Organization and Nucleotide Sequence of a Densovirus Genome Imply a Host-Dependent Evolution of the Parvoviruses

HISANORI BANDO,^{1,2}†* JUN KUSUDA,¹ TAKASHI GOJOBORI,³ TAKEO MARUYAMA,³ AND SHIGEMI KAWASE2

Invertebrate Section, Genetic Stock Research Center,¹ and Department of Evolutionary Genetics,³ National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan; and Laboratory of Sericultural Science, Faculty of Agriculture, Nagoya University, Chikusa-ku, Nagoya, 464 Japan2

Received 11 June 1986/Accepted 8 October 1986

The genome structure of a densovirus from a silkworm was determined by sequencing more than 85% of the complete genome DNA. This is the first report of the genome organization of an insect parvovirus deduced from the DNA sequence. In the viral genome, two large open reading frames designated 1 and 2 and one smaller open reading frame designated 3 were identified. The first two open reading frames shared the same strand, while the third was found in the complementary sequence. Computer analysis suggested that open reading frame 2 may encode all four structural proteins. The genome organization and a part of the nucleotide sequence were conserved among the insect densovirus, rodent parvoviruses, and ^a human dependovirus. These viruses may have diverged from a common ancestor.

Many kinds of viruses have been isolated from a wide variety of organisms. Those viruses are classified into more than 50 families according to their characteristic traits (18). Viruses containing ^a linear, single-stranded DNA belong to the family Parvoviridae (29), which consists of three genera, parvovirus, dependovirus, and densovirus. Parvovirus and dependovirus are both vertebrate viruses, although the former is able to replicate autonomously and the latter is replication defective. On the other hand, densovirus is an invertebrate autonomous virus.

Members of the genus densovirus are commonly called densonucleosis virus (DNV) (17). In 1968 a viral disease of the silkworm Bombyx mori was discovered in sericultural farms in the suburbs of Ina City, Japan. Based on cytopathological, chemical, and physical characteristics (12), the causative agent of this disease was determined to be a Bombyx DNV.

The *Bombyx* DNV has a diameter of roughly 22 nm. Protein analysis of the *Bombyx* DNV showed that the virion possesses four structural proteins (22). The surface of the Bombyx DNV particle has the form of an icosahedron consisting of 12 capsomers, and viral protein 1 (VP1; molecular weight, 50,000) is thought to be the repeating unit in the capsid structure (23). The function of the remaining three proteins (VP2, VP3, and VP4) are not known, although the relative concentrations of each amino acid are known to be very similar among all four proteins (2) .

The DNA of Bombyx DNV is composed of ^a linear, single-stranded molecule with a molecular weight of about 1.7×10^6 , and two complementary strands are contained in different particles of the DNV (21). Each strand seems to possess an inverted repeat at each terminus (21). DNA synthesis occurs predominantly in infected nuclei of columnar cells of the silkworm midgut, indicating that the virus multiplies in the nucleus (33). The mode of replication of DNV DNA, however, is still obscure.

* Corresponding author.

Recently, the nucleotide sequences of the complete genomes of two species of the genus parvovirus (1, 26) and of one species of the genus dependovirus (31) were determined. These sequencing studies showed that the genome organization of parvovirus is similar to that of dependovirus, even though one is replication defective and the other is not. The genome organization of densovirus has not been characterized, however, because of the absence of suitable cell systems for DNV multiplication. With the aim of elucidating the genome organization, the evolutionary origin, and the replication process of the invertebrate virus, we determined more than 85% (4,277 nucleotides) of the complete genome DNA sequence of ^a densovirus (Ina isolate) from the silkworm B. mori.

MATERIALS AND METHODS

Enzymes and reagents. Restriction endonucleases, the Klenow fragment of Escherichia coli DNA polymerase I, the 17-mer synthetic M13 primer, nucleotides, dideoxynucleotides, and T4 polynucleotide kinase were purchased from Takara Shuzo Co. Ltd. Bacterial alkaline phosphatase came from Worthington Diagnostics (Freehold, N.J.). The M13 vectors mp8 to mpll were obtained from P-L Biochemicals, Inc. (Milwaukee, Wis.). $[\alpha^{-32}P] dCTP$ (specific activity, >3,000 Ci/mmol) was purchased from Amersham Corp. (Arlington Heights, Ill.).

DNA sequencing. The virion DNA extracted from purified DNV particles (21) was used as the source of DNV DNA. The nucleotide sequences were determined by the dideoxy chain-termination method (29) by using ^a range of M13 vectors (20). The sequencing strategy is shown in Fig. 1. Sequencing reactions and gel conditions have been described elsewhere (27). The nucleotide sequence that was determined was confirmed by shotgun sequencing by using the plasmid clone (pCBg42) containing the virus DNA (H. Bando, J. Kusuda, and S. Kawase, Arch. Virol., in press).

RESULTS

Nucleotide sequence of Bombyx DNV genome. We determined 4,277 nucleotides of the Bombyx DNV DNA se-

^t Present address: Department of Microbiology, Mie University School of Medicine, Edobashi, Tsu, Mie, 514 Japan.

FIG. 1. Sequencing strategy used to determine the nucleotide sequence of the Bombyx DNV genome. The reference line represents the DNV genome of 4.9 kilobase pairs (scale between each mark on the axis, 1,000 nucleotides). The vertical bars represent the restriction sites, and the arrows indicate the direction and extent of the sequence obtained. The restriction endonuclease sites are abbreviated as follows: A, AccI; B, BglII; D, DraI; E, EcoRI; H, HindIII; P, PstI; S, SalI; Sc, ScaI.

____**__ORF 1**
___ _____OO
GATCTTGTAA GGTGCATGCG CAATAGAGGT GAACACGAAG AATATACAGA CTTGCTTGAT
CTAGAACATT CCACGTACGC GTTATCTCCA CTTGTGCTTC TTATATGTCT GAACGAACTA 80 90 100 110 120 AATGTTAATA CTCCTCGTGA ACTTTTGTCT GCTATGGAGT CTTTTGAAGA GTCTTCAAAT TTACAATTAT GAGGAGCACT TGAAAACAGA CGATACCTCA GAAAACTTCT CAGAAGTTTA 140 150 160 170 180 GAAATTCCAA GTCAAGCAGC GACCGGAGTA TTTGCAGAAA CTGGAAACCG TTATGATGTC CTTTAAGGTT CAGTTCGTCG CTGGCCTCAT AAACGTCTTT GACCTTTGGC AATACTACAG 200 210 220 230 240 CCACGATCCC AAGATTGGGA TACTATGATG AAAGTTGTGA GCGCAGAAGA AGAAATGATA GGTGCTAGGG TTCTAACCCT ATGATACTAC TTTCAACACT CGCGTCTTCT TCTTTACTAT 250 260 200
GAATICITAA CCGAGGAAT CTCAGCATI ATTCAAAG ATTGCAAA CAACTCGTTG
CTTAAGAATT GCCTCCTTAA GAGTCG<u>TAAA TAA</u>GTTTTIC TAACGTTTTI GTTGAGCAAC
20 310 220 300 300 300 300
GAAGAATTGA TCGAAACAGT GGACGAGCCG AAATTGGTAC ATTCAAGAAC 380 390 400 410 420 GAGTATTACT TAACGCTCAT CTTACAGAAC AAAGCAAAAG AAACTACGGA TTTGAAAAGG CTCATAATGA ATTGCGAGTA GAATGTCTTG TTTCGTTTTC TTTGATGCCT AAACTTTTCC 440 450 460 470 480 CCATTCCCTG TCAGTTATCA ACAAATGGAA CTCTGTCATA AAGACGCTTT GAACTCATCT GGTAAGGGAC AGTCAATAGT TGTTTACCTT GAGACAGTAT TTCTGCGAAA CTTGAGTAGA 500 510 520 530 540 GAACGAGAAA GAAGCCTCGA CCAATTCTTT CGAGCGGTGG AACCGAGAAT GGGAATCCTC CTTGCTCTTT CTTCGGAGCT GGTTAAGAAA GCTCGCCACC TTGGCTCTTA CCCTTAGGAG 550 560 560 570
GCAGATCAGT TATGCAGACG TTTTGAGAAT AATACAGTA ATGCAAGA AATTTACACG
CGTCTAGTCA ATACGTCTGC AAAACTCTTA TTATGTCAAT TACGTTCTTC TTA<u>AAT</u>GTGC Terminator
660 620 630 640 650 660 ACTCCATTGA GTTTAGACCA GGCCACAGAT TTCGTGAAAT GGTTTACAAA TAGAGGAGAG TGAGGTAACT CAAATCTGGT CCGGTGTCTA AAGCACTTTA CCAAATGTTT ATCTCCTCTC 680 690 700 710 720 CTCTTGATAG CAGAGGTTCG TACGCTATTT GGTTCCATGA AACAAATAAA GGAACGGAAG GAGAACTATC GTCTCCAAGC ATGCGATAAA CCAAGGTACT TTGTTTATTT CCTTGCCTTC 740 750 760 770 780 ACTACCCACA AGACATTAGT AGAGGATACC CACCCGACTA CGATTTCGGA GTCATACCGA TGATGGGTGT TCTGTAATCA TCTCCTATGG GTGGGCTGAT GCTAAAGCCT CAGTATGGCT 840 800 810
CTAGATACCG ATACTGCGGA AAGCCAGCAA TCGAGCCAT CAGAATCAC GAGGGAGGA
CTACTATGGC TATGACGCCT TTCGGTCGTT AGCTCGGTTA GTCTTAGTCT CTCCCTCCIC 860 870 880 890 900 GAGAATCGCA GGAACCGACC ArCGAGTTCA AGACATATAA ATACTACTCG GAAAAGAAAG CTCTTAGCGT CCTTGGCTGG TAGCTCAAGT TCTGTATATT TATGATGAGC CTTTTCTTTC 920 930 940 950 960 TCTATCACAA CATCGAAGGG CGTGCTTACG AAGAAAAAAA GTTTGAAGAA CCAACCCAGG AGATAGTGTT GTAGCTTCCC GCACGAATGC TTCTTTTTTT CAAACTTCTT GGTTGGGTCC 970 980 1010
GCCATTTCCA CATACTTCAC GCCTGCAGT GGTACATAA TGAATGG TGCCTCGGGA
GGGTAAAGGT GTATGAAGTG CGGACGTCTA CCATGTTATT ACTTACGTCC ACGGAGCCCT 1030 1040 1050 1060
GGAGCTTTGA CGTCAATCCA AGGAAGAATA AGGCAATECC TATCGACTCG GTCGACACCG
GCTCGAAACT GCAGTTAGGT TCCTTCTTAT TCCGTTAGGG ATAGCTGGGC CAGCTGTGGC

quence, extending over 85% of the complete genome (Fig. 2). The nucleotide sequence that was obtained accounts for all known restriction sites of Bombyx DNV DNA (Bando et al., in press). In particular, the computer analysis of the nucleotide sequence predicted no recognition site for the restriction enzymes HpaI, KpnI, SmaI, StuI, XbaI, and XhoI; only one site each for BamHI, HindIII, PvuII, and Sall; two sites for PstI; and four sites for DraI. These predictions are consistent with the actual restriction maps of the viral DNA that have been described elsewhere (Bando et al., in press). The nucleotide sequence of Bombyx DNV DNA is presented in Fig. 1.

Location of open reading frames. Open reading frames (ORFs) in the DNA sequence were determined by computer analysis. The Bombyx DNV genome contains three major ORFs: ORFi and ORF2 lie in one strand, while the third

FIG. 2. Nucleotide sequence of the Bombyx DNV genome. Major ORFs (ORF1, 2, and 3), putative signals of transcription and translation (CAAT box, TATA box, initiation codon, terminator, polyadenylation signal), possible splicing junctions (Acp; acceptor site, Don; donor site), and the conserved region are highlighted and are discussed in the text.

ORF, ORF3, is in the complementary strand (Fig. 3). ORF1 consists of 1,290 nucleotides (Fig. 2). If ORFi is translated into ^a polypeptide without RNA splicing, ^a protein of 43,000 daltons is produced. The molecular weight of this protein is larger, however, if the initiation codon is located in an upstream region of ORFi where the sequence has not been completed. ORF2 contains 887 codons and is located between nucleotides 1546 and 4207 shown in Fig. 2; it encodes a protein of at most 89,000 daltons. ORF3 is in the complementary strand and consists of 501 nucleotides or 167 codons.

To clarify further the genome organization and coding potential of the Bombyx DNV genome, we analyzed the distribution of initiation signals for transcription and translation in the sequence. In eucaryotes one of the control regions of transcription by RNA polymerase II is often characterized by ^a TATA box, (M. L. Goldberg, Ph.D. thesis, Stanford University, Stanford, Calif., 1979) which

Initiation codon2 Don 2170 2180 2190 2200 2210 2220 TGATAAGTAT GTAACCGATT TGTTAATCTT AATCAAA=A TATAATCACA GG3TGATGGG ACTATTCATA CATTGGCTAA ACAATTAGAA TTAGTTTTAC ATATTAGTGT CCCACTACCC 2240 2250 2260 2270 2280 TAATAAAATA TGTAATCTTA TCAAAATGTA TAAATATAGG GTGATGGTAA TAAATATATT ATTATTTTAT ACATTAGAAT AGTTTTACAT ATTTATATCC CACTACCATT ATTTATATAA TATA box 2300 2300 2310
2310 2330 2330 2310 2310 2310
CAGTGCATTC ATATGCCTGA TCTTAATTTT CCTTACTAAT AATTATCTTG GTCCGGGACT
GTCACGTAAG TATACGGACT AGAATTAAAA GGAATGATTA TTAATAGAAC CAGGCCCTGA 2360 2370 2380 2390 2400 GTATACATGT AAATCCATAG ACGAGACGAC GCTATCCGAG GCCGTAGTAA TTTGGCCTTC CATATGTACA TTTAGGTATC TGCTCTGCTG CGATAGGCTC CGGCATCATT AAACCGGAAG 2420 2430 2440 2450 2460 AGATAAAGTA ACCAATCATA AGGAAGTTTT TCAAGCTGAT AAACAGGCCC GTGACGAGTT TCTATTTCAT TGGTTAGTAT TCCTTCAAAA AGTTCGACTA TTTGTCCGGG CACTGCTCAA 2480 2490 2500 2510 2520 TTTTACTTCA TTTGTGCATA TCGGAAACGT GCATAGTTTA ATTGGCGGTA TTGGACTTGG AAAATCAAGT AAACACGTAT AGCCTTTGCA CGTATCAAAT TAACCGCCAT AACCTGAACC Inittaton codon ²⁵³⁰ ²⁵⁴⁰ ²⁵⁵⁰ ²⁵⁶⁰ ²⁵⁷⁰ 2o ⁸⁰ AACTAAAAAT TTGGTAGAAG AACATGTATT AGGTAAACCC TTGTACGGAA TGGCAAAAG TTGATTTTTA AACCATCTTC TTGTACATAA TCCATTTGGG AACATGCCTT ACCCGTTTTC 2600 2610 2620 2630 2640 AAAATCAACT GAAAAAGATT GGGCCAAAAT CAAAAGGATT AATAGAGCTA GAGCCGCAAG TTTTAGTTGA CTTTTTCTAA CCCGGTTTTA GTTTTCCTAA TTATCTCGAT CTCGGCGTTC 2660 2670 2680 2690 2700 ACGAGAAAAC CAAGAAAACC AACCAGATAT TAGAGAATTT GGACACGTAG CTGGACAAAA TGCTCTTTTG GTTCTTTTGG TTGGTCTATA ATCTCTTAAA CCTGTGCATC GACCTGTTTT 2720 2730 2740 2750 2760 TATTAACGCA GACCAAGAAG TAAATTTGGC TGACTTTCCT GACTTTTTAC AAGACTTTGA ATAATTGCGT CTGGTTCTTC ATTTAAACCG ACTGAAAGGA CTGAAAAATG TTCTGAAACT 270 2810 2800 2800 2800
2820 2810 2820 2820 2830 2840
TGCCGAAGCA GGACCAAGTG GAACTCAACC AGTAACAACAAT CTCCTTGTTA GAGGAGGTTG
ACGGCTTCGT CCTGGTTCAC CTTGAGTTGG TCAGCTTTGT CGTGTTGTTA GAGGAGGTTG 2840 2850 2860 2870 2880 AATGTCTGAA GATATACAAC CAATGGAAAC CGTCGGGGCC ACTGATACCG GAGGAGGAGC TTACAGACTT CTATATGTTG GTTACCTTTG GCAGCCCCGG TGACTATGGC CTCCTCCTCG 2900 2910 2920 2930 2940 TCAAGTCGAT CCACGTACTG GAGGACAAGC AGCTGGAGGA TCCGAAATGG GAGCTGGTGG AGTTCAGCTA GGTGCATGAC CTCCTGTTCG TCGACCTCCT AGGCTTTACC CTCGACCACC 2950 2950 2960 2970 2980
ATCAGCTAAT GATGGTAGAG AAGACATTIT TICTGGAGCA CCACAA ATCAACATCA
TAGTCGATTA CTACCATCTC TTCTGTAAAA AAGACCTCGT GGTGTTGGTT TAGTTGTAGT 3020 3030 3040 3050 3060 TACATTAGTA TATGGAAAAA GCTACCATTT CACAATAACA AAATGGTTTA CTGAATTTCG ATGTAATCAT ATACCTTTTT CGATGGTAAA GTGTTATTGT TTTACCAAAT GACTTAAAGC 3070 3080 3090 3090
ACATTTAGCA ACAACGAACT CGGGCTATTA CGCTCAACAA CGTTTTAAAC ATATACATGG
TGTAAATCGT TGTTGCTTGA GCCCGATAAT GCGAGTTGTT GCAAAATTTG TATATGTACC 3140 3150 3160 3170 3180 AATTCCATGG GAAAGACTAC TAATGTATGT AAGTGAAGGC GAACTCCTCC GAATGTTTAG TTAAGGTACC CTTTCTGATG ATTACATACA TTCACTTCCG CTTGAGGAGG CTTACAAATC 3190 3200 3200 3210 3240
AGATTATACT TCATTGAAAG TGGAAGAAGT AGTATGTA GTCTATAGTC TCGGAGTACG
TCTAATATGA AGTAACTTTC ACCTTCTTCA TCATACACTT CAGATATCAG AGCCTCATGC occurs approximately 30 nucleotides upstream of the cap site of the mRNA. Another sequence in the -70 to -80 nucleotide region has also been implicated in transcriptional control of both viral and nonviral genes (3, 6, 10), and its consensus sequence is GGPyCAATCT, which is commonly called the CAAT box. In eucaryotic nuclear genomes, the only initiation signal of translation is AUG, and it usually has purines at its -3 and 4 positions (i.e., three nucleotides upstream and four nucleotides downstream from the AUG codon). In particular, a purine in position -3 has a dominant effect (16).

Relying on these criteria, we searched for possible initiation signals in the Bombyx DNV sequence. Three sets of possible initiation signals were found for ORF2. The first set, located in the 5' region of ORF2, contained ^a CAAT box (at nucleotide 1383), ^a TATA box (at nucleotide 1447), and ^a A**AUG sequence (at nucleotide 1553). The other two sets contained only Pu**AUG (Pu is purine) sequences and ^a TATA box (Fig. ¹ and 2). If all three of these initiation signal

3250 3260 3270 3280 ATTACCTTTT GTAACTTCAG CCACTACCAG TTCAGTTGCT TAATGGAAAA CATTGAAGTC GGTGATGGTC AAGTCAACGA TTGCGATTGC 3360 3310 3330 3340
CATCGATGTT TTTCATTTTG ATGAAGCTTA TGAAACCAAC
GTAGCTACAA AAAGTAAAAC TACTTCGAAT ACTTTGGTTG 3370 3400 3980 3990 3400
AGACATCATA AATAAAGCTC TTGGAACA ATGGAAAAT GCTACACGGC CTACTGCTCC
TCTGTAGTAT TTATTTCGAG AACCTTGACT TACCTTTTTA CGATGTGCCG GATGACGAGG 3430 3440
3480 3470 3480 3440 3450 3450 3460
ACATTGTTGT CGAACCAGTC TTGTTAAAGG CTTATATAGC CGTAGTAGAT GCTCATCCCT 3490 3500 3510 3520 3540 TATAAACAAT CCCGTAATCG TTGATTATTC TCTTCCATAT TTTGAAAATA ATGTGCCTAA ATATTTGTTA GGGCATTAGC AACTAATAAG AGAAGGTATA TACACGGATT 3550 3560 3570 3560
AGACGTCGGA ATATATGACT ACGTTGACAT TAAAAATGGA ACTACTGCTT ACGGTAAATG
TCTGCAGCCT TATATACTGA TGCAACTGTA ATTTTTACCT TGATGACGAA TGCCATTTAC 3610 3620 3630 3640 CTGGGAAAAA CGATTTAAAC CTACGAATGG GACCCTTTTT GCTAAATTTG GATGCTTACC TGAAAATATA GAAACTTTCC 3670 3680 3690 3700 AAACGTAGTC ACTCCGCTTG CAGCACCAAC CTGGATTAGA TTTGCATCAG TGAGGCGAAC GTCGTGGTTG ATTATATTAC TGTGGTTATG GACCTAATCT 3730 3740 3740 3750 3760 3760 3760 3760
AAATGGGTAT TTCATGAGCA ATGACCAAT AAGAGAAGCA CGAGACCTAA CTACAAGTGT
TTTACCCATA AAGTACTCGT TACTGGTTTA TTCTCTTGCT GCTCTGGATT GATGTTCACA 3790 3800 3810 3820 3840 ACCACCTGAT GCTCTAACAG CTACAAAATT AAATCAAAGT GCTTCTAATA ATTTAAATGC TGGTGGACTA CGAGATTGTC GATGTTTTAA TTTAGTTTCA CGAAGATTAT TAAATTTACG 3850 3860 3870 3880 3900 ATTTGTGGAT TACATGGGTT ATAATTATTT CGGCGAACAA AAAGCGCCGC AATCAATGCC TAAACACCTA ATGTACCCAA TATTAATAAA GCCGCTTGTT TTAGTTACGG 3910 3920 3930 3940 3950 3960 TAAGTTTATG ATTGGATTTG TAAACATTAG AAACGAAGAC AATTCTCTAC TTAATGCTAA ATTCAAATAC TAACCTAAAC 3970 3980 3990 3990 4000 4000 4000 4010 4010
ATGGGACATT TTAATTAAAA CTCGAATTAG ACTCACTGGA CTTCAATCTA CTAGGGAATG
TACCCTGTAA AATTAATTTT GAGCTTAATC TGAGTGACCT GAAGTTAGAT GATCCCTTAC 4030 4040 4050 4060 4070
GGTTGCTAGA ACGGATAGAA TTCCGCCACA ATATTTCACA TCACAATATA CGCAGTTCCC
CCAACGATCT TGCCTATCTT AAGGCGGTGT TATAAAGTGT AGTGTTATAT GCGTCAAGGC 4100 4100 4100
CTATCCTAAT ATAAATGACA CTCCTCTACT AAGGTCTTTG GGTACATTTA AGTTACCAAC
GATAGGATTA TATTTACTGT GAGGAGATGA TTCCAGAAAC CCATGTAAAT TCAATGGTTG 4150 4160 4170 4180 4190 4200
TAAACGACCA GGAATGGATT CAAGAATAGC GCTGGGCGAG CTGCAGAAAC AACGCAAGAT
ATTTGCTGGT CCTTACCTAA GTTCTTATCG CGACCCGCTC GACGTCTTTG TTGCGTTCT4 Terminator Acp 4220 4230 4240 4250 4260 GAATTTATAG 'ATGCACAAAA GAAACTTCGT CTTAAATATC TACGTGTTTT CTTTGAAGCA

4270 4277 AAACCTGTAA CTAGATC TTTGGACATT GATCTAG

FIG. 2-Continued.

FIG. 3. ORFs in the Bombyx DNV genome. The upper three lines correspond to the three ORFs in a genome strand and the lower three lines are those in the complementary strand. The vertical lines represent the position of the termination codon in each frame. kb, Kilobases.

sets are functional, proteins with molecular weights of about 89,000 (I-1 [initiation signal 1]), 67,000 (I-2), and 55,000 (1-3) would be produced from ORF2 (Table 1). We could not identify any likely initiation signals for transcription and translation of ORF1; it is possible that its promoter is located in the upstream region, for which the sequence has not been determined. ORF3 appears to have one set of signals for transcription and translation (Fig. 1). Because only one strand is thought to be the sense strand for mammalian parvoviruses (4), ORF3 in the complementary strand may encode a protein that is unique to densoviruses.

Codon usage and amino acid composition. There were differences in the codon frequencies among these ORFs. In particular, aspartic acids and asparagine were more frequent in ORF2 than in ORF1. In Table ¹ are shown the amino acid compositions of the structural proteins obtained previously (2) and those of the three ORFs predicted from the nucleotide sequence. The amino acid frequencies in all the structural proteins are similar to those predicted from ORF2 but different than those predicted from ORFi and ORF3. This suggests that all structural proteins may be coded by a single ORF, ORF2. Recently, Nakagaki and Kawase (22) have reported that the four structural proteins VP1, VP2, VP3, and VP4 have molecular weights of 50,000, 57,000, 70,000, and 77,000, respectively. If the total molecular weight represents the products of individual genes, it would exceed the coding capacity of the Bombyx DNV DNA. Because ^a similar situation occurs for the other parvoviruses, it is possible that ORF2 codes for all four structural proteins.

Homology among parvovirus, dependovirus, and densovirus genomes. To ascertain whether any homologous domains exist among the nucleotide sequences of the parvovirus, dependovirus, and densovirus genomes, the genome sequences of rodent parvoviruses H-1 virus (H-1) and minute virus of mice (MVM); a human dependovirus, adenoassociated virus ² (AAV-2); and Bombyx DNV were examined by a dot matrix method. The sequence comparisons presented in Fig. 4 indicate that there is a region showing more than 80% homology between H-1 and MVM. In the nucleotide sequence of 440 bases in that region, a subregion

Amino acid	Composition (mol%) of the following:									
	ORF1	ORF ₂								
		$I-1a$	$I-2^a$	$I-3a$	ORF3	VP1	VP2	VP3	VP ₄	
Asp	8.9	13.2	13.5	14.0	14.0	14.1	14.4	14.2	13.7	
Thr	11.1	7.7	8.9	9.1	2.1	8.3	8.3	8.1	7.6	
Ser	9.3	5.7	5.9	6.2	15.2	8.0	10.3	6.1	10.0	
Glu	14.0	11.1	13.1	11.7	3.3	10.4	9.7	12.7	11.7	
Gly	3.1	5.8	8.0	7.7	3.8	8.3	8.6	8.2	8.4	
Ala	4.7	6.6	7.5	8.4	5.4	10.3	9.9	8.4	8.2	
Val	4.8	5.9	6.2	5.5	16.7	7.2	6.7	7.4	6.9	
Met	2.9	2.9	2.8	2.6	2.0	1.2	1.6	1.6	1.5	
Ile	5.1	6.7	6.7	5.7	6.8	6.8	6.6	6.7	6.3	
Leu	7.0	7.1	7.1	6.6	10.1	7.5	7.7	8.6	8.2	
Tyr	2.9	5.4	4.7	4.6	1.3	4.4	4.1	3.3	3.3	
Phe	4.0	5.5	5.1	4.9	11.9	3.9	3.4	4.5	4.1	
Lys	6.4	6.3	5.8	5.3	4.1	3.1	2.7	3.1	2.8	
His	2.2	1.9	2.0	1.6	0.5	1.1	1.0	1.0	1.1	
Arg	9.0	5.2	5.4	6.2	6.0	5.7	5.0	6.3	6.2	

TABLE 1. Amino acid composition of proteins studied

 a Tentative reading frames from initiation codons 1, 2, and 3 (Fig. 2).

FIG. 4. Dot matrix comparison of the sequence homologies between H-1 and MVM, H-1 and AAV, and H-1 and DNV. A dot represents a segment of 120 bases in which there are more than 48 bases that are identical between the two sequences being compared. The lines show the coding regions of probable structural proteins (SP) or nonstructural proteins (NS). Scale between each mark on both axes, 1,000 nucleotides. Double-headed arrows indicate commonly conserved regions (see text).

of more than 40% homology existed between H-1 and AAV-2. Furthermore, within that region the Bombyx DNV genome was also shown to have 35% homology with the H-1, MVM, and AAV-2 genomes in ^a 300-base sequence (Fig. 4). We take the homologies found among the viral genomes as evidence for a common origin of these viruses. Therefore, the nucleotide sequence in this region has been conserved in each of the evolutionarily diverged members of the Parvoviridae family. Interestingly, this conserved region is located not only within ORF2 of the Bombyx DNV genome but also within sequences encoding the nonstructural proteins of the other viruses examined. Thus, this domain may code an important active site in the gene product.

These regions of the four viruses were aligned with each other to estimate the number of nucleotide substitutions (Fig. 5). We computed the proportion of different nucleotides in these homologous regions among the different viruses. By using these proportions, the total number of nucleotide substitutions per site at all codon positions (Table 2) was estimated by the four-parameter method (32). It is clear from the results shown in Table 2 that the substitution number at the third position of a codon is much larger than those at the first two positions. In fact, many synonymous substitutions were observed in this region. Hydrophobicity in this region was also conserved during the evolution of these viruses. The observations indicate that this conserved

1721
DNV TYR GLU ILE LEU GLU LYS AG CAT CAA AAA ACA AAC AAC ATA TTC CAA ATA GTA AGT CCA CCA AGC GCA GGA AAA AAC TTT TTT ATA
1407 TYR GLU ILE LEU GLU LYS LYS HIS GLN LYS THR ASN THR PHE GLN ILE VAL SER PRO PRO SER ALA GLY L H-i 1404
MVM TGC TGT GTT TTA AAC AGA CAA GGA GGC AAA AGA AAT ACT GTT TTA TTT CAT GGA CCA GCC AGC ACA GGC AAA TCT --- ATT ATT
CYS CYS VAL LEU ASN <u>ARG</u> GLN GLY GLY L<u>YS ARG ASN THR</u> VAL LEU PHE HIS <u>GLY PRO ALA SER THR GLY LYS</u> 1271
AAV CTG GGA TGG GCC ACG AAA AAG TTC GGC AAG AGG AAC ACC ATC TGG CTG TTT GGC CCT GCA ACT ACC GGG AAG ACC --- AAC ATC
LEU GLY TRP ALA THR <u>LYS</u> LYS PHE GLY <u>LYS ARG ASN THR</u> ILE TRP LEU PHE <u>GLY PRO ALA</u> THR <u>THR GLY LY</u> GAA ACA GTA CTT GCA TTT
GLU THR VAL LEU <u>ALA</u> PHE GCA CAA GCC ATA GCA CAA
<u>ALA</u> GLN <u>ALA ILE ALA</u> GLN GCA CAA GCC ATA GCA CAA
<u>ALA</u> GLN <u>ALA ILE ALA</u> GLN GCG GAG GCC ATA GCC CAC
<u>ALA</u> GLU <u>ALA ILE ALA</u> HIS * ** * *** ** ** ** * ** * * *** ** DNV TAT TGG AAC ACT GGG GTC ATT CAA AAT TTT AAC CGA TAC AAC AAT TTC CCG TTA ATG GAA GCT GTT AAT AGA AGA GTA AAC TAT TGG GAT GAA CCA AAC TTT
TYR TRP ASN THR GLY VAL ILE GLN <u>ASN</u> PHE ASN ARG TYR <u>ASN ASN PHE PRO</u> LEU MET GL H−1 GCA GTT GGT AAT GTT GGT TGT TAC AAT GCT GCC AAT GTG AAC −−− TTT CCA TTT AAT GAC −−− TGT ACC AAC AAA AAC TTG ATT TGG GTG GAA GAA GCT GGT
ALA <u>VAL</u> GLY ASN VAL <u>GLY CYS</u> TYR <u>ASN</u> ALA ALA <u>ASN</u> VAL <u>ASN PHE PRO PHE ASN </u> MUM GCA GTT GGC AAT GTT GGT TGC TAT AAT GCA GCC AAT GTA AAC --- TIT CCA TIT AAT GAC --- TGT ACC AAG AAG TAG AT TGG GTA GAA GAA GCT GGT
ALA <u>VAL</u> GLY ASN VAL <u>GLY CYS</u> TYR <u>ASN</u> ALA ALA <u>ASN</u> VAL <u>ASN PHE PRO PHE ASN ASP C</u> AAV ACT GTG CCC TTC TAC GGG TGC GTA AAC TGG ACC AAT GAG AAC --- TIT CCC TTC AAC GAC --- TGT GTC GAG ATG GTG ATC TGG TGG GAG GGG AAG GAG GAG GAG GGG AAG GGG AAG GAT LA THE PROPHETICAL TRANSPORT OF THE REAL ALA ALSO PHETICAL TITLE OF THE REAL AND THE REAL AND THE REAL AND THE TREAT OF THE REAL AND T 1701
H-1 AAC TTT GGC CAG CAA GTA AAC CAA TTC AAA GCT ATT TGT TCT GGC CAA ACC ATA CGC ATT GAA AGA GGA AAA GGC AAC AAC
ASN PHE GLY GLN GLN <u>VAL</u> ASN GLN PHE <u>LYS ALA ILE</u> CYS SER GLY GLN THR ILE <u>ARG</u> ILE <u>ASP GLN LYS</u> GLY <u></u> AAC TIT GGA CAG CAA GTA AAC CAG TIT AAA GCC ATT TGC TCT GGT CAA ACT ATT CGC ATT GAT CAA AAA GGA AAA GGC AGC AAA CAG ATT GAA CCA ACA
ASN PHE GLY GLN GLN <u>VAL</u> ASN GLN PHE <u>LYS ALA ILE</u> CYS SER GLY GLN THR ILE <u>ARG</u> ILE <u>ASP</u> AAV ATG ACC GCC AAG GTC GTG GAG TCG GCC AAA GCC ATT CTC GGA GGA AGC AAG GTG CGC GTG GAC CAG AAA TGC AAG TCC TCG GCC CAG ATA GAC CCG ACT
HET THR ALA LYS VAL <u>VAL</u> GLU SER ALA <u>LYS ALA ILE</u> ILE GLY GLY SER LYS VAL <u>ARG</u> VAL ASN PHE GLY GLN GLN <u>VAL</u> ASN GLN PHE <u>LYS ALA</u> ILLE CYS SER GLY GLN THR ILLE <u>ARG</u> ILLE A<u>SP GLN LYS</u> GLY <u>LYS GLY SER LYS GLN ILLE GLU PRO THR</u>
AAC TIT GGA CAG CAA GTA AAC CAG TIT AAA GCC ATT TGC TCT GGT CAA ACT ATT CGC

FIG. 5. Comparisons of nucleotide sequences and predicted amino acid sequences in the conserved regions of the parvovirus genome. The alignment was made with the computer program described by Wilbur and Lipman (34). Underlining indicates the amino acids that were conserved in at least three viruses. Asterisks indicate the nucleotides that were conserved in the sequences of all four viruses. The numbers refer to the nucleotide number in the DNV, H-1 (26), MVM (1), and AAV-2 (31) genomes.

		Proportion of nucleotide sites of the following viruses ^a :					
Virus	Codon position	DNV	$H-1$	MVM			
$H-1$		0.571(1.124)					
		0.612(1.292)					
		0.663(2.332)					
	All	0.616(1.341)					
MVM		0.571(1.143)	0.020(0.021)				
		0.612(1.292)	0.000(0.000)				
		0.684(2.452)	0.245(0.319)				
	All	0.622(1.365)	0.088(0.095)				
$AAV-2$		0.633(1.503)	0.398(0.579)	0.388(0.558)			
		0.612(1.317)	0.347(0.483)	0.347 (--)			
		0.755 (-	0.714 (-)	0.714 (-			
	All	0.667 (--)	0.486(0.791)	0.483 (-			

TABLE 2. Proportion of nucleotide sites with different nucleotides between two compared species and estimated number of nucleotide substitutions in the conserved regions of H-1, MVM, AAV-2, and the Bombyx DNV

^a Values in parentheses are nucleotide substitutions and were estimated by the four-parameter method (32). —, inapplicable case because of too many nucleotide changes. Ninety-eight codons were compared in each case, and all gaps were excluded from the computation.

region may be important in biological functions (8, 15). Finally, based on the homologies of the regions, we constructed a phylogenetic tree for these four viruses, including the Bombyx DNV, by the unweighted pairwise group method (Fig. 6). For this purpose, we used the number of nucleotide substitutions at the first and second positions of a codon, because the nucleotide substitutions at the third position were often too numerous to be estimated by the four-parameter method (Table 2). The phylogenetic tree shows that the divergence order of the virus genomes is exactly the same as that of the host genomes. Thus, these viruses may represent a typical example of host-dependent evolution.

DISCUSSION

We determined the nucleotide sequence of 4,277 nucleotides comprising more than 85% of the complete Bombyx DNV genome. Our computer analysis showed that the Bombyx DNV genome possesses at least three major ORFs. Although the precise locations of splice junctions and the starting signals of transcription and translation remain unknown, we tentatively identified ORF2 as the coding region for all viral structural proteins. This conclusion is consistent with our previous observation that the peptide map shows considerable similarity among the amino acid sequences of

FIG. 6. The phylogenetic tree for the parvoviruses H-1, MVM, AAV-2, and the Bombyx DNV.

the four structural proteins (2). The difference in molecular weights between the structural proteins and those predicted from ORF2, however, suggests that the four structural proteins result from differential mRNA splicing and protein processing.

The rodent parvoviruses (H-1 and MVM) and the human dependovirus (AAV-2) have a common feature in their genomes. They have two large ORFs that are located separately in the right and left halves of the genome (1, 26, 31). All structural proteins of these mammalian parvoviruses are known to be coded by the ORF in the right half (ORF-R) of the genome (24, 31). If ORF-R of the mammalian parvoviruses corresponds evolutionarily to ORF2 of the Bombyx DNV, the ORF in the left half (ORF-L) of the genome of the mammalian parvoviruses may be related to ORF1 of the Bombyx DNV. Because the two nonstructural (NS) proteins NS-1 and NS-2 are coded by ORF-L in mammalian parvoviruses (7), ORFi may code for nonstructural proteins of the Bombyx DNV.

ORF3 of the Bombyx DNV has ^a coding capacity of ¹⁶⁷ amino acids. Because only four structural proteins are known to be produced by the Bombyx DNV, ORF3 may code for a nonstructural protein that has presently gone undetected. This ORF may allow ^a clear distinction between insect and mammalian parvoviruses, because only one of the viral DNA strands seems to be ^a sense strand in the mammalian parvovirus.

The sequence comparisons among the four virus genomes (H-1, MVM, AAV-2, and DNV) in the family Parvoviridae revealed that the Bombyx DNV has ^a sequence of ³⁰⁰ nucleotides that is homologous to sequences in the other three viruses. Although this conserved region is located in ORF2 of the Bombyx DNV, the corresponding regions of the mammalian parvoviruses are located in ORF-L. Searching the sequence signals of splice junctions in ORF2, we found possible acceptor and donor sites that were located near the beginning and the end of the conserved region of the Bombyx DNV (Fig. 2). Furthermore, ^a possible donor site also exists at the ³' end of ORFi and an acceptor site at the ³' end of ORF2. Thus, a potential splicing pattern could join the ⁵' end of ORFi with the conserved region of ORF2 in phase, then eliminate the bulk of the remaining coding sequence of ORF2, and then pick up a translational stop codon. Such a splicing gives rise to a protein that is equivalent in size to the larger of the mammalian parvovirus nonstructural proteins. Interestingly, there is convincing evidence to suggest that the NS-1 protein coded by ORF-L may be essential for early rounds of replication form replication (7, 19, 24) and transactivation of the promoter for the structural proteins (25). Thus, the active sites of the nonstructural protein may be located in the conserved region between the mammalian parvoviruses and the Bombyx DNV.

Although the significance of the conserved region still remains obscure, we tentatively conclude that the rodent parvoviruses (H-1 and MVM), a human dependovirus (AAV-2), and an insect densovirus (Bombyx DNV) share ^a common ancestor. The phylogenetic tree shows that the branching order of these viruses is the same as that of their hosts, suggesting a host-dependent evolution of these viruses (30). Such host-dependent evolution of the viruses may be easily explained if the viral DNAs are integrated into the host genome. In fact, it is known that AAV-2 DNAs are integrated into the human genome in latently infected cells (5, 11). Moreover, the viral genomes of the family Parvoviridae contain a palindromic sequence arrangement at both termini, although the fine structure varies with the virus. It is of particular interest to know whether terminal repeats exist in the Bombyx DNV DNA because most of the integrated viruses and transposons have long terminal repeats. It must be noted that the terminal organization of the genome of some other densoviruses appears to resemble that of the dependoviruses (13).

The rate of evolution of these DNA viruses is of interest, because it is not clear whether these virus genomes evolve at the same rate as the host genomes (9). If we use the divergence time (80 million years) between the orders Rodentia and Primates as that of the rodent parvoviruses H-1 and MVM and the human dependovirus AAV-2, the nucleotide substitution rate of these viruses can be estimated to be about 3.5×10^{-9} per year per site. This rate is almost comparable to that of eucaryotic genes such as hemoglobin (14). This conclusion must be taken with caution, however, because the estimation of the evolutionary rate depends on whether the assumption of host-dependent evolution is reasonable.

In view of the widespread distribution of severe parvovirus disease among domestic animals and insects, it is of considerable importance to understand the evolutionary relationships among the viruses in the family Parvoviridae. As shown by the results of our study, viruses found in invertebrates can be useful references in comparative studies of viruses found in vertebrates.

LITERATURE CITED

- 1. Astell, C. R., M. Thomson, M. Merchlinsky, and D. C. Ward. 1983. The complete DNA sequence of minute of mice, an autonomous parvovirus. Nucleic Acids Res. 11:999-1018.
- 2. Bando, H., N. Kondo, and S. Kawase. 1984. Molecular homology among the structural proteins of densonucleosis virus from silkworm. Bombyx mori. Arch. Virol. 80:209-218.
- 3. Benoist, C., K. O'Hare, R. Breathnach, and P. Chambon. 1980. The ovalbumin gene-sequence of putative control regions. Nucleic Acids Res. 8:127-142.
- 4. Carter, B. J., C. A. Laughlin, and C. J. Marcus-Sekura. 1984. Parvovirus transcription, p. 153-208. In I. B. Kenneth, and W. W. Hauswirth (ed.), Parvoviruses. Plenum Publishing Corp., New York.
- 5. Cheung, A. K.M., M. D. Hoggan, W. W. Hauswirth, and K. I. Berns. 1980. Integration of the adeno-associated virus genome into cellular DNA in latently infected human Detroit ⁶ cells. J. Virol. 33:739-748.
- 6. Corden, J., B. Wasylyk, and A. Buchwalder. 1980. Promoter

sequences of eukaryotic protein-coding genes. Science 209:1406-1414.

- 7. Cotmore, S. F., L. J. Sturzenbecker, and P. Tattersall. 1983. The autonomous parvovirus MVM encodes two nonstructural proteins in addition to its capsid polypeptides. Virology 129:333-343.
- 8. Go, M., and S. Miyazawa. 1980. Relationship between mutability and exteriority of amino acid residues in protein evolution. Int. J. Peptide Protein Res. 15:211-224.
- 9. Gojobori, T., and S. Yokoyama. 1985. Evolution rates of c- and v-mos genes. Proc. Natl. Acad. Sci. USA 82:4198-4201.
- 10. Grosschedl, R., and M. L. Birnsteil. 1980. Identification of regulatory sequence in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. Proc. Natl. Acad. Sci. USA 77:1432-1436.
- 11. Handa, H., K. Shiroki, and H. Shimojo. 1977. Establishment of characterization of KB cell lines latently infected with adenoassociated virus type 1. Virology 82:84-92.
- 12. Kawase, S. 1985. Pathology associated with densoviruses, p. 197-231. In K. Maramorosh and K.E Sherman (ed.), Viral insecticides for biological control. Academic Press, Inc., London.
- 13. Kelly, D. C., and H. M. Budd. 1978. Densonucleosis virus DNA: analysis of fine structure by electron microscopy and agarose gel electrophoresis. J. Gen. Virol. 40:33-43.
- Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature (London) 267:275-276.
- 15. Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111-120.
- 16. Kozak, M. 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 44:283-292.
- 17. Kurstak, E., and C. Vago. 1967. Transmission du virus de la densonucleose par le parasitisme d'un Hymenoptere. Rev. Can. Biol. 26:311-316.
- 18. Matthews, R. E. F. 1982. Fourth report of the International Committee on Taxonomy of Viruses: classification and nomenclature of viruses. Intervirology 17:1-199.
- 19. Merchlinsky, M. J., P. J. Tattersall, J. J. Leary, S. F. Cotmore, E. M. Gardiner, and D. C. Ward. 1983. Construction of an infectious molecular clone of the autonomous parvovirus minute virus of mice. J. Virol. 47:227-232.
- 20. Messing, J., and J. Vieira. 1982. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19:269-276.
- 21. Nakagaki, M., and S. Kawase. 1980. DNA of ^a new parvo-like virus isolated from the silkworm, Bombyx mori. J. Invertebr. Pathol. 35:124-133.
- 22. Nakagaki, M., and S. Kawase. 1980. Structural proteins of densonucleosis virus isolated from the silkworm, Bombyx mori, infected with the flacherie virus. J. Invertebr. Pathol. 36:166-171.
- 23. Nakagaki, M., and S. Kawase. 1982. Capsid structure of the densonucleosis virus of the silkworm, Bombyx mori. J. Seric. Sci. Jpn. 51:420-424.
- 24. Paradiso, P. R., K. R. Williams, and R. L. Costantino. 1984. Mapping of the amino terminus of the H-1 parvovirus major capsid protein. J. Virol. 52:77-81.
- 25. Rhode, S. L., III. 1985. trans-Activation of parvovirus P38 promoter by 76K noncapsid protein. J. Virol. 55:886-889.
- 26. Rhode, S. L., III, and P. R. Paradiso. 1983. Parvovirus genome: nucleotide sequence of H-1 and mapping of its genes by hybridarrested translation. J. Virol. 45:173-184.
- 27. Sanger, F., A. R. Coulson, B. G. Barrell, A. J. H. Smith, and B. A. Roe. 1980. Cloning in single stranded bacteriophage as an aid to rapid DNA sequencing. J. Mol. Biol. 143:161-178.
- 28. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. USA 74:5463-5467.
- 29. Siegl, G., R. C. Bates, K. I. Berns, B. J. Carter, D. C. Kelly, E. Kurstak, and P. Tattersall. 1985. Characteristics and taxonomy

of Parvoviridae. Intervirology 23:61-73.

- 30. Soeda, E., T. Maruyama, J. R. Arrand, and B. E. Griffin. 1980. Host-dependent evolution of three papova viruses. Nature (London) 285:165-167.
- 31. Srivastava, A., E. W. Lusby, and K. I. Berns. 1983. Nucleotide sequence and organization of the adeno-associated virus 2
- genome. J. Virol. 45:555-564. 32. Takahata, N., and M. Kimura. 1981. A model of evolutionary

base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641-657.

- 33. Watanabe, H., and S. Maeda. 1979. Multiplication of a densonucleosis virus in the silkworm. Bombyx mor L., reared at high temperature. Jpn. J. Appl. Entomol. Zool. 23:151-155.
- 34. Wilbur, W. J., and D. J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. Proc. Natl. Acad. Sci. USA 80:726-730.