Proc. Natl. Acad. Sci. USA
Vol. 80, pp. 2902–2906, May 1983
Biochemistry

Gene and protein sequences of adenovirus protein VII, a hybrid basic chromosomal protein

(histones and protamine/chromatin structure)

MICHAEL T. SUNG, TIN M. CAO, ROGER T. COLEMAN, AND KIM A. BUDELIER

Department of Chemistry and Biochemistry, Southern Illinois University, Carbondale, Illinois 62901

Communicated by J. Herbert Taylor, February 22, 1983

ABSTRACT **The sequences of both the gene and the corresponding protein of adenovirus major core protein VII have been determined. The precise location of this gene is between 43.37 and 44.90 map coordinates on the viral genome. Protein VII is 173 residues long and has a molecular weight of 19,258. Detailed analysis of its sequence has revealed four basic domains separated by several predicted α helices. It is proposed that intrachain folding of protein VII is driven by hydrophobic interactions of the α helices, leaving the basic domains of the protein to interact with DNA phosphates. Protein monomers may further associate with each other in the formation of hexameric nucleosome-like particles. The displacement and replacement of protein VII during the viral infectious cycle in the host cell appears to mimic the biology of nucleoprotamine during the processes of spermatogenesis and fertilization. The presence of a protamine-like domain affirms a hybrid histone/protamine molecular structure for protein VII, although it may resemble the protamine in function.**

DNA is associated with various classes of basic proteins in the nuclei of eukaryotic cells. These proteins provide structural stability as well as functional organization of the DNA into chromosomes. As an example of the different and specific roles of the nuclear proteins, full complements of somatic histones in genetically active germ-line cells are progressively replaced by the arginine-rich protamine of mature sperm (1). During the process of fertilization, protamine is removed, histones regain their associations with the DNA, and the compact chromatin of the sperm decondenses. The "life cycles" of the two classes of basic proteins in relation to the life cycle of sperm cells indicate their distinctive interactions with DNA. Among the histones, additional distinctions can be made for their role in the structure and expression of genomic DNA.

The core histones, H2a, H2b, H3, and H4, are functionally different from the spacer histone, H1 (2). Because the charged residues in the histones are asymmetrically distributed, it is generally believed that the carboxyl ends of the core histones interact to form a globular unit. These globular units on the various histones interact to form nucleosomal particles, whereas the basic domains of the histones form ionic bonds with 145 base pairs of DNA (3). By contrast, histone H1 is lysine-rich, binds to the 50 base pairs of spacer DNA, and is involved in the higher order of chromatin structure (4). During avian erythropoiesis, an additional spacer histone, H5, is added to the repertoire of chromosomal basic proteins. It is believed that H5 is responsible for the genetic inactivity of the avian erythrocyte chromatin (5). Clearly, the basic proteins exert coarse control on genetic activity and are therefore important in the growth and development of cells.

Morphogenesis of adenovirus type 2 in human cells produces yet another class of basic nuclear proteins. Late in virus infection, two virus-encoded basic proteins are synthesized. There are 1,080 copies of core protein VII (M_r = 18,000) and 180 copies of the minor core protein V (M_r = 45,000) involved in the formation of the compact virus core (6). Micrococcal nuclease digestion of the virus core has resulted in a 150-base-pair subunit DNA fragment (7), but no nucleosome repeat pattern has been observed. As early as 3 hr after infection, intranuclear adenovirus DNA assumes a nucleosomal repeat pattern similar to that of cellular chromatin (8–11). This suggests that, during virus uncoating to produce transcriptionally active viral DNA, the adenovirus 2 specific basic proteins VII and V are replaced by histones. Viral DNA synthesis, which begins at 6 hr after infection, initiates the transcription of late genes coding for viral structural polypeptides. At the late stage of infection, progeny viral DNA once again assumes the viral chromatin structure (10). The replacement of basic nuclear proteins during the growth and development of the adenovirion appears to mimic the life cycle of histones and protamine. It is in this context that the major core protein VII is a convenient model system for study of the molecular biology of eukaryotic chromatin.

In an effort to understand the protein VII–DNA interactions, we studied its primary structure. In particular, we wondered whether VII is similar in structure to the histones and whether the protein sequence can shed light on the molecular organization of the adenovirus core.

MATERIALS AND METHODS

Cells, Virus Protein, and DNA. Adenovirus type 2 was propagated in suspension cultures of KB cells. The methods of virus purification, extraction, and chromatography of protein VII and its precursor (pro-VII) on a phosphocellulose column were done as described (12). DNA was isolated according to Pettersson and Sambrook (13).

Protein Sequence Determination. Chemical cleavage of protein VII with cyanogen bromide, *N*-bromosuccinimide, and *N*-chlorosuccinimide/urea and the separation of peptides have been described (12, 14, 15). Sequential degradation of peptides was performed with a Beckman 890C sequencer. The phenylthiohydantoin (>PhNCS) amino acids were determined with a Waters HPLC system; the results were confirmed by thin-layer chromatography.

Cloning of Viral DNA Fragments. Adenovirus DNA was digested with *Hin*dIII and the DNA fragments were separated on a 1% agarose gel. *Hin*dIII fragment D (map units 41.0–50.1) was eluted from gel slices and cloned into pBR322 (16). All experiments involving recombinant DNA were conducted under

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: pro-VII, precursor to major core protein VII (the designation pro conforms with the slow kinetics of processing of pro proteins; by contrast, the processing of pre protein is rapid and occurs in nascent chains on polyribosomes); >PhNCS, phenylthiohydantoin.

P1/EK1 containment in accordance with current National Institutes of Health guidelines.

DNA Sequence Determination. The procedure of Maxam and Gilbert (17) was used for sequence analysis and end-labeling of DNA.

RESULTS

Previous work in our laboratory established protein sequences of parts of the adenovirus core protein VII—the NH_2 terminus (12), a protamine-like internal cyanogen bromide fragment (15), and the pro-VII (unpublished data). In addition, we have other sequence information and this composite data can be briefly summarized. The entire sequence of protein VII was determined by automatic protein sequencer studies of the four cyanogen bromide-, three *N*-bromosuccinimide-, and one *N*-chlorosuccinimide-cleaved peptide fragments. Each peptide was analyzed at least twice and the resultant >PhNCS derivatives were identified by HPLC and thin-layer chromatography. As indicated by the arrows in Fig. 1, the composite data delimit actual sequence information obtained for each fragment. Thus, ≈97% of the amino acid sequence of the entire molecule was established.

To complement and to verify the protein sequence information, we constructed a recombinant DNA plasmid that carried *Hin*dIII fragment D (map position 41.0–50.1) of adenovirus type 2 DNA. The major core protein VII is known to be encoded within this DNA fragment (18). Knowledge of the protein sequence enabled us to locate the DNA encoding region. Fig. 2 shows our DNA sequence analysis strategy spanning an 800-base-pair segment within *Hin*dIII fragment D. The pro-VII gene is located at map coordinates 43.37–44.90. About 90% of the nucleotide sequence of the pro-VII gene was determined by analysis of both complementary strands. Our nucleotide sequence complements the protein sequence data in unequivocally establishing the complete gene and protein sequences of the major core protein VII and pro-VII (Fig. 1).

DISCUSSION

mRNA coding for the adenovirus late structural polypeptides are transcribed from the r strand of the viral genome and can be grouped into five families (L1–L5) of sequences (19). Each family has in common the position of their 3′ polyadenylylation sites and the identical 5′ tripartite leader sequences which are acquired through splicing of the segment of sequences from map positions 16.3, 19.7, and 26.5 (18, 20). The L2 family encompasses map coordinates 39.0–49.0 and codes for three viral structural polypeptides: III, V, and pro-VII. In the present study, we have elucidated the physical location of the pro-VII gene. We have also completed nucleotide sequences flanking the pro-VII-encoding regions (data not shown). Based on knowledge of the amino acid sequences of the minor core protein V (unpublished data), we have deduced that nucleotide sequences downstream from the pro-VII gene code for protein V and sequences upstream should code for protein III. The order of the L2 family of genes is: III (39.0–43.4), pro-VII (43.4–44.9), and V (44.9–49.0).

There is an A-G sequence located five nucleotides upstream

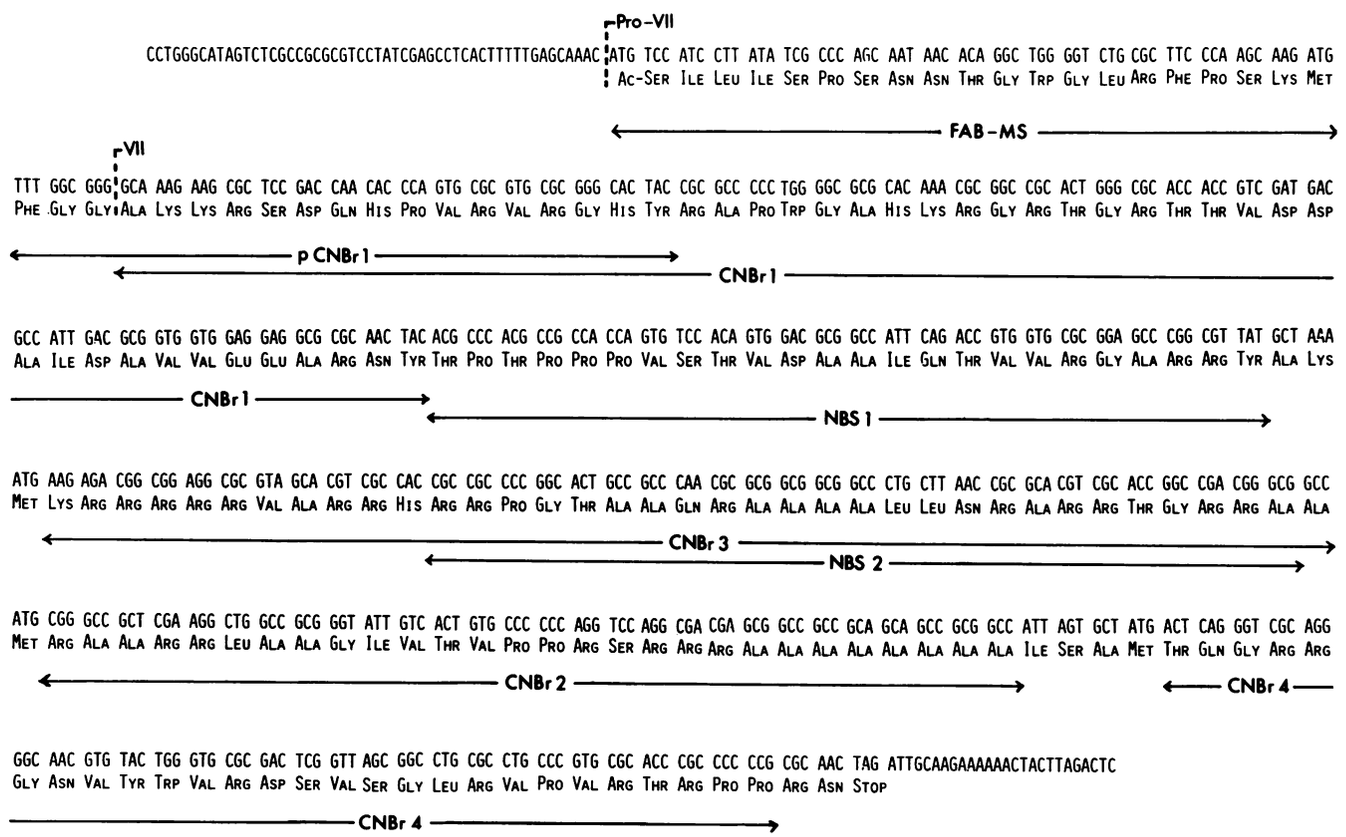

FIG. 1. DNA and protein sequences of major core protein VII. The DNA sequence shown is a region in cloned *Hin*dIII fragment D (map position 41.0–50.1) of adenovirus 2 DNA. The pro-VII sequence was determined by: (*i*) fast atom bombardment/mass spectrometry (FAB-MS) of the CNBr precursor fragment and (*ii*) spinning-cup sequence analysis of the largest CNBr fragment of pro-VII (pCNBr1). The processing site is indicated. The sequence of protein VII was determined by automatic protein sequence analysis of the four CNBr fragments (CNBr1, CNBr3, CNBr2, and CNBr4), *N*-bromosuccinimide fragments (only NBS1 and NBS2 are shown), and *N*-chlorosuccinimide/urea-cleaved fragments (not shown). The arrows delimit actual sequence information obtained for each fragment. NBS2 is due to a fortuitous cleavage at histidine. Duplicate analytical runs were usually performed for each of the isolated peptide fragments.

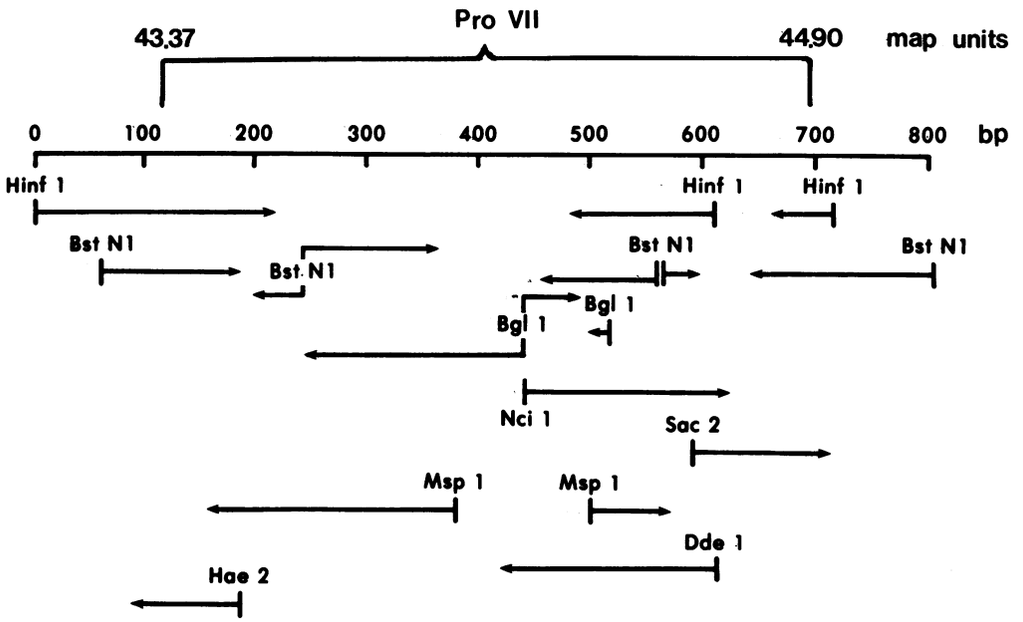

FIG. 2. Strategy for DNA sequence analysis of the pro-VII-encoding region. A schematic drawing of the *Hinf*I–*Bst*NI fragment (map position 41.0–43.0) illustrates the restriction enzyme cleavage sites and the strategy. Arrows indicate the direction and the amount of sequence information used from individual restriction cleavage sites. The pro-VII-encoding region is located at map position 43.37–44.90.

from the first ATG of the pro-VII gene (Fig. 1). The adjacent sequence T-C-A-C-T-T-T-T-T-G-A-G seems to fit the consensus splice junction acceptor site sequence of T-Py-Py-A-A-T-T-T-Py-G-C-A-G (21). The A-G sequences at −20 and −40 and their respective adjacent sequences are less homologous to the consensus sequence. However, the precise splice site for joining the tripartite leader to the pro-VII mRNA must await further characterizations. The putative poly(A) adding site for the L2 family mRNA is located at map position 49.0. The mature pro-VII mRNA should be approximately 2,500 nucleotides long, considering the lengths of the tripartite leader, the long mRNA body spanning map positions 43.4–49.0, and the poly(A) tail.

One of the significant conclusions about the gene structure is that there are no introns within the coding sequences. This is evidenced by the observed colinearity of the nucleotide sequence and the amino acid sequence. Another interesting feature of the gene is its 71% G+C content; the G+C content of the viral DNA is 58% (22). This high G+C content is due to the prevalence of arginine, alanine, glycine, and proline (Table 1), the codons of which all are rich in G and C. However, the nonrandom codon usage further contributes to the high G+C content. For example, the CGC codon for arginine is used 23 times in comparison to the other five arginine codons: CGU, 3 times; CGA, 4; CGG, 5; AGA, 1; and AGG, 5. Furthermore, there is a considerable bias for codons ending with C or G compared to codons ending with A or U. This is exemplified by the ratios of NN^{C}_{C} to NN^{A}_{U} for alanine, proline, and glycine—25/8, 11/4, and 11/4, respectively.

Pro-VII contains 196 residues, whereas protein VII is 176 residues long. Their respective molecular weights are 21,749 and 19,258 (Table 1). During maturation assembly of the adenovirion, the 23-residue precursor fragment (M_r, 2,509) is removed by a virus-encoded endoprotease. In Table 1 there is a correction for the COOH-terminal residue from lysine (12) to asparagine. In the end-group determination experiment, the free amino acid recovered after hydrazinolysis was identified by amino acid analysis. Lysine and asparagine have identical chromatographic behaviors, which caused the error of our previous identification (12).

The elucidation of the primary structure of this basic protein immediately raises the question of whether protein VII is similar in structure to the histones. Our analysis as described below suggests that it has a distinctive molecular architecture.

The protein is rich in both arginine (23%) and alanine (19%) (Table 1), and the arginine and alanine residues are clustered in the sequence. The repetition of alanine residues at positions 91–98 and 133–140 (Fig. 3) is most interesting and suggests a unique conformation. We therefore applied the Chou and Fasman method (23) to predict the secondary structure of these and other regions in protein VII. The 91–100 and 133–144 regions have Pα average values of 1.37 and 1.34 (Fig. 3, segments c and d), respectively; accordingly, we assigned these as α-heli-

Table 1. Amino acid composition of pro-VII, protein VII, and the precursor fragment

	Pro-VII	Protein VII	Precursor fragment
Asp	6	6	
Asn	6	4	2
Thr	13	12	1
Ser	10	6	4
Glu	2	2	
Gln	4	4	
Pro	14	12	2
Gly	15	11	4
Ala	33	33	
Val	17	17	
Met	4	3	1
Ile	6	4	2
Leu	6	4	2
Tyr	4	4	
Phe	2	0	2
Trp	3	2	1
Lys	6	5	1
His	4	4	
Arg	41	40	1
COOH-terminal	Asn	Asn	Gly
NH_2-terminal	Ac-Ser	Ala	Ac-Ser
M_r	21,749	19,258	2,509
Total residues	196	173	23

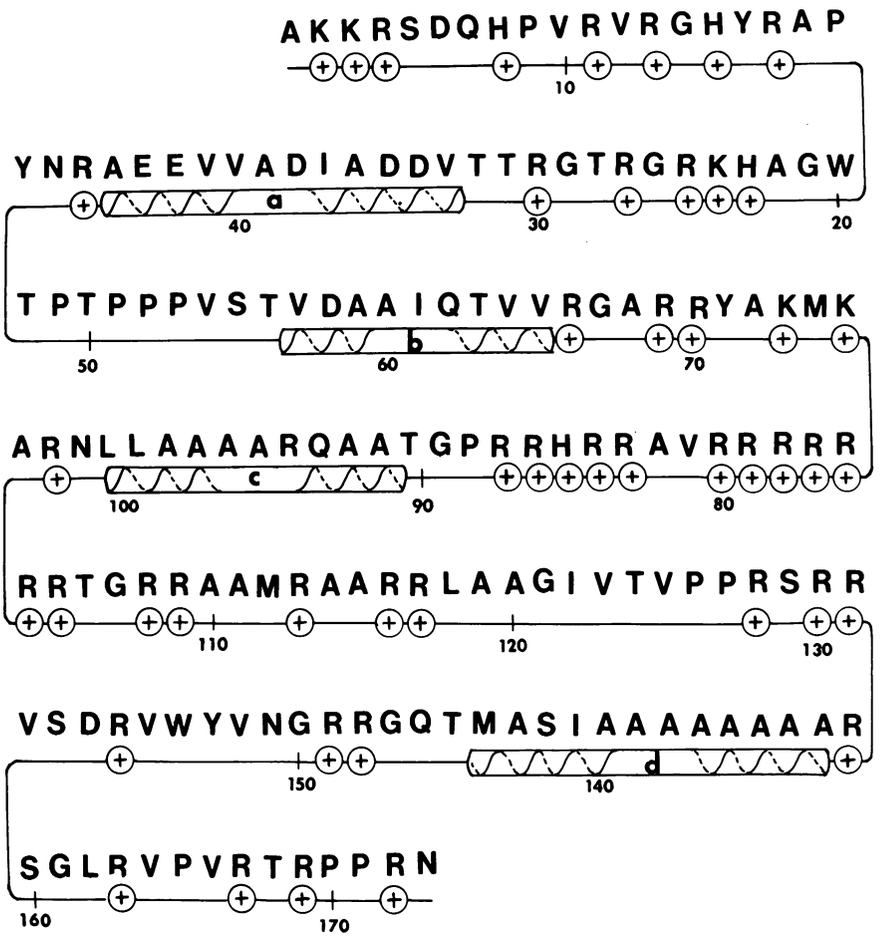

FIG. 3. Structural characteristics of protein VII. Schematic drawing shows the distribution of basic amino acids (⊕-⊕—⊕—) and potential α helical segments (segments a, b, c, and d) along the linear sequence of protein VII. The average values of Pα and Pβ for the α-helical segments according to Chou and Fasman (23) are as follows: segment a, Pα = 1.23, Pβ = 1.13; segment b, Pα = 1.14; Pβ = 1.30; segment c, Pα = 1.37, Pβ = 1.04; segment d, Pα = 1.34, Pβ = 1.06.

cal regions. Peptide segment 33–44 (Fig. 3, segment a) with a Pα value of 1.23 is also most likely to be in an α-helical conformation. The regions composed of residues 32–41 and 56–65 are partially duplicated regions with respect to both nucleotide and protein sequences. The Pα = 1.14 of residues 57–65 (Fig. 3, segment b) is a weak assignment for α helix. However, it also shows a strong Pβ value, 1.30. Therefore, this region will certainly have secondary structure.

The distribution of basic residues in protein VII is nonrandom and different from that of nucleosomal histones. This distribution is especially evident when the α-helical segments are assigned. Clearly, four basic domains are segregated by α helices (Fig. 3). We have already described the NH_2-terminal sequence of protein VII (residues 1–30) as histone-like (24). This domain is composed of ≈40% basic residues. The protamine-like domain (15), residues 66–90, is 60% basic. The last two histone-like basic domains (101–132 and 145–173) are 40% and 25% basic, respectively. Because DNA/basic protein interactions are principally ionic in nature, the four basic domains described above are likely to be involved in DNA binding.

A subunit organization of adenovirus chromatin of six molecules of protein VII and 200 base pairs of DNA has been proposed by Corden *et al.* (25). In a limit digest of virus cores, virus nucleosome-like monomers sediment as 11S particles, and the corresponding nuclease-resistant fragment contains 150 base pairs (7). Furthermore, a beaded chromatin structure has been observed by Mirza and Weber (7). Thus, it appears that adenovirus chromatin possesses a subunit nucleosome-like structure.

Our proposed architecture of the viral nucleosome is modeled after cellular nucleosomes. According to the principles of protein chemistry, the hydrophobic regions of protein VII should coalesce to give a tertiary structure. Specifically, we propose that the α-helical segments are juxtaposed by intrachain folding. The α-helical segments of individual proteins, as in the globular domains of the histones, can further participate in the formation of hexameric nucleosomal particles (25). Virus nucleosome monomers are known to organize 150 base pairs of the DNA (10). The four basic domains of protein VII are similar in function to the basic arms of the histones. In virus nucleosomes they are in combination with the $1^1/_2$ turns of the DNA phosphates. This model is not new in principle and appears to be realistic.

The strongest argument against the subunit model of adenovirus chromatin organization is the lack of an indication of a nucleosomal repeat pattern from numerous nuclease digestions (8–10). Our observation of a protamine-like domain in protein VII is relevant to these unusual nuclease digestion results. Honda (26) has observed that nucleoprotamine in spermatid-stage chromatin is resistant to micrococcal nuclease digestion, whereas nucleohistone chromatin is not. Warrant and Kim (27) observed protamine to be in a helical configuration when bound to a double-stranded region of tRNA. This led them to postulate that a protamine molecule contains four α-helical segments lying in the grooves of the DNA double helix. The

basic arginyl side chains may interact with adjacent DNA double helices, and the segments may act cooperatively to condense DNA. There is only indirect evidence to support the view that the protamine-like domain in protein VII also functions cooperatively to condense DNA. We studied the reconstitution of protein VII and adenovirus DNA, using the dialysis conditions for histones and DNA (28). The resultant chromatin was examined by electron microscopy. It revealed a highly aggregated nucleoprotein complex and, even at low input ratios of protein VII to DNA, crosslinked DNA fibers were observed. *In vivo*, the assembly of protein VII along the virus DNA is clearly nonrandom and is involved in the formation of nucleosome-like particles. A mechanism similar to that described above may involve the protamine-like domain binding to adjacent nucleosome regions. Two consequences may follow: (*a*) it produces variable lengths of spacer DNA, or (*b*) it renders the spaced DNA less susceptible to nuclease attack. This may account for the lack of a nucleosomal repeat pattern as described above.

The metabolism of protein VII early in the viral infection is also intriguing. Upon entry into a host cell nucleus, viruses are uncoated. The transcriptionally active and replicating adenovirus DNA molecules are devoid of protein VII. Instead, the virus DNA is probably associated with histones in transcription complexes and with the 72,000-dalton single-stranded-DNA-binding proteins of the replication complex (29). The replacement of protein VII, in the process of virus uncoating, resembles the decondensation of nucleoprotamine in the biological process of fertilization. Indeed, in electron micrographs of infected cells, the core exhibits a paracrystalline appearance which is similar to that seen in sperm. All of the above considerations lead us to conclude that protein VII is a hybrid histone/protamine basic protein, although it may resemble the protamine in function.

This work has been supported by Grant AI16571 from the National Institutes of Health and Equipment Grants GM27425 from the National Institutes of Health and PCM77-17808 from the National Science Foundation.

1. Marushige, K. & Dixon, G. M. (1969) *Dev. Biol.* **19**, 397–414.
2. Kornberg, R. (1977) *Annu. Rev. Biochem.* **46**, 931–954.
3. McGhee, J. D. & Felsenfeld, G. A. (1980) *Annu. Rev. Biochem.* **49**, 1115–1156.
4. Thoma, F., Koller, T. & Klug, A. (1979) *J. Cell Biol.* **83**, 403–427.
5. Sung, M. T. (1977) *Biochemistry* **16**, 286–289.
6. Everitt, E., Sundquist, B., Pettersson, U. & Philipson, L. (1973) *Virology* **52**, 130–147.
7. Mirza, M. A. & Weber, J. (1982) *Biochim. Biophys. Acta* **696**, 76–86.
8. Sergeant, A., Tigges, M. A. & Raskas, H. (1979) *J. Virol.* **29**, 888–897.
9. Tate, V. E. & Philipson, L. (1979) *Nucleic Acids Res.* **6**, 2769–2785.
10. Brown, J. & Weber, J. (1980) *Virology* **107**, 306–310.
11. Daniell, E., Groff, D. E. & Fedor, M. J. (1981) *Mol. Cell. Biol.* **1**, 1094–1105.
12. Sung, M. T., Lischwe, M. A., Richards, J. C. & Hosokawa, K. (1977) *J. Biol. Chem.* **252**, 4981–4987.
13. Pettersson, U. & Sambrook, J. (1973) *J. Mol. Biol.* **73**, 125–130.
14. Lischwe, M. A. & Sung, M. T. (1977) *J. Biol. Chem.* **252**, 4976–4980.
15. Cao, T. M. & Sung, M. T. (1982) *Biochem. Biophys. Res. Commun.* **108**, 1061–1066.
16. Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L., Boyer, H. W., Crosa, J. H. & Falkors, S. (1977) *Gene* **2**, 95–113.
17. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
18. Chow, L., Roberts, J. M., Lewis, J. B. & Broker, T. R. (1977) *Cell* **11**, 819–836.
19. Ziff, E. B. & Fraser, N. (1978) *J. Virol.* **25**, 897–906.
20. Berget, S. M., Moore, C. & Sharp, P. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3171–3175.
21. Akusjarvi, G. & Persson, H. (1981) *J. Virol.* **38**, 469–482.
22. Piña, M. & Green, M. (1965) *Proc. Natl. Acad. Sci. USA* **54**, 547–551.
23. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 211–222.
24. Lischwe, M. A. & Sung, M. T. (1977) *Nature (London)* **267**, 552–554.
25. Corden, J., Engelking, H. M. & Pearson, G. D. (1976) *Proc. Natl. Acad. Sci. USA* **73**, 401–404.
26. Honda, B. (1973) Dissertation (Univ. of British Columbia, Vancouver, BC).
27. Warrant, R. W. & Kim, S. H. (1978) *Nature (London)* **271**, 130–135.
28. Oudet, P., Gross-Bellard, M. & Chambon, P. (1975) *Cell* **4**, 281–289.
29. Kedinger, C., Brison, O., Perrin, F. & Whilelm, J. (1978) *J. Virol.* **26**, 364–379.