Volume 17 Number 12 1989 Nucleic Acids Research

Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis

Alexander E.Gorbalenya*, Eugene V.Koonin, Alexei P.Donchenko and Vladimir M.Blinov

Institute of Poliomyelitis and Viral Encephalitides, USSR Academy of Medical Sciences, 142782 Moscow region, USSR

Received January 6, 1989; Revised and Accepted May 22, 1989

ABSTRACT

Amino acid sequences of 2 giant non-structural polyproteins (F1 and F2) of infectious bronchitis virus (IBV), a member of Coronaviridae, were compared, by computer-assisted methods, to sequences of a number of other positive strand RNA viral and cellular proteins. By this approach, juxtaposed putative RNA-dependent RNA polymerase, nucleic acid binding ("finger"-like) and RNA helicase domains were identified in F2. Together, these domains might constitute the core of the protein complex involved in the primer-dependent transcription, replication and recombination of coronaviruses. In F1, two cysteine protease-like domains and a growth factor-like one were revealed. One of the putative proteases of IBV is similar to 3C proteases of picornaviruses and related enzymes of como-, nepo- and potyviruses. Search of IBV F1 and F2 sequences for sites similar to those cleaved by the latter proteases and intercomparison of the surrounding sequence stretches revealed 13 dipeptides Q/S(G) which are probably cleaved by the coronavirus 3C-like protease. Based on these observations, a partial tentative scheme for the functional organization and expression strategy of the non-structural polyproteins of IBV was proposed. It implies that, despite the general similarity to other positive strand RNA viruses, and particularly to potyviruses, coronaviruses possess a number of unique structural and functional features.

INTRODUCTION

Coronaviruses are enveloped positive strand RNA viruses having by far the largest genome in this virus class (1-3). Recently, the genome sequence of the type member of Coronaviridae, avian infectious bronchitis virus (IBV), has been completed (4). The total length of IBV genome is 27 608 nucleotides, excluding 3'-terminal poly(A). Of these, about 8 000 nucleotides at the 3'-end are dedicated to coding virion and some small non-structural proteins, expressed as a nested set of 3'co-terminal mRNAs, with only the 5'-terminal "unique" part probably translated in each (2). The 5'-terminal part of genomic RNA (approx. 20 000 nucleotides) contains two large ORFs, potentially encoding two non-structural polypeptides (F1 and F2) of 441 and 300 kD, respectively. As no subgenomic mRNA corresponding to the F2 polypeptide has been detected, it was

suggested that the two ORFs are expressed as a single giant polyprotein, via ribosome frame-shifting (4). Subsequently, experimental evidence has been obtained corroborating this hypothesis (5).

Functional organization of the F1-F2 polyprotein of IBV remained, until very recently, completely obscure. Only a short region of F2 has been shown to possess a considerable similarity to non-structural proteins of alphaviruses and certain plant viruses (4). We demonstrated that this segment in fact comprised a part of a domain containing an NTP-binding sequence motif and belonging to a vast superfamily of positive strand RNA viral proteins in which this motif is the most conserved sequence (6). Moreover, it has been shown that one of the three protein families constituting this superfamily, the IBV domain included, possessed highly significant sequence similarity to DNA helicases (7-9). We suggested that proteins of this family could be RNA helicases involved in duplex unwinding during viral RNA replication (7,8). Encouraged by these observations, we performed a systematic search of the sequences of the large non-structural polypeptides of IBV for sequence stretches similar to highly conserved proteins of positive strand RNA viruses and to certain cellular proteins. Here we report the results of this study and discuss implications for functional organization and expression strategy of IBV genome.

METHODS

Amino acid sequence comparisons

Amino acid sequences were from current literature; for abbreviations and references see legends to figures. Comparisons were done by programs MULDI (MULtiple DIagon) and OPTAL (OPTimal ALignment). Program MULDI is a modification of standard DIAGON (10) designed to reveal highly conserved segments in amino acid sequences. Groups of aligned amino acid sequences are compared in a diagonal plot, utilizing the MDM78 amino acid residue comparison matrix (10). What results, may be considered a superposition of several pairwise local similarity maps in which only streaks corresponding to highly conserved segments are filtered out. MULDI is principally similar to the program recently described by Argos (11). Program OPTAL (6, 12), based on the original algorithm of Sankoff (13), performs stepwise optimal alignment of multiple amino acid sequences and its statistical assessment by a Monte Carlo procedure. Adjusted alignment score is calculated in standard deviation (SD) units: $AS = S_O - S_r/\sigma$ where S_O is the score obtained for a given comparison utilizing MDM78 scoring matrix, S_r is the mean score obtained upon intercomparison of 25 randomly jumbled sequences (or sequence sets) identical to the real ones in amino acid composition, and σ is the standard deviation. The programs were written in FORTRAN77 and run on a ES-1060 computer. The statistical significance of manual alignments was assessed by program SCORE. Average per residue score was computed for a query sequence versus a group of aligned sequences and AS was calculated by the above equation using 300 randomly scrambled versions of the query sequence (E.V.K. et al, in preparation). The probabilitiy of chance similarity between

two sequences aligned without gaps ('double matching probability') was calculated using the algorithm of McLachlan (14).

RESULTS AND DISCUSSION

Approach

As the first step to identification of functional domains in coronavirus polyproteins, it was natural to try to find coronaviral counterparts of the most highly conserved proteins of positive strand RNA viruses. Such proteins are, in the order of decreasing conservation: i) RNA-dependent RNA polymerases present in all viruses of this class and always having a similar central segment (15,16); ii) NTP-binding motif-containing proteins involved in RNA replication some of which are similar to helicases; proteins of this type were identified in all eukaryotic positive strand RNA viruses whose genome lengths exceed 6.3 kb [(6-9) and manuscript in preparation]; iii) 3C proteases of picornaviruses and similar enzymes revealed in como-, nepo- and potyviruses (17-23). Clearly, at least for the first and the second groups of enzymes, the case for existence of coronaviral homologs seemed quite strong.

Alignments of conserved fragments of these three groups of viral proteins were used as probes to screen sequences of F1 and F2 polypeptides of IBV by program MULDI. Segments of these proteins best matching the probes were fitted into respective alignments by program OPTAL (or visually) and the significance of the observed similarity was correspondingly assessed. Additional search by the same procedure was made for segments of coronaviral proteins similar to different classes of cellular proteases and to certain other sequence motifs conserved in cellular proteins. Identification of the putative helicase was described previously (see Introduction); other results are presented below.

RNA-dependent RNA polymerase

In F2 polypeptide two segments similar to the two most conserved sequence blocks of (putative) positive strand RNA viral RNA polymerases were detected. Inspection of the neighboring regions of F2 revealed also putative counterparts of other conserved stretches of polymerases. As can be seen in the resulting alignment (Fig.1), this part of F2 contained all the amino acid residues invariant in other viral polymerases, except one, as well as many partially conserved residues. A notable exception is the substitution of S for G in the so called GDD site considered to be the most characteristic sequence of positive strand RNA viral RNA polymerases (15,16, 22). Presumably, it was this substitution that prevented other investigators from identification of the IBV polymerase. Evaluation of the alignment of the 4 picked segments of F2 with the conserved segments of 40 (putative) polymerases of positive strand RNA viruses by program SCORE showed significance at the 9.2 SD level. Lengths of variable spacers separating conserved fragments in the putative polymerase of IBV are generally within the limits set by other polymerases although the coronavirus one appears to be among the longest. Unexpectedly, a 19 amino acid residue segment of F2 has been shown to possess

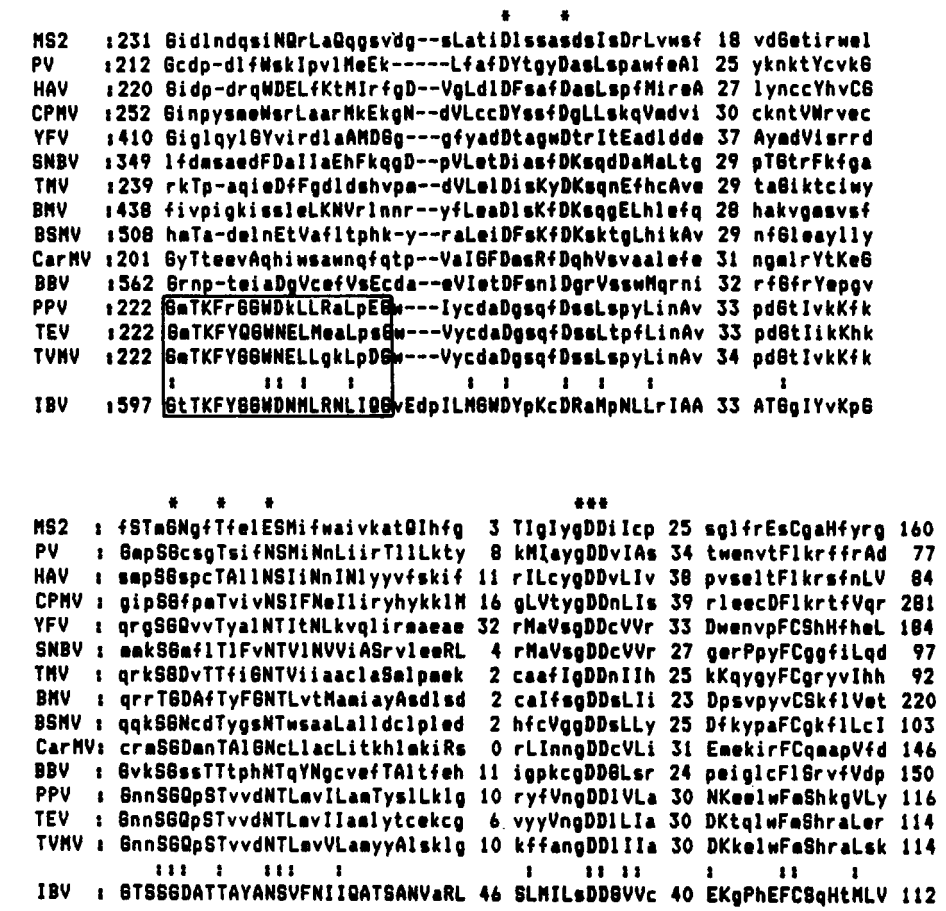

Fig.1. Alignment of a fragment of putative RNA-dependent RNA polymerase of IBV with evolutionary conserved fragments of selected (putative) polymerases of other positive strand RNA viruses.
The sampling of the (putative) polymerases was compiled so as to represent the main groups of positive strand RNA viruses and the entire range of sequence variability of this protein (cf.16). Abbreviations: MS2, MS2 bacteriophage; PV, poliovirus type 1, HAV, hepatitis A virus (picornaviruses); CPMV, cowpea mosaic virus (a comovirus); YFV, yellow fever virus (a flavivirus); SNBV, Sindbis virus (an alphavirus); TMV, tobacco mosaic virus (a tobamovirus); BMV, brome mosaiv virus (a tricornavirus); BSMV, barley stripe mosaic virus (a hordeivirus); CarMV, carnation mottle virus; BBV, black beetle virus (a nodavirus); PPV, plum pox virus, TEV, tobacco etch virus, TVMV, tobacco vein mottling virus (potyviruses). The lengths of the terminal regions and of the variable spacers separating the conserved segments are designated by numbers. For IBV, the bounderies of the polymerase were predicted from analysis of the putative

a very remarkable similarity to a segment of RNA polymerases of potyviruses which is relatively variable among positive strand RNA viruses in general (Fig.1). For this segment, the similarity between IBV and the potyviruses is comparable to that between potyviruses themselves, and unprecedented for positive strand RNA viruses of different families. Taken together, these observations strongly suggest that the pinpointed region of F2 is the core domain of IBV RNA-dependent RNA polymerase. As for the aforementioned substitution in the 'GDD box', it is relatively conservative in nature and, more importantly, includes a residue which obviously plays a structural, rather than catalytic, role. It is perhaps relevant that polymerases of MS2 and related phages, for which the activity had been firmly established, also bear a substitution of an otherwise conserved residue, i.e. Glu for Asn (cf. Fig.1).

Two types of RNA-synthesizing complexes greatly differing with respect to enzymatic properties and products synthesized were isolated from coronavirus-infected cells (24). Also, coronaviruses are known to have a unique mechanism of subgenomic RNA synthesis quite distinct from that of genome replication (3). Thus, it is not unlikely that IBV could have more than one RNA polymerase. However, our search did not reveal any segments of F1 or F2 significantly similar to viral polymerases except that shown in Fig.1; though some sequences of marginal similarity could be detected in C-terminal parts of both polyproteins. Thus, if IBV genome encodes a 2nd RNA polymerase, its sequence should be very different from those of other positive strand viral polymerases.

3C-like protease

In F1 polypeptide, sequence stretches similar to all three conserved segments of 3C-like proteases (19) were detected. Alignment of a 188 residue piece of F1 with 14 viral proteases proved to be significant at the 5.7 SD level. Notably, His, Asp(Glu) and Cys residues conserved in 3C-like proteases and thought to constitute their catalytic triad (19) were identified also in the coronavirus sequence (Fig.2). The putative coronavirus protease contains one replacement of a residue invariant in other 3C-like proteases. This is the substitution of Tyr for Gly in the sequence GXH in the vicinity of the proposed catalytic Cys residue (Fig.2). It is notable that, just like the replacement in the putative polymerase

Fig.1 legend cont.....

cleavage sites (see text and Fig.6); the sequence shown is residues 549 to 780 of the F2 polypeptide (4). The PPV sequence is from (39), and the BSMV one from (40). For sources of the other sequences see (16). Capitals: residues identical or similar to respective residues of IBV; colons: positions where residues identical or similar to those of IBV are observed in more than a half of included sequences. Residues belonging to one of the following groups were regarded similar: L,I,V,M,; A, G; S,T; D,E,N,Q; K,R; F,Y,W. Asterisks: consensus residues of positive strand RNA viral polymerases (15,16,22). Boxed: region of high local similarity between putative polymerases of IBV and potyviruses.

```
      Ref.                        *     *                *
PV    (41): 24 ftMlGV-hdNvailPtH-29-lEITiitlkrnE-62-AGqCGg-vitct-G---kvigMH-Vgg  19
HRV   (42): 24 ftglGV-ydrfVvvPtH-29-lEITvlkldrnE-62-sGyCGg-vlyki-G---QvlgIH-Vgg  19
EMCV  (43): 32 Qtcl1V-rGrTLvvnRH-32-tDVSfirlsSgp-65-kGwCGSalladl-GgskkilgIH-sag  25
FMDV  (44): 32 ccatGV-fGtaylvPRH-36-sDaalmvlMrgN-65-AGyCGgavlakD-Gadtfivgth-sag  29
HAV   (45): 32 mNalGV-kdDwLlvPsH-38-qDVvlmkvpTIp-74-pGmCGGalvssNqsIqnailgIH-Vag  23
CPMV  (46): 24 lQIvmV-pGrrflacKH-34-sELvlysmpSLE-71-pedCGSLviahi-GgkhkivgVH-Vag  21
TBRV  (21): 22 vsamqy-knkSVrmtRH-36-sEIvTwlApSLp-73-nddCGmIilcqi-kgkmrvvgMl-Vag  19
TEV   (33):217 tsLyGIgfGpfIitnKH-34-rDMiiirmpkd--56-dGqCGSplvstrdG---fivgIHsasn  71
                  ::  :  :  :::     ::        :     : ::        :       :: :
IBV    (4): 24 NNLnGLwLGDTIycPRH-21-fEVTTqhGVTLN-65-AGaCGSVgfniEkGVv-NffyMHhLel 142
```

Fig.2. Alignment of a fragment of putative 3C-like protease of IBV with conserved fragments of selected cysteine proteases of other positive strand RNA viruses.
The representative sampling of (putative) proteases was generated as indicated in the legend to Fig.1. Additional abbreviations: HRV, human rhinovirus type 2; EMCV, encephalomyocarditis virus, FMDV, foot-and-mouth disease virus (picornaviruses); TBRV, tomato black ring virus (a nepovirus). The boundaries of the putative protease of IBV were predicted as indicated in the legend to Fig.1; the sequence shown is residues 2804 to 2945 of the F1 polypeptide (4). Source references for the other sequences are given in parentheses before each sequence. Asterisks: putative catalytic residues; other designations as in Fig.1.

```
                                                                  *
SP  QPVVKSLLDSKGIHYNQGNPYNLL-------TPVIEKVKPGEQSFVGQAATGHCVATATAQIMKYHNYPDKGLK
          .      .   :: ::       :   .   :   :::   .:::   :  :..      :  :
IBV SNCPTCGANNTDEVIEASLPYLLLFATDGPATVDCDEDAVGTVVFVGSTNSGHCYTQAAGQAFDNLAKDRKFGK

SP  NYTYTLSSNPDYFDHPKNLFAAISTRQYDWNNILPTYS----GRQSQNVKMAISELMADVGISVDMDYGPSSGS
      :      .                    .   :          .   :   :   :   .     :
IBV KSPYITAMYTRFAFKNETSLPVAKQSKGKSKSVKEDVSNLATSSKASFDNLTDFEQWYDSNIYESLKVQESPDN

SP  AG-----SSRVQRALKENFGYNQSVHQINRGDFSKQDWEAQIDKELSQNQPVYY------------EGVGK-V
          ... .  :            ..             : .     ::::             :: .  :
IBV FDKYVSFTTKEDSKLPLTLKVRGIKSVVDFRSKDGFIYKLTPDTDENSKAPVYYPVLDAISLKAIWVEGNANFV

      *
SP  GGHA-FVIDD-------GAGRNFYHVDWGWGGVSDGFFRLDALNPSALGTGGGAGGFNGYESAVVGIKP
     ::  .           . ::  .    :::. :..: . ::   :             : ::
IBV VGHPNYYSKSLHIPTFWENAENFVKMGDKIGGVTMGLWRAEHLNKPNLERIFNIAKKAIVGSSVVTTQC
```

Fig.3. Alignment of the putative second cysteine protease domain of IBV with the protease of Streptococcus pneumoniae. The IBV sequence is residues 1385 to 1677 of the F1 polypeptide (4). The S. pneumoniae protease sequence was from (47). Colons: identical residues; dots: similar residues; asterisks: putative catalytic residues. The alignment generated by program OPTAL (see Methods) was slightly corrected to improve local similarity around the catalytic His residue of the bacterial protease and the corresponding residue of IBV.

```
                        2     3             4 5
     factor VII    spCqngggC---kDql-qaYiCfClp
     factor IX     npCLnggaC---kDdi-naYeCwCpf
II   factor X      spCqnQgkC---kDgl-geYtCtCle
     prC           slCcghgtC---iDgi-gaFaCDCrS
     prZ           qpCLnNgaC---qDat-IGYACtCap

     uPA           --CLnggtCvSnkyfa-nihwCNCpk
     tPA           prCfnggtCqqqlyfa-dfv-CQCpe
I    vaccinia 19K GYCLhgd-CiharDid-gmY-CrCch
     TGF           qFC-fhgtC-rflvqe-dkpACvChS
     EGF           GYCLnggVC-mhiEld-saYtCNCvi
                     ::    :    :     : : :
     IBV F1        GFCLrNkVC-TVCQcw-IGYGCQCDS
                     :     :  ::     :: : :
III  LDL R exon 7 --CLdNggCshVCNdlkIGYeClCpd
     LDL R exon 8 --CqdpddCsqLCpdlegGYkCQCEe
                     2     3   1       4 5
```

Fig.4. Alignment of a cysteine-rich segment of the F1 polypeptide of IBV with receptor-binding domains. The IBV sequence was from residue 3894 to 3917 (4). For sources of the other sequences see (25). Abbreviations: factors VII-X, respective human coagulation factors; prC, human plasma protein C; prZ, human plasma protein Z; uPA, urokinase-type plasminogen activator; tPA, tissue-type plasminogen activator; vaccinia 19K, growth factor-like protein of vaccinia virus; TGF, transforming growth factor; EGF, epidermal growth factor; LDL R, low density lypoprotein receptor. The grouping of the EGF-like domains and the numbering of Cys residues is according to (25). Disulfide bonds Cys 1-3, Cys 2-4, Cys 5-6 are expected to form but Cys 6 having no counterpart in the IBV sequence is not shown. Other designations as in Fig.1.

discussed above, this one includes a Gly residue which cannot be directly involved in catalysis. Another conserved Gly residue is substituted by Glu in the CPMV protease, the activity of which was determined in unequivocal experiments (cf. 23).

2nd cysteine protease

Upon comparison of the sequences of F1 and F2 with those of cellular proteases, a segment of F1 has been revealed remarkably similar to a fragment of the catalytic center of Streptococcus pneumoniae cysteine protease. Alignment of the respective portion of F1 with this protease (Fig.3) is significant at approx. 5 SD level. The two most prominent regions of similarity (N- and C-terminal) include segments of the bacterial protease around the catalytic Cys and His residues. Corresponding residues could be identified in IBV , emphasizing the possibility that this segment of F1 could be an authentic protease.

Cysteine-rich segments

An interesting feature of F1 and F2 polypeptides is the presence of several segments with anomalously high content of Cys residues. One of these segments resides in the C-terminal

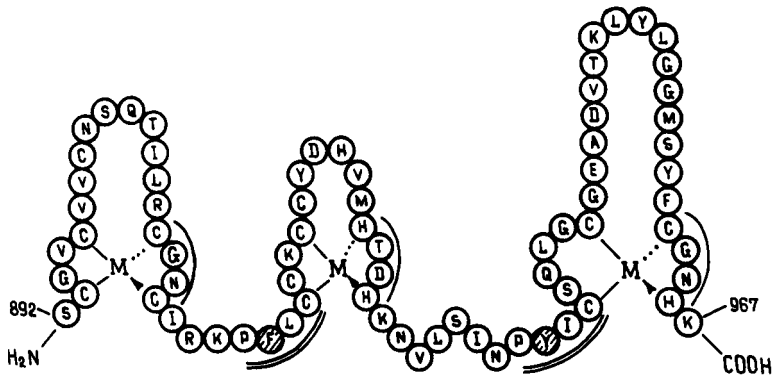

Fig.5. A model of possible organization of the putative metal-binding ("finger") domain of the F2 polypeptide of IBV. Amino acid residue numbering is indicated. Alternative configurations involving other pairs of Cys and His residues are also possible. M, metal (probably Zn^{2+}) cation. Highlighted: similar sequence stretches adjacent to putative metal-binding residues; aromatic residues conserved in TFIIIA-like fingers.

part of F1. It was shown to be significantly similar to the receptor-binding site of murine epidermal growth factor (probability of fortuitous similarity approx. 10^{-10}). Recently EGF-like domains have been divided into three groups differing in cystein residues arrangement and the lengths of spacer segments (25). While bearing the most significant similarity to group 1 domains (EGF, uPA etc.), the IBV domain contains counterparts to only 4 of 6 Cys residues (residues 2-5 in Fig.4) which are highly conserved within this group and are thought to form three disulfide bonds. On the other hand, one of the additional Cys residues present in the IBV sequence could be aligned with Cys 1 of the group 3 domains (LDL R and some other) to which the IBV domain is also considerably similar (Fig.4). Cys 6, however, is absent from this domain. It may be speculated that disulfide bonding might occur between Cys 5 and some more distant Cys residue; several such residues are available in F1 to the N-side of the EGF-like domain. Thus, IBV appears to possess a novel type of EGF-like domain.

Another cysteine-rich segment lies in F2, between the putative RNA polymerase and the RNA helicase. This 30 residue stretch contains 9 Cys and 4 His residues, conforming to the formula of the so called "finger" Zn^{2+}-binding motif ($C\text{-}X_{2\text{-}4}\text{-}C\text{-}X_{2\text{-}15}\text{-}a\text{-}X_{2\text{-}4}\text{-}a$ where a is C or H, and X is any amino acid residue) characteristic of numerous DNA- and RNA-binding proteins (26-28). It is potentially capable of forming three "fingers" supported by Cys and His residues which might tetrahedrally coordinate Zn^{2+} cations (Fig.5), suggesting classification as a class I (i.e. multi-finger) domain (28). No general consensus for finger domains beyond the (putative) metal-binding Cys and His

No	PUTATIVE CLEAVAGE SITES -7 -6 -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 +6 +7 +8 +9	NC	Coordinates (Q)	PROTEIN FUNCTION
3	I G V S R L Q/S G F K K L V S P		F1 2779	MP1
4	I G G V R L Q/S S F V R K A T S		F1 3086	3CL
10	R A P T T L Q/S C G V C V V C N		F2 891	POL
11	D S E T S L Q/G T G L F K I C N		F2 1492	HEL
1	a d V L R f Q/S A r V i a e d v	7	F1 440	
2	a m V I K f Q/G v F K a y A T T	10	F1 2583	
5	t A f k c V Q/G C Y M n a f n T	8	F1 3214	MP2
6	a T N I l I Q/G i G g d R V l P	10	F1 3365	
7	K r a T v L Q/S v t q e f a h i	5	F1 3462	
8	a n V V v L Q/S k G h e t e e v	6	F1 3784	
9	Q p k S S V Q/S v A g a a d f D	8	F1 3928	GFL
12	K S f S a L Q/S i d n i a y n m	6	F2 2012	
13	t c y p q L Q/S A W t C g y n m	6	F2 2350	

Fig.6. Putative cleavage sites in F1 and F2 polyproteins of IBV. The sites are numbered beginning from the N-terminus of F1. The 4 sites which were identified first (3, 4, 10 and 11) and constituted the reference set for identification of the other putative sites are shown in the upper 4 rows. In the other sequences capitals highlight residues having identical or homologous counterparts in at least one of the sequences of the reference set. NC: number of residues having counterparts in the reference sequences. MP1, MP2, putative membrane proteins; POL, putative RNA-dependent RNA polymerase; HEL, putative RNA helicase; 3CL, putative 3C-like protease; GFL, growth factor-like domain. In the 'protein function' column proteins are indicated whose C-terminus may be flanked by the given site.

residues can be derived (26-28), and the putative finger domain of IBV does not appear to bear significant sequence similarity to any particular finger domain of other proteins. Specifically, it does not contain a more strict consensus typical of classical TFIIIA-like fingers (29), although two of the residues thought to be important for proper folding of the latter are present (highlighted in Fig.5). Nevertheless, the conservation of the typical "polarity" of finger domains, with the N-terminal pair of consensus residues represented by Cys_2, and the C-terminal pair by any possible combination of Cys and His, in all the three coronavirus fingers is notable. Also of interest is the similarity between short sequence stretches adjacent to some of the candidate metal-binding residues (Fig.5). Moreover, two of these stretches flanking the Cys residues from the N-side strikingly resembled respective sequences in the finger domains of yeast transcription activator ADR1 (30; data not shown). Thus, whereas the finger-like structures of IBV may not be close structural analogs of TFIIIA-like fingers (cf.29), it seems likely that they constitute an authentic metal-binding and nucleic acid-binding domain.

Putative cleavage sites

We have tentatively identified two protease domains in F1 polypeptide of IBV. Of these, the cleavage specificities of 3C-like proteases have been studied in considerable detail (for reviews see Refs. 31,32). They primarily cleave at dipeptides Q,E/G,S,A. Cleavage occurs selectively and, unfortunately, the requirements for a site to be utilized are not fully understood, probably differing considerably in different viruses. Nevertheless, in potyviruses a clear consensus (though unique for each virus) for the sequences flanking the cleavage sites has been derived (20,33). This encouraged comparison of the sequence stretches centering at Q/S,G dipeptides in the polyproteins of IBV. At the first step, we compared those sites which could flank the putative protease, polymerase and helicase domains. We observed that the distances between highly conserved sequence stretches and protein termini vary to a rather limited extent in most enzymes of each class (Figs.1,2 and data not shown). Thus, three Q/S and one Q/G site were identified in the respective regions of the IBV polyproteins, i. e. sites 3, 4, 10 and 11 in Fig.6. Sites 3 and 4 flank the putative protease, and sites 10 and 11 the putative helicase, site 10 being also the probable C-terminus of the polymerase; the site flanking the polymerase from the N-side was less easily determined (see below). Sequences around these 4 sites bear considerable similarity to each other. Especially pronounced is the similarity between consecutive sites delineating each domain. It could be calculated that the probability of the similarity between sites 3 and 4 being fortuitous was about 10^{-6}, and for sites 10 and 11 about 10^{-5}. It could be shown that the similarity within these two pairs was most prominent among all sequence stretches surrounding Q/G,S dipeptides in F1 and F2.

Based on these observations, we further compared sequences flanking all the Q/S,G dipeptides contained in the F1 and F2 polyproteins to those surrounding the 4 tentatively identified cleavage sites (Fig. 6). Thus, 9 additional putative cleavage sites bearing some resemblance to the first 4 were identified (Fig. 6). A notable feature of all the 13 detected sites is the presence of a hydrophobic residue (mostly L) in position -1 which is thought to be most important for cleavage by 3C-like proteases (31). Also of interest are peculiarities of sites 3 and 4 flanking the putative 3C-like protease (F in position +3 and a positively charged residue in position -3) shared by site 2. It is tempting to speculate that these may be specific requirements for intramolecular cleavage. Some of the sequences shown in Fig. 6 bear additional similarities to each other (for example, sites 12 and 13), emphasizing the case for their authenticity. Finally, a striking resemblance is observed between some of the putative cleavage sites of IBV (especially sites 1, 2 and 4) and the consensus (VRFQ/S,G) derived for the polyprotein cleavage sites of one of the potyviruses, TVMV (20). However, contrary to what is observed in potyviruses (34,35), the C-flanking sequences of the putative coronavirus cleavage sites are also somewhat similar to each other (Fig. 6) and, by implication, might be important for processing.

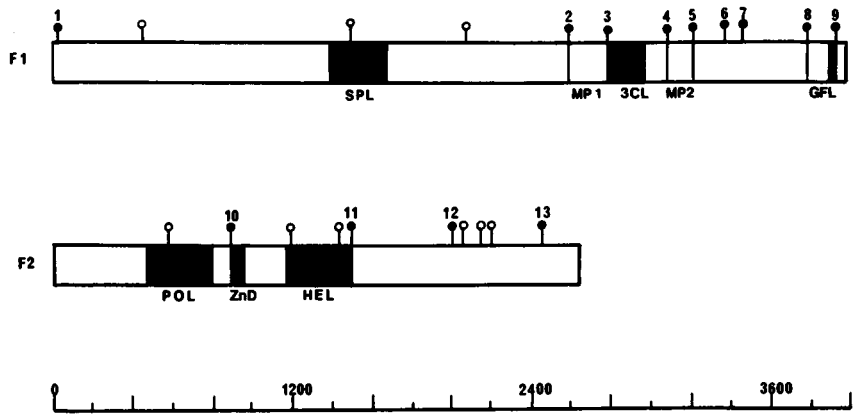

Fig.7. A scheme depicting possible organization of functional domains in the non-structural polyprotein of IBV.
F1 and F2 polypeptides are shown to scale. Filled circles: putative cleavage sites numbered as in Fig.6; empty circles: Q/G,S dipeptides whose flanking sequences bear no significant resemblance to those of the reference set (Fig.6) and which are thought to be not utilized. SPL, putative protease similar to the S. pneumoniae protease; ZnD, putative Zn^{2+}-binding domain. Regions of significant sequence similarity to respective viral or cellular proteins are shown in black. Other designations as in Fig.6.

Implications for coronavirus polyprotein organization and expression

Fig. 7 schematically summarizes what could be derived from amino acid sequence of the coronavirus non-structural polyprotein(s) by computer-assisted comparisons. The approx. 6600 amino residue polyprotein (provided F1 and F2 are actually joined by translation frame-shift) may be provisionally separated into two vast regions, the N-terminal one of about 2600 residues, and the approx. 4000 residue C-terminal one. The C-terminal part encompasses the putative cleavage sites for the 3C-like protease which are predicted to be cleaved (see above) and it is tempting to suggest that expression of this region of the polyprotein might be completely controlled by the 3C-like protease. In principle, cleavage at sites shown in Fig. 6 might be sufficient for the generation of all the proteins of IBV essential for genome replication and expression. Organization of the complex of these proteins (domains) is principally similar to that observed in other positive strand RNA viruses but certain interesting unique features are also present. Specifically, the putative 3C-like protease may be flanked by two relatively small proteins having long N-terminal stretches of hydrophobic amino acid residues, presumably membrane-spanning domains, which might influence the intracellular topology of the protease. Localization of the 3C-like protease relative to the polymerase is also typical of

other viruses having enzymes of this family, though a large domain of unknown function is inserted between the two conserved domains. This domain contains the EGF-like sequence unique for IBV. The best guess concerning the function of this domain is that it might be involved in some special kind of protein-protein interaction. In fact, it is not clear what may be the N-terminus of the polymerase, the size of this enzyme varying to a large extent among positive strand RNA viruses (cf. Fig.1 and refs. 16,22). Still, site 9 separating the EGF-like sequence from the putative polymerase and bearing some resemblance to site 10 (Fig. 6) is the most plausible candidate.

What is unusual in the organization of the putative complex of replication proteins of IBV, is the mutual orientation of the polymerase and helicase domains which is conversed as compared to that observed in other positive strand RNA viruses (6). An interesting possibility is that this array could arise as a result of a recombinational event, high frequency of recombination being a salient feature of coronavirus reproduction (1-3). Another unique feature of coronaviruses is the presence of the "finger" domain which, provided the cleavage sites are determined correctly, may be the N-terminal part of the helicase (Fig. 7). It has recently been demonstrated that small finger proteins of retroviruses possess RNA annealing activity and mediate positioning of the replicative primer (a specific tRNA) on the viral genome (36). A similar role in the primer-dependent transcription and recombination of the coronavirus genome (cf. 3) is plausible for the finger domain of IBV. These functions might be performed in conjunction with the helicase domain. A putative single-finger domain has been identified also in the N-terminal portion of the polyprotein of another coronavirus, MHV (37). No obvious similarity between this region and any IBV sequence could be revealed. Evaluation of the significance of this observation awaits complete sequencing of the MHV genome.

In the N-terminal portion of the polyprotein, the only domain for which a function could be proposed is the putative 2nd thiol protease. Possibly, it may control processing of this region whose pathway as well as functions of the products remain obscure. A possible exception is a 440 residue domain at the very N-terminus of F1 flanked by one of the putative cleavage sites of the 3C-like protease (Fig. 7). Some sequence similarity has been detected between a portion of this region and the replication initiation protein of the R6K plasmid (4, 38). In the absence of any data about functional sites of the latter, it is, however, difficult to assess the significance of this observation.

Generally, the identification, in IBV, of putative homologs of three conserved domains of positive strand RNA viruses suggests an evolutionary relationship between coronaviruses and other groups of this class. On the other hand, the unique biochemistry of coronaviruses seems to be reflected in the unusual arrangement of these domains, and in the presence of additional specific ones.

CONCLUDING REMARKS

The findings reported here may prove important in several dimensions. First, and most obvious, one may hope that the predictions made might catalyze studies directed on experimental dissection of coronavirus non-structural polyprotein(s). Second, the putative polymerase, helicase and 3C-like protease of IBV, while related to the similar enzymes of other positive strand RNA viruses at a statistically significant level, loosened the respective consensus patterns, thus providing a new groundwork for probing newly sequenced viral genomes. Finally, the general approach utilized also might be helpful in analysis of other genomes.

ACKNOWLEDGEMENTS

The authors are grateful to Professor V. I. Agol for encouragement and critical reading of the manuscript, to Dr. K. M. Chumakov for help with programming, and to Dr. S. Yu. Morozov for helpful suggestions.

*To whom correspondence should be addressed

REFERENCES

1. Siddel,S.G., Anderson,R., Cavanagh,D., Fujiwara,K., Klenk, H.D., MacNaughton,M.R., Pensaert,M., Stohlman,S.A., Sturman, L. and Van der Zeist,B.A.M. (1983) Intervirology 20, 181-189.
2. Sturman,L.S. and Holmes,K. (1983) Adv.Virus Res. 28, 35-112.
3. Lai,M.M.C. (1986) BioEssays 5, 257-260.
4. Bournell,M.E.G., Brown,T.D.K., Foulds,I.J., Green,P.F., Tomley,F.M. and Binns,M.M. (1987) J.gen.Virol. 68, 57-77.
5. Brierley,I., Boursnell,M.E.G., Binns,M.M., Bilimoria,B., Blok,V.C., Brown,T.D.K. and Inglis,S.C. (1987) EMBO J. 6, 3779-3785.
6. Gorbalenya,A.E., Blinov,V.M., Donchenko,A.P. and Koonin,E.V. (1988a) J.mol.Evol. 28, in press.
7. Gorbalenya,A.E., Koonin,E.V., Donchenko,A.P. and Blinov,V.M. (1988b) Nature 333, 22.
8. Gorbalenya,A.E., Koonin,E.V., Donchenko,A.P. and Blinov,V.M. (1988c) FEBS Lett. 235, 16-24.
9. Hodgman,T.C. (1988) Nature 333, 22-23.
10. Staden, R. (1982) Nucleic Acids Res. 10, 2951-2961.
11. Argos, P. (1987) J.mol.Biol. 193, 385-396.
12. Pozdnyakov,V.I. and Pankov,Yu.A. (1981) Int.J.Peptide Prot.Res. 17, 284-291.
13. Sankoff,D. (1972) Proc.Nat.Acad.Sci.USA 69, 4-6.
14. Mclachlan,A.D. (1971) J.Mol.Biol. 61, 409-424.
15. Kamer, G. and Argos,P. (1984) Nucl.Acids.Res. 12, 7269-7282.
16. Koonin,E.V., Gorbalenya,A.E., Chumakov,K.M., Donchenko,A.P.

and Blinov,V.N. (1987) Molek. Genetika No.7, 27-39 (in Russian).
17. Argos,P., Kamer,G., Nicklin,M.J.H. and Wimmer,E. (1984) Nucl.Acids Res. 12, 7251-7267.
18. Gorbalenya,A.E., Blinov,V.N. and Donchenko,A.P. (1986) FEBS Lett. 194, 253-257.
19. Gorbalenya,A.E., Donchenko,A.P., Blinov,V.N. and Koonin, E.V. (1989) FEBS Lett., in press.
20. Domier,L., Shaw,J.G. and Rhoads,R.E. (1987) Virology 158, 20-27.
21. Greif,C., Hammer,O. and Fritsch,C. (1988) J.Gen.Virol. 69, 1517-1529.
22. Zimmern,D. (1988) In Holland,J.J., Domingo,E. and Ahlquist, P. (ed.) RNA Genetics, Boca Raton, FL, in press.
23. Krausslich,H.-G. and Wimmer,E. (1988) Ann.Rev.Biochem. 57, 701-754.
24. Brayton,P.R., Stohlman,S.A. and Lai,M.M.C. (1984) Virology 133, 197-201.
25. Appella,E., Weber,I.T. and Blasi,F. (1988) FEBS Lett. 241, 1-4.
26. Berg,J.M. (1986) Science 242, 485-487.
27. Evans,R.M. and Hollenberg,S.M. (1988) Cell 52, 1-3.
28. Payre,F. and Vincent,A. (1988) FEBS Lett. 244, 245-250.
29. Frankel,A.D. and Pabo,C.A. (1988) Cell 53, 675.
30. Hartshorne,T.A., Blumberg,H. and Young,E.T. (1986) Nature 320, 283-287.
31. Palmenberg,A.C. (1987) J.Cell.Biochem. 33, 191-198.
32. Wellink,J. and Van Kammen,A. (1988) Arch.Virol. 98, 1-26.
33. Allison,R.F., Johnston,R.E. and Dougherty,W.G. (1986) Virology 154, 9-20.
34. Dougherty,W.G., Carrington,J.C., Cary,S.M., Parks,T.D. (1988) EMBO J. 7, 1281-1287.
35. Carrington,J.C. and Dougherty,W.G. (1988) Proc.Nat.Acad. Sci.USA 85, 3391-3395.
36. Prats,A.C., Sarih,L., Gabus,C., Litvak,S., Keith,G. and Darlix,J.L. (1988) EMBO J. 7, 1777-1783.
37. Soe,L.H., Shieh,C.-K., Baker,S.C., Chang,M.-F. and Lai,M.M.C. (1987) J.Virol. 61, 3968-3976.
38. Boursnell,M.E.G., Brown,T.D.K., Foulds,I.J., Green,P.F., Tomley,F.M., Binns,M.M. (1987) In Lai,M.M.C. and Stohlman,S.A.(ed.). Coronaviruses. Plenum Press, New York, pp.15-29.
39. Lain,S., Riechmann,J.L., Mendez,E. and Garcia,J.A. (1988) Virus Res. 10, 325-342.
40. Gustafson,G., Hunter,B., Hanau,R., Armour,S. and Jackson, A.O. (1987) Virology 158, 394-406.
41. Racaniello,V.C. and Baltimore,D. (1981) Proc.Nat.Acad.Sci.USA 78, 4887-4891.
42. Skern,T., Sommergruber,W., Blaas,D., Gruendler,P., Fraundorfer,F., Pieler,C., Fogy,I. and Kuechler,E. (1985) Nucleic Acids Res. 13, 7859-7875.
43. Palmenberg,A.C., Kirby,E.M., Janda,M.R., Drake,N.I.,

Potratz,K.F. and Collett,M.C. (1984) Nucleic Acids Res. 12, 2969-2985.
44. Carrol,A.R., Rowlands,D.J. and Clarke,B.E. (1984) Nucleic Acids Res. 12, 2461-2472.
45. Najerian,R., Caput,D., Gee,W., Potter,S.J., Renard,A., Merryweather, J., Van Nest,G. and Dino,D. (1985) Proc.Nat.Acad.Sci.USA 82, 2627-2631.
46. Lomonosoff,G.P. and Shanks,M. (1983) EMBO J. 2, 2253-2258.
47. Tai,J.Y. and Liu,T.-Y. (1976) J.Biol.Chem. 251, 1955-1959.