sequences (*53–56*). Recent studies have altered this picture somewhat by indicating that a region close to the viral origin of replication is particularly sensitive to nuclease digestion (*57–60*). The simplest interpretation of the presence of a nuclease-sensitive region is that it contains a peculiar arrangement of protein structures about this region.

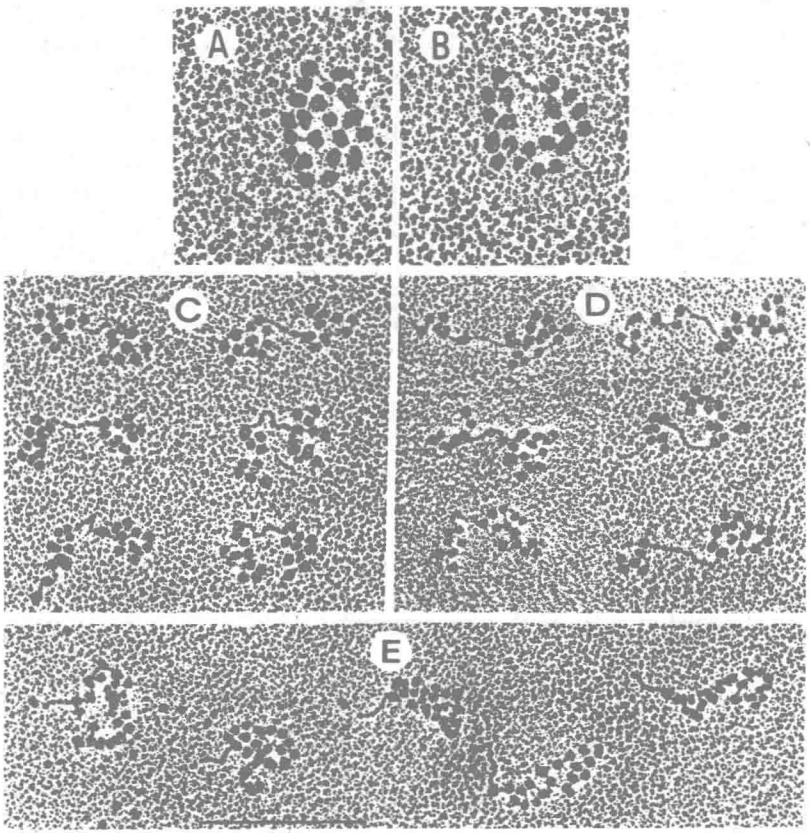

FIG. 2. Electron microscope visualization of SV40 minichromosomes. (A) Minichromosome without a gap. (B) Minichromosome with a gap. (C, D, and E) Minichromosomes with gaps cleaved with *Bam*HI (0.15 map unit), *Eco*RI (0.0 map unit), and *Bgl*I (0.67 map unit), respectively.

We have analyzed SV40 minichromosomes in the electron microscope in an attempt to determine whether alteration in the gross nucleoprotein structure of this region could be visualized. About 25% of the SV40 minichromosomes observed contain a region of DNA between 0.67 and 0.75 on the map that is not organized into the typical nucleosome beaded structure (see Fig. 2) (*61*). This is the same region where the initiation of late transcription occurs (*42*) (see Fig. 1).

It was therefore of interest to investigate whether there is a cor-