Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses

Gregory Kamer and Patrick Argos*

Purdue University, Department of Biological Sciences, West Lafayette, IN 47907, USA

Received 14 May 1984; Revised and Accepted 4 September 1984

ABSTRACT

Possible alignments for portions of the genomic codons in eight different plant and animal viruses are presented: tobacco mosaic, brome mosaic, alfalfa mosaic, sindbis, foot-and-mouth disease, polio, encephalomyocarditis, and cowpea mosaic viruses. Since in one of the viruses (polio) the aligned sequence has been identified as an RNA-dependent polymerase, this would imply the identification of the polymerases in the other viruses. A conserved fourteen-residue segment consisting of an Asp-Asp sequence flanked by hydrophobic residues has also been found in retroviral reverse transcriptases, a bacteriophage, influenza virus, cauliflower mosaic virus and hepatitis B virus, suggesting this span as a possible active site or nucleic acid recognition region for the polymerases. Evolutionary implications are discussed.

INTRODUCTION

In the last five years the nucleotide sequence of genomic RNAs and DNAs have been determined for several animal, bacterial, and plant viruses. Table 1 lists several viruses which will be discussed in this report as well as references describing their codon sequences.

Franssen et al. (1) have recently observed homology in the amino acid sequences of the RNA-dependent RNA polymerases and segments of protein 2C (formerly known as P2-X (2,3)) from polio virus which is an animal picornavirus and from cowpea mosaic virus (CPMV) which is a plant comovirus. Argos et al. (4) have aligned the primary structures for the 2C protein, RNA polymerase, VPg protein, and the polyprotein proteinase from three picornaviruses (polio, foot-and-mouth disease (FMD), and encephalomyocarditis virus (EMC)) and CPMV. Haseloff et al. (5) have found similarities in the amino acid sequences for non-structural proteins encoded by the subgenomic RNA2 of brome mosaic (BMV) and alfalfa mosaic (AMV) viruses (plant bromo viruses) and by the genomic RNA of tobacco mosaic virus (RMV, a plant virus) and sindbis virus (SNBV, an animal alphavirus). They suggest that the homologous proteins are involved with RNA replication. Toh et al. (6) have observed sequence homology between

Table 1. A list of the viruses discussed in this report and references for their genome nucleotide sequence.

Virus Name	Abbreviation Used	Host	References
Tobacco Mosaic	TMV	plant	(24)
Brome Mosaic	BMV	plant	(25,26)
Alfalfa Mosaic	AMV	plant	(27-29)
Sindbis	SNBV	animal	(30)
Cowpea Mosaic	CPMV	plant	(38,39)
Polio	Polio	animal	(41-44)
Foot-and-Mouth Disease	FMD	animal	(22)
Encephalomyocarditis	EMC	animal	(23)
Rous Sarcoma	RSV	animal	(45)
Hepatitis B	HBV	animal	(46)
Maloney Murine Leukemia	MuLV	animal	(47)
Adult T-cell Leukemia	ATLV	animal	(35)
Phage MS2	MSV	bacteria	(34)
Influenza	FLU	animal	(31-33)
Cauliflower Mosaic	CaMV	plant	(48)

portions of retroviral reverse transcriptases (Maloney murine leukemia virus (MuLV), Rous sarcoma virus (RSV), and T-cell leukemia virus (ATLV)) and the putative RNA polymerases of hepatitis B virus (HBV) and cauliflower mosaic virus (CaMV). Though Toh et al. (6) were not able to align the sequences for the entire polymerases and transcriptases, they show an alignment for a 94-residue segment shared by the various viruses.

In the present report, possible identification of the RNA-dependent polymerases as well as alignment of their entire sequences in TMV, AMV, BMV, SNBV, FMD, EMC, polio, and CPMV is given. A conserved Asp-Asp sequence flanked by generally hydrophobic residues was found in all the putative polymerases and is suggested as the probable active and/or recognition site region. Within the central portion of the 94-residue segment relating reverse transcriptases or polymerases from HBV, RSV, MuLV, ATLV, and CaMV, there exists a 14-residue stretch that also conformed to the Asp-Asp region of the previously mentioned viral polymerases whose entire sequences were aligned. This putative active site segment was also found in the influenza virus (FLU) protein PA, suggesting its function as an RNA polymerase. Furthermore, a homologous active site span was observed in the RNA replicase of bacteriophage MS2 (MSV). Evolutionary implications are discussed through an analysis of the extent of homology and residue physical characteristic correlations of aligned amino acids.

METHODS

Searches for homologous amino acid sequences were performed by the method of Jukes and Cantor (7). In comparing two sequences, every amino acid span of length L residues from the first protein is aligned with all stretches of length L in the second protein. The total minimum base difference (MBD) for each of the possible oligopeptide alignments is determined by summing the minimum base change per codon (MBC/C) between paired amino acids in the aligned L-residue spans. The length L was chosen as 10 residues, a number which allows statistical significance and yet makes reasonable allowances for possible gaps. Significance is tested by calculating the ratio (P_{obs}/P_{calc}) for all possible MBD values resulting from a comparison of the two proteins. P_{obs} is the frequency with which a given MBD is observed in comparing all segments from the two proteins while P_{calc} is the expected frequency calculated from the amino acid compositions of the proteins compared. As the frequency ratio becomes increasingly larger than one, the significance of the homology becomes greater. The mean MBD value and associated standard deviation were also determined for each pairwise comparison of the several viral polymerases. Ten residue alignments were used that displayed a MBD value three or more standard deviations below the mean value.

After alignment of the proteins from the various viruses, certain characteristics of residues were selected to assess the degree of structural homology. The parameters included the experimental hydration potential of Wolfenden et al. (8); the bulk hydrophobic character, statistically determined by Manavalan and Ponnuswamy (9); a measure of hydrophobicity based on the amino acid mutability matrix of Dayhoff (10); the Chou-Fasman (11, 12) parameters for perference of alpha-helical, beta-strand, and reverse-turn configurations, as calculated by Palau et al. (13); and the residue polarity listed by Jones (14). These characteristics were selected as they represent the major forces thought to be required for proper protein folding (15, 16).

Three measures of hydrophobicity were used, as several have been calculated or empirically determined, but they do not necessarily correlate well (17). The measures of hydrophobicity represent an empirical and two theoretical attempts to quantify this character for the amino acids. The bulk hydrophobicities of Manavalan and Ponnuswamy were calculated by averaging the Nozaki-Tanford transfer-free energies (18) for residues surrounding a given amino acid type within known structures of soluble proteins. The values of Wolfenden et al. (8) result from measured vapor-water partition coefficients for model compounds corresponding to each of the amino acids. The values of Dayhoff were calculated from the relative frequency with which amino acids exchange in members of aligned primary structures in several protein families (10). The combination of all three measures of hydrophobicity, as used in this report, should provide an adequate sample of the possible measures.

After the sequences of the individual proteins from EMC, polio, FMD, CPMV, SNBV, TMV, AMV, and BMV had been aligned, they were compared pairwise, and the parameters for each residue were determined, and correlation coefficients (14) were calculated. These values should indicate the extent of structural and evolutionary relatedness amongst the three viruses (cf. 19).

RESULTS AND DISCUSSION

Sequence Alignment

The alignment of the major portion of the codons in AMV and BMV subgenomic RNA2s and codon segments of the genomic RNAs of SNBV and TMV has been reported by Haseloff et al. (5). In the present work a MBC/C search was performed on the aforementioned viral sequences and a strong homology was also found. The alignment presented here (Figure 1) is essentially similar to that of Haseloff et al. (5) except for segments about 40 residues long that relate the N- and C-terminal portions of SNBV to AMV, BMV, and TMV. The homologies in these spans were weaker and were aligned visually with the constraints of maintaining matched charged and hydrophobic residues described in the caption of Figure 1. The alignments of the RNA-dependent RNA polymerases for EMC, FMD, polio and CPMV were taken from Argos et al. (4) where, once again, the residue correlations were strong. In these two sets, each involving the alignments of sequences from four different viruses, a completely conserved Gly-Asp-Asp segment was visually observed in roughly the same location along the sequences from the N- to C-termini. This observation prompted a search to determine if the rest of the primary structures could be aligned thereby allowing the identifications of the RNA-dependent RNA polymerase in AMV, BMV, TMV, and SNBV.

A MBC/C search using a probe length of 10 residues was performed for all possible pairwise comparisons among the four-viral sequences in each of the two sets of aligned residues, resulting in a total of 16 pairwise searches. All aligned ten-residue spans that displayed a MBD value of seven or less were compiled. The smallest MBD value observed was four in a Gly-Asp-Asp segment, heretofore referred to as the GDD span. The Pobs/Pcalc ratio for MBD=7 averaged near 1.5 in all the pairwise comparisons while the lowest MBD=4 corresponded to a ratio near 25. The probability of detecting a seven MBD value was of the order of 10^{-4} which decreased to 10^{-6} for a MBD value of four. The mean MBD for each of the pairwise sequence comparisons was near 14 with a standard deviation of 2 such that MBD=7 was removed by 3.5σ from the mean while MBD=4 corresponded to 5.0σ. From the compiled list of segment alignments, several were chosen that would allow a contiguous alignment of the sequences from each of the four-viral sets. If one alignment was found between any two viruses in each of the sets, then the alignment for the remaining viruses in the region was implied by virtue of the excellent match of each of the four-viral sets. Figure 1, which displays the alignment of the sequences from all eight viruses, shows these "marker" spans through underscoring of the amino acid symbols. If more than two spans are underlined, then more than one pair of spans had a MBD value of seven or less. Such matched pairs were sufficient in number and sequence distribution to allow the visual alignment of the remaining residues sandwiched between the marker segments. Constraints involving the preservation of hydrophobicity, charge, and turn preference were applied in the visual alignment. Gaps were kept to a minimum. There were no marker segments for the C-terminal region (Figure 1); the alignments given there are suggestive.

Correlation coefficients were calculated between aligned residues over seven residue physical characteristics (mentioned in the Methods section) for each pairwise sequence comparison. The resulting mean correlation matrix is shown in Figure 2; the values are all positive and range from 0.63 to 0.12. The mean number of aligned residues over each pairwise comparison was 438. There are 171 alignment positions for which the amino acids display conservative characteristics (e.g. hydrophobicity, see caption of Figure 1) in six or more of the eight sequences, resulting in a 40% conservation.

Possible Identification of RNA Polymerases

Not all the terminal positions of the protein sequences of Figure 1 are known. For the polio virus, the primary structure given is known to be an RNA-dependent RNA polymerase; the terminal segments are composed of Gln-Gly residues which have been observed as polyprotein cleavage sites (2, 20, 21). The sequences for the FMD and EMC polymerases have been determined by homology with the polio genomic structure (22, 23). The CPMV polymerase has been inferred once again by homology with the three

```
TMV                        Q L Q - I D S V F - - - - - - - K G S N L F V A A P
AMV                  M T L G R - I I P T T - P V P T I R - - - - - D V F F S G
BMV        M N P I E H R V S R - V I D T - H C H P D N P - - - - - D I S - T G
SNBV   D Q P E C Y K I T Y P Q P L Y S S S - - V P A N Y S D P Q F A V A V C N
FMD                                      G L I V D T R D V E E - - R V H V M R
EMC                                      G A - - - L E R L P D G P R I H V P R
Polio                      G E I Q W M R P - - - - S K E V G Y - P I I N A P S
CpMV   G A E E Y F D F L P A E E N V S S G V A M - V A G L K Q G V Y I P L P T
                                                                     *   +

TMV    K T G - - D I S D M - Q F Y Y - - D K C L P G N S T M M N N F D A - V T
AMV    L S R H G S - P E V I Q N A L - - D E F L - P L H H S I D D L Y F Q E W
BMV    P I Y M E R - V S L A R T E A T S H S I L - P T H A Y F D D S Y H Q A L
SNBV   N Y L H E N Y P T V A S Y Q I T - D E Y - - D A Y L D M V D G T - - - -
FMD    K T K L - - A P T V A Y G V F N - P E F G - P A A L S N K D P R L N E G
EMC    K T A L R - - P T V A R Q V F Q - P A Y A - P A V L S K F D P R T E A D
Polio  K T K L E - - P S A F H Y V F E - G V K E - P A V L T K N D P R L K T D
CPMV   K T A L - - V E T P S E W H L D T P C D K V P S I L V P T D P R I P A Q
         +   +           + +     + *     +   +     *   + +       *     +

TMV    M R - L T D I S L N V K D C I - L D M S K - - - - S V P A A P K D Q I K
AMV    V E - - - - - - - T S D K S L D V D P K R - I D L S V F N N W Q S S E N
BMV    V E - - - - - - - N G D Y S M D F D R I R - L K Q S D V D W Y R D P D K
SNBV   V A C L D T A - - - - - - T F C P A K L - - - - R S Y P K K Y - - E Y R
FMD    V - V L D D V - - I F S K H K G A D K M T E E D K A L F R R C A A D Y -
EMC    V D E V - - - - - A F S K H T S - N Q E S L P P V F R M - - V A K E Y -
Polio  F E E A - - - - - I F S K Y V G - N K I T E V D E Y M - K E A V D H Y -
CpMV   H E G Y D P A K S G V S K Y S Q - P M S A L D P E L L - G E V A N D V -
       * +   +                 +         +                 +       +       +

TMV    P L I P M V R T A A E M P R Q T G L L - E N L V A M I K R N F N A P E L
AMV    C Y E P R F K T G A L S T R K - G T Q T E A L L A I K K R N M N V P N L
BMV    Y F Q P K M N I G S A Q R R V - G T Q T E N L T A L K K R N A D V P E M
SNBV   A P N I R S A V P S A M Q - - - N T L Q N V L I A A T K R N C N V T Q M
FMD    A S R L H S V L G T A - - - - - N A A P L S I Y E A I K G V D G L D A M
EMC    A N R V F T L L G K - - - - - - D N G R L T V K Q A L E G L E G M D P M
Polio  A G Q L M S L - - - - - - - - - N T E Q M C L E D A M Y G T D G L E A L
CpMV   L E L W H D C A V D W - - - - D D F G E V S L E E A L N G C E G V E Y M
       +                               *           *     * + +       * *     *

TMV    S G I I D I E N T A S L V V D K F F D S Y L L K E - - K R K P N K N V S
AMV    G Q I Y D V N S V A N S V V N K L L T T V I D P D - - K L C M - - F P D
BMV    G D A I N M K D T A K A I A K R F R S T F L N V D - G E D C L R A S M D
SNBV   R E L - P T L D S A T F N V E C F R K Y A C N D E Y W E E F A R K P I R
FMD    E P - - - - D T A P G L P W - - - - - - - - - - - - A L Q G K R R G A L
EMC    D R - - - - N T S P G L P Y - - - - - - - - - - - - T A L G M R R T - D
Polio  D - - - - L S T S A G Y P Y - - - - - - - - - - - - V A M G K K K R - D
CpMV   E R I - P L A T S E G F P H - - - - - - I L S R N G K E K G K R R F V Q
                       +     +   *                                 + +     +
```

Fig. 1

```
TMV     L F S R E S L N R W L E K Q E Q V T I G Q L A D F D F V D L P A V D Q Y
AMV     F I S E G E V S Y F Q D Y I V G K N P D P E L Y S D P L G V R S I D S Y
BMV     V M T K C L E - Y H K K - - W G K H M D L Q G V N - V A A E T D L C R Y
SNBV    I T T E F V T A Y V A R L K G P K A A A L F A - - - - - K T Y N L V P L
FMD     I D F E N G T V G - - - - - P E V E A A L K L - - - - - - M E K R E Y -
EMC     V D W E S A T L I - - - - - - - - F A A E R L R - - - - - M N E G D F S
Polio   I L N K Q T P - - - - - - - - - - R D T K E M Q - - - - K L L D T Y G I
CpMV    - G D D C V V S L I P G T - T V A K A Y E E L E - - - - A S A H R F V P
        +     *                             +       +

TMV     R H M Y K A Q P K Q K L D T S I - Q T E Y P A L Q T - - - - I V Y H S K
AMV     K H M I K S V L K P V E D N S L - H L E R P M P A T - - - - I T Y H D K
BMV     Q H M L K S D V K P V V T D T L - H L E R A V A A T - - - - I T F H S K
SNBV    Q E V P - - - M D R F V M D M K R D V - K V T P G T K H T E E R P K V Q
FMD     K F A - - - - C Q T F L K D E I R P M E K V - - - - - - - - R A G K T R
EMC     E V V - - - - Y Q T F L K D E L R P I E K V - - - - - - - - Q A A K T R
Polio   N L P - - - - L V T Y V K D E L R S K T K V - - - - - - - - E Q G K S R
CpMV    A L V - - - - G I E C P K D E K L P M R K V F - - - - - - - D K P K T R
          + +         +     + + + +   +     + + + +                         + *

TMV     K I N A I F G P L F S E L T R Q L L D S V D S S R F L F F T R K T P - A
AMV     D I V M S S S P I F L.A A A A R L M L I L R D K - I T I P S G K F H Q L
BMV     G V T S N F S P F F T A C F E K L S L A L K S R - F I V P I G K I S S L
SNBV    - V I Q A A E P L A T A Y L C G I H R E L V R R L T A V L L P N I H T L
FMD     - I V D V L P V E H I L Y T K M M I G R F C A Q M H S N N G P Q I G S A
EMC     - I V D V P P F E H C I L G R Q L L G K F A S K F Q T Q P G L E L G S A
Polio   - L I E A S S L N D S V A M R M A F G N L Y A A F H K N P G V I T G S A
CpMV    - C F T I L P M E Y N L V V R R K F L N F V - R F I M A N R H R L S C Q
          * +   +         +   + *   +   + +     *         +     +       +     +

TMV     Q I E D - - - - F F G D L D S H Y - - - P M D V L E L D I S K Y D K S -
AMV     F S I D A E - - A F D A S - - - - - - - - - H F K E I D F S K F D K S -
BMV     E L K N V R - - L N N R Y - - - - - - - - - - F L E A D L S K F D K S -
SNBV    F D M S A E - - D F D A I I A E H F K Q G D P V L E T D I A S F D K S -
FMD     V G C N P D - V D W Q R F G T H F A Q Y - R N V W D V D Y S A F D A N H
EMC     I G C D P D - V A W T A F G V A M Q G F - E R V Y D V D Y S N F D S T H
Polio   V G C D P D - L F W S K I P V L M E E K L - - F A - F D Y T G Y D A S L
CpMV    V G I N P Y S M E W S R L A A R M K E K G N D V L C C D Y S S F D G L L
        +   + +   +       +     +       +             * + + + * * +   *   * +
                                                                  *

TMV     Q N E F H C A V E Y E I W R R L G F E D F L G E V W K Q G - - H R K T T
AMV     Q N E L H H L I Q E R F L K Y L G I P N E F L T L W F N A - - H R K S R
BMV     Q G E L H L E F Q R E I L L A L G F P A P L T N W W S D F - - H R D S Y
SNBV    Q D D A M A L T G L M I L E D L G V D Q P L L D L I E C A - - F G E I S
FMD     C S D A M - - - N I M F E E V F R T D F G F H P N A E W I L K T L V N -
EMC     S - V A M - - F R L L A E E F F T P E N G F D P L T R E Y L E S L A I -
Polio   S - P A W - - F E A L - - K M V L E K I G F - G D R V D Y I D Y L N H -
CpMV    S K Q V M D V I A S M I N E L C G G E D Q L K N A R R N L L - M A C C -
            + * *     + +     +   + + *     +     *   + +     + +     +
TMV    L K D Y T - A G I K T C I W Y Q R K S G D V T T F I G N T V I I A A C L
AMV    I S D S K - N G V F F N V D F Q R R T G D A L T Y L G N T I V T L A C L
BMV    L S D P H - A K V G M S V S F Q R R T G D A F T Y F G N T L V T M A M I
SNBV   S T H L P - T G T R F K F G A M M K S G M F L T L F V N T V L N V V I A
FMD    T E H A Y - E N K R I T V E G G M P S G C S A T S I I N T I L N N I Y V
EMC    S T H A F - E E K R F L I T G G L P S G C A A T S M L N T I M N N I I I
Polio  S H H L Y C K N K T Y C V K G G M P S G C S G T S I F N S M I N N L I I
CpMV   S R H A I - K N T V W R V E C G I P S G F P M T V I V N S I F N E I L I
             +       +     +   *           * *     + *   *   * * * *     * * *
                                             *       *       *

TMV    A S M L - - - - - - - - - - - - - - - P M E K I I K G A F C G D D S L L
AMV    C H V Y D L M D P N - - - - - - - - - - - - - - - F V V A S G D D S L I
BMV    A Y A S D L S D - - - - - - - - - - - - C D - - - C A I F S G D D S L I
SNBV   S R V L E - - - - - - - - - - - - - E R L K T S R C A A F I G D D N I I
FMD    L Y A L R R H Y E G V E - - - - - - - - L D T Y T - M I S Y G D D I V V
EMC    R A G L Y L T Y K N F E - - - - - - - - F D D V K - V L S Y G D D L L V
Polio  R T L L L K T Y K G T D - - - - - - - - L D H L K - M I A Y G D D V I A
CpMV   R Y H Y K K L M R E Q Q A P E L M V Q S F D K L I G L V T Y G D D N L I
           + +                                 + +         + *   + * * *   * *
                                                                   * * *

TMV    Y F P K G C E F P D V Q H S A N L M W N F E A K L F K Q Y - - - - - G Y
AMV    G T V E - E L P R D Q E F L F T T L F N L E A K F P H N Q - - - - - P F
BMV    I S K V - K P V L D T D M - F T S L F N M E I K V M D P S V - - - - P Y
SNBV   H G V V S D K E M A E R - - C A T W L N M E V K I I D A V I G E R P P Y
FMD    A S D Y - - - - - - - - - - - - - - Y D L D F E A L K P H F K S L G Q T
EMC    A T N Y - - - - - - - - - - - - - - Y Q L D F D K V R A S L A K T G Y K
Polio  S Y P H - - - - - - - - - - - - - - H E V D A S L L A Q S G K D Y G L T
CpMV   S V N A - - - - - - - - - - V V T P Y - F D G K K L K Q S L A Q G G V T
             +                             * + * * + + + + + +

TMV    F C G R Y V I H H - - - - - D R G C I V Y Y D P L K L I S K L G A K H I
AMV    I C S K F L I T M P T T S G G K V V L P I P N P L K L L I R L G S K K V
BMV    V C S K F L V E T E M G N L V S - - - - V P D P L R E I Q R L A K R K I
SNBV   F C G G F I L Q D S - - - V T S T A C R V A D P L K R L F K L G K P L P
FMD    I T - P A D K S D K G F V L G Q S I T D V - - F L K R H F H M D Y - G T
EMC    I T - P A N T T S T F P L N S T L - E D V V - F L K R K F K - K E - - G
Polio  M T - P A D K S A T F E T V - T W - E N V T - F L K R F F R A D E K Y P
CpMV   I T D G K D K T S L E L P F R R L - E E C - D F L K R T F - V - Q R S S
       *       +                 +             *       * *   + + + +
                                                       *

TMV    K D W E H L E - E F R R S L C D V A V S L - - - N N C A Y Y T Q L D D A
AMV    N A D I F D E - - W Y Q S W I D I I G G F - - - N D H H V I R C V A A M
BMV    L R D E Q M L R A H F V S F C D R M K F I - - - N Q L D E K M I T T L C
SNBV   A D D E Q D E - D R R R A L L D E T K A W F R V G I T G - - - - - T L A
FMD    G F Y K P V M - - A S K T L - E A I L S F A R R G T I Q E K L I S V A G
EMC    P L Y R P V M - - N R E A L - E A M L S Y Y R P G T L S E K L T S I T M
Polio  F L I H P V M - P M K E I H - E S I R W T K D P R N T Q D H V R S L C L
CpMV   T I W D - A - - P E D K A S - L W S Q L H Y V N C N N L E K E V A Y L T
             +   +           +   +   +   +     +       +                   + +
TMV     V W E V H K T A - - - - P P G S F V Y - K S L V K Y L - - - - - - - - -
AMV     T A H R Y L R R P S L Y - E A A L E S L G K I F A G K T L C K E C L F N
BMV     H E V R Y L K Y G K E - K P W I F E E V - R A A L A A F S L Y S E C L N
SNBV    V A V T T - R Y E V D N I T P V L L A L R T - F A Q S K R A F Q A I R G
FMD     L A V H S G P D E - - - Y R R L F E P F Q - - G L F E I P S Y R S - - -
EMC     L A V H S G K Q E - - - Y D R L F A P F R E - V G V V V P S F E S - - -
Polio   L A W H N G E E E - - - Y N K F L A K I R S - V P I G - R A L L L - - P
CpMV    N V V N V L R E L Y M H S P R E A T E F R R K V L K K V S W I T S - - G
        + + +       +                 + *     +       +             +

TMV     - - - S D K V L F R S L F I D G S S C
AMV     E K H E S N V K I K P R R V K K S H S D A R S R A R R A
BMV     F L R F S D C Y C T E G I R V Y Q M S D P V C K F K R T
SNBV    E I K H L Y - - - - - - - - G G P K
FMD     - - - - L Y L R W V N A V C - G - - D A
EMC     - - - V E Y - R W R S L F W
Polio   E Y S T L Y R R W L D S F
CpMV    D L P T L A L - - L Q E F Y E Y Q R Q Q
                        +       +
```

Figure 1. Alignment of the amino acid sequences of the putative polymerase regions in eight viruses. Abbreviations for the viral names are given in Table 1. Underscored segments indicate spans used to align one or more of the top four viruses with one or more of the lower four. The conservation of residues is shown according to the following scheme.

(⁑) Residue conserved in all eight sequences.

(*) Residues at this position in all eight sequences are members of one of the following sets of residues: D,E,Q,N. (acidic and polar); K,R,D,E (charged); C,F,I,L,M,V,A,W,H,Y (hydrophobic); P,G,N,D (strong turn formers (2)).

(+) As (*) except conserved in six or seven of the eight sequences.

picornaviruses (1, 4); the N-terminal Gly is flanked by a Gln residue and therefore provides a suspected cleavage site by picornaviral analogy (4). It is possible that the C-terminus of the putative CPMV polymerase shown in Figure 1 contains a further 176 residues as discussed by Argos et al. (4). In TMV the terminal residues given correspond to known protein termini (24). In BMV the C-terminal amino acid shown is the 3'-terminal codon of RNA2; the N-terminal Met is codon 194 from the 5' end of BMV RNA2 (25, 26). It is possible that the extra N-terminal residues are part of the BMV polymerase suggested here or represent another protein of unknown function. The situation is similar for the AMV RNA2 where a further 262 codons exist at the 5' terminus (27, 28, 29). However, it is noteworthy that the homologies ended in the vicinity of a Met in both AMV and BMV. For SNBV the C-terminal residue shown in Figure 1 corresponds to the 3'-terminal codon of the first open reading frame in the SNBV genome while the N-terminal Asp given is 97 residues removed from the suspected N-terminus of the fourth SNBV non-structural protein (30).

	EMC	FMD	Polio	CPMV	SNBV	BMV	AMV	TMV
EMC	XX	63	54	42	29	19	19	18
FMD	63	XX	47	35	31	22	17	14
Polio	54	47	XX	39	25	15	14	12
CPMV	42	35	39	XX	26	16	13	14
SNBV	29	31	25	26	XX	34	30	25
BMV	19	22	15	16	34	XX	45	34
AMV	19	17	14	13	30	45	XX	36
TMV	18	14	12	14	25	34	36	XX

Figure 2. Mean correlation coefficients (x100) of seven residue characteristics for aligned amino acids in a given protein pair. A random correlation would be 0.00. The symmetric matrix is given in its entirety for ease of comprehension. Viral names are abbreviated as listed in Table 1.

It is suggested that the sequences represent all or a major part of the primary structures of the RNA-dependent polymerases in the eight viruses. Only in polio virus has the polymerase been definitely assigned. Given the positive correlations of residue characteristics, the even distribution of "marker" spans, and the statistical significance of MBD values seven or less, it is certainly possible that the sequences are RNA-dependent RNA polymerases and share a similar tertiary fold.

Toh *et al*. (6) were able to align 94-residue spans in the reverse transcriptases of MuLV and RSV with putative polymerases of CaMV and HBV. For the individual pair, MuLV and CaMV, they were further able to match about 330 residues which contain the 94-residue segment with a Tyr-Val-Asp-Asp (YVDD) sequence near position 181 of the RSV reverse transcriptase. In the eight sequences shown in Figure 1, the GDD span occurs at approximately position 350 which roughly corresponds to the Asp-Asp sequence position in CaMV, HBV, and MuLV. The 350 position is considerably removed from the 181 site in RSV. Nonetheless the Asp-Asp segments flanked by about six hydrophobic residues on either side are all clearly homologous for all the viruses (Figure 3). The Asp-Asp sequence in three further viruses has also been observed; position 478 of the PA polypeptide of influenza virus (31, 32, 33); position 340 of the RNA-dependent RNA polymerase (β chain) of bacteriophage MS2 (34); and position 189 of ATLV (35) as also noted by Toh *et al*. (6). These segments are also listed in

```
TMV     I K G A F C G D D S L L Y F
AMV     N F V V A S G D D S L I G T
BMV     D C A I F S G D D S L I I S
SNBV    R C A A F I G D D N I I H G
FMD     Y T M I S Y G D D I V V A S
EMC     V K V L S Y G D D D L L V A
Polio   L K M I A Y G D D V I A S Y
CPMV    I G L V T Y G D D N L I S V
ATLV    C T I L Q Y M D D I L L A S
MSV     G T I G I Y G D D I I C P S
FLU     N A S C A A M D D F Q L I P
HBV     C L A F S Y M D D V V L G A
RSV     L C M L H Y M D D L L L A A
MuLV    L I L L Q Y V D D L L L A A
CaMV    K F C C V Y V D D I L V F S
```

Figure 3. Alignment of regions for several viruses around the Asp-Asp sequence. Viral names are abbreviated as listed in Table 1. The hydrophobic character of residues flanking the Asp-Asp pair is evident.

Figure 3. This consistent homology may point to a common nucleic acid recognition site in the various polymerases and/or to an active processing region.

Attempts to align the entire polymerase sequences of RSV, HBV, CaMV, MuLV, FLU, MSV and ATLV are presently underway, both by the MBC/C criteria or by the physical characteristics of the amino acids which have been suggested as more sensitive criteria for residue alignment (36). There are many examples of known protein tertiary architectures which display similar folding patterns and active sites and yet a random MBC/C between amino acids whose side chains are associated with spatially equivalenced C_α positions (cf. 37). Though the position of the Asp-Asp sequence in the RSV and ATLV (pol) gene product would appear too N-terminal as compared with the other viruses, these two viruses perhaps possess only a recognition domain.

Evolution

Are all the viruses mentioned here related by divergent evolution; i.e., have they derived from a common ancestral virus? The authors "feel" that the answer to the question is yes, an affirmation that is far from proven. The pros and cons of the query's answer will be subsequently discussed.

Figure 2 shows the average correlation coefficient over seven residue physical characteristics for each pairwise comparison of the putative polymerases for eight viruses. It is clear that FMD, polio and EMC are strongly related as they should be, given their many similar properties that classify them as picornaviruses: a single genome that is encapsidated by an icosahedrally symmetric protein capsid composed of four coat protein types and that is similarly organized in protein coding and function (see (4) for a discussion). CPMV, a plant virus, is best related to the three animal picornaviruses. Though CPMV has a divided genome and requires two particles for infection,each containing a subgenome with one (M RNA) coding for structural proteins and the other (B RNA) for non-structural proteins (38, 39), its B RNA is similarly organized in protein coding and function as the picornaviruses (see (1) and (4) for a discussion). It would appear then that CPMV, FMD, EMC, and polio virus are divergently related, allowing for genetic recombination and separation. BMV and AMV display a close correlation, again as expected from their viral properties such as requiring several icosahedral particles for infection, each made up of the same capsid protein and each containing one of three different but similarly-sized RNA subgenomes (see (5) for a discussion). TMV is best related to AMV and BMV and displays about the same correlation to these latter two viruses as does CPMV to the picornaviruses. TMV, however, uses a single genome encapsidated in a rod-shaped cluster of identical copies of coat protein. Haseloff et al. (5) have observed that the codons of AMV and BMV RNA1 are also homologous to the 5'-most portion of the TMV RNA and that the AMV and BMV RNA2 codons are homologous with a TMV segment just following the region homologous with RNA1. The 3'-most part of the TMV genome codes for the coat protein as does RNA3 of AMV and BMV. Once again, assuming a facile ability for genetic recombination, it would appear that TMV, AMV, and BMV are divergently related as also suggested by Haseloff et al. (5). SNBV appears about as closely related to CPMV and the picornaviruses as it does to TMV and the bromoviruses with a somewhat favorable correlation with BMV. The Sindbis virion contains a single genome of two open reading frames with the 5'-most portion coding for non-structural proteins and the 3'-most handling the structural proteins. SNBV also requires polyprotein processing as the picornaviruses and CPMV. If all the viruses proceeded from a common ancestor, SNBV would provide the link between the probable divergent groups.

The reverse transcriptases are another matter since their homology with each other in the (pol) gene products and with the other viruses is presently observed to be limited. If the residue physical characteristic correlations prove significant such that their entire polymerase sequences can be aligned with all the other viruses, then perhaps divergent evolution will be plausible as juxtaposed to the convergent evolution of active site processing (cf. 40). Nonetheless, it is possible that all the viruses share an active and/or recognition site important in copying an RNA template into DNA or an oppositely-stranded RNA.

The alternative evolutionary pathway for viral development is a convergent one. The viruses could have independently evolved from their host cells taking those genes that are amenable for transfer and necessary for viral replication. The sequence similarities observed here may thus be a result of common viral transfer mechanisms and the structural and functional constraints on host cell proteins as RNA-dependent polymerases.

Hopefully the alignments of Figures 1 and 3 will be useful in discovering the polymerase sequences of still further viruses.

ACKNOWLEDGEMENTS

The authors are grateful for considerable computing assistance, both in cost and advice, from the Purdue University Cyber 205 Computer Center.

*To whom reprint requests should be sent

REFERENCES

1. Franssen, H., Leunissen, J., Goldbach, R., Lomonosoff, G., and Zimmern, D. (1984) EMBO Journal 3, 855-861.
2. Flanegan, J.B., and Baltimore, D. (1977) Proc. Natl. Acad. Sci. USA 74, 3677-3680.
3. Rueckert, R.R., and Wimmer, E. (1984) J. Virol. 50, 957-959.
4. Argos, P., Kamer, G., Nicklin, M.J.H., and Wimmer, E. (1984) Nucl. Acids Res., preceding paper.
5. Haseloff, J., Goelet, P., Zimmern, D., Ahlquist, P., Dasgupta, R., and Kaesberg, P. (1984) Proc. Natl. Acad. Sci. USA, in press.
6. Toh, H., Hayashida, H., and Miyata, T. (1983) Nature 305, 827-829.
7. Jukes, T. H., and Cantor, C.R. (1969) In Mammalian Protein Metabolism (ed. Munro, H.N.) Vol. III, pp. 21-132 Academic Press, New York.
8. Wolfenden, R.V., Cullis, P.M., and Southgate, C.C.F. (1979) Science 206, 575-577.
9. Manavalan, P., and Ponnuswamy, P.K. (1978) Nature 275, 673-674.
10. Sweet, R.M., and Eisenberg, D. (1983) J. Mol. Biol. 171, 479-488.
11. Chou, P.Y., and Fasman, G.D. (1974) Biochemistry 13, 211-221.
12. Chou, P.Y., and Fasman, G.D. (1974) Biochemistry 13, 222-245.
13. Palau, J., Argos, P., and Puigdomenech, P. (1982) Int. J. Peptide Protein Res. 19, 394-401.
14. Jones, K.K. (1975) J. Theor. Biol. 50, 167-183.
15. Creighton, T.E. (1978) Biophys. Mol. Biol. 33, 231-297.
16. Ghelis, C., and Yon, J. (1982) In Protein Folding, pp. 136-176. Academic Press, New York.
17. Argos, P., and Palau, J. (1982) Int. J. Peptide Protein Res. 19, 380-393.
18. Nozaki, Y., and Tanford, C. (1971) J. Biol. Chem. 246, 2211-2217.
19. Keim, P., Heinrikson, R.L., and Fitch, W.M. (1981) J. Mol. Biol. 151, 179-197.
20. Hanecak, R., Semler, B.L., Anderson, C.W., and Wimmer, E. (1982) Proc. Natl. Acad. Sci. USA 79, 3973-3977.
21. Semler, B.L., Hanecak, R., Anderson, C.W., and Wimmer, E. (1981) Virology 114, 589-594.

22. Carrol, A.R., Rowlands, D.J., and Clarke, B.E. (1984) Nucl. Acids Res. 12, 2461-2472.
23. Palmenberg, A.C., Kirby, E.M., Janda, M.R., Drake, N.L., Duke, G.M., Potratz, K.F., and Collett, M.S. (1984) Nucl. Acids Res., in press.
24. Goelet, P., Lomonosoff, G.P., Butler, P.J.G., Akam, M.E., Gait, M.J. and Karn, J. (1982) Proc. Natl. Acad. Sci. USA 79, 5818-5822.
25. Ahlquist, P., Dasgupta, R., and Kaesberg, P. (1984) J. Mol. Biol. 172, 369-383.
26. Ahlquist, P., Luckow, V., and Kaesberg, P. (1981) J. Mol. Biol. 153, 23-28.
27. Cornelissen, B., Brederode, E., Mooreman, R., and Bol, J. (1983) Nucl. Acids Res. 11, 1253-1265.
28. Cornelissen, B., Brederode, E., Veeneman, G., van Boom, J., and Bol, J. (1983) Nucl. Acids Res. 11, 3019-3025.
29. Barker, R., Jarvis, N., Thompson, D., Loesch-Fries, L., and Hall, T. (1983) Nucl. Acids Res. 11, 2881-2891.
30. Strauss, E.G., Rice, C.M., and Strauss, J.H. (1984) Virology 133, 92-110.
31. Fields, S., and Winter, G. (1982) Cell 28, 303-313.
32. Bishop, D.H.L., Jones, K.L., Huddlestone, J.A., and Brownlee, G.G. (1982) Virology 120, 481-489.
33. Robertson, J.S., Robertson, M.E.S.C., and Roditi, I.J. (1984) Virus Res. 1, 73-79.
34. Fiers, W., Contreras, R., Duernick, F., Haegeman, G., Iserentant, D., Merregaert, J., MinJou, W., Molemans, F., Raeymaekers, A., van den Berghe, A., Volckaert, G., and Ysebaert, M. (1976) Nature 260, 500-507.
35. Seiki, M., Hattori, S., Hirayama, Y., and Yoshida, M. (1983) Proc. Natl. Acad. Sci. USA 80, 3618-3622.
36. Argos, P., Hanei, M., Wilson, J.M., and Kelley, W.N. (1983) J. Biol. Chem. 10, 6450-6457.
37. Rossmann, M.G., and Argos, P. (1978) Mol. Cell. Biochem. 21, 161-182.
38. van Wezenbeek, P., Verver, J., Harmsen, J., Vos, P., and van Kammen, A. (1983) EMBO Journal 2, 941-946.
39. Franssen, H., Moerman, M., Rezelman, G., and Goldbach, R. (1984) J. Virol. 50, 183-190.
40. Argos, P., Garovito, R.M., Eventoff, W., Rossmann, M.G., and Branden, C.L. (1978) J. Mol. Biol. 126, 141-158.
41. Kitamura, N., Semler, B L., Rothberg, P.G., Larsen, G.R., Adler, C.J., Dorner, A.J., Emini, E.A., Hanecak, R., Lee, J.J., van der Werf, S., Anderson, C.W., and Wimmer, E. (1981) Nature 291, 541-553.
42. Racaniello, V.R., and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA 78, 4887-4891.
43. Nomoto, A., Omata, T., Toyoda, H., Kuge, S., Horie, H., Kataoka, Y., Genba, Y., Nakano, Y., and Imura, N. (1982) Proc. Natl. Acad. Sci. USA 79, 5793-5797.
44. Stanway, G., Cann, A. J., Hauptman, R., Hughes, P., Clarke, L.D., Mountford, R.C., Minor, P.D., Schild, G.C., and Almond, J.W. (1983) Nucl. Acids Res. 11, 5629-5643.
45. Schwarts, D.E., Tizard, R., and Gilbert, W. (1983) Cell 32, 853-869.
46. Ono, Y., Onda, H., Sasada, R., Igarashi, K., Sugino, Y., and Nishioka, K. (1983) Nucl. Acids Res. 11, 1747-1757.
47. Shinnick, T.M., Lerner, R.A., and Sutcliffe, J.G. (1981) Nature 293, 543-548.
48. Gardner, R.C., Howarth, A.J., Hohn, P., Brown-Luedi, M., Shepherd, R.J., and Messing, J. (1981) Nucl. Acids Res. 9, 2871-2888.