Volume 243, number 2, 103–114 FEB 06691 January 1989

Discussion Letter

Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases

A distinct protein superfamily with a common structural fold

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

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

Received 11 October 1988

Evidence is presented, based on sequence comparison and secondary structure prediction, of structural and evolutionary relationship between chymotrypsin-like serine proteases, cysteine proteases of positive strand RNA viruses (3C proteases of picornaviruses and related enzymes of como-, nepo- and potyviruses) and putative serine protease of a sobemovirus. These observations lead to re-identification of principal catalytic residues of viral proteases. Instead of the pair of Cys and His, both located in the C-terminal part of 3C proteases, a triad of conserved His, Asp(Glu) and Cys(Ser) has been identified, the first two residues resident in the N-terminal, and Cys in the C-terminal β-barrel domain. These residues are suggested to form a charge-transfer system similar to that formed by the catalytic triad of chymotrypsin-like proteases. Based on the structural analogy with chymotrypsin-like proteases, the His residue previously implicated in catalysis, together with two partially conserved Gly residues, is predicted to constitute part of the substrate-binding pocket of 3C proteases. A partially conserved ThrLys/Arg dipeptide located in the loop preceding the catalytic Cys is suggested to confer the primary cleavage specificity of 3C toward Glx/Gly(Ser) sites. These observations provide the first example of relatedness between proteases belonging, by definition, to different classes.

Serine protease; Cysteine protease; Amino acid sequence comparison; Protein fold prediction; Positive strand RNA virus

Correspondence address: A.E. Gorbalenya, Institute of Poliomyelitis and Viral Encephalitides of the USSR Academy of Medical Sciences, 142782 Moscow Region, USSR

Abbreviations: PV1, poliovirus type 1; HRV1a, human rhinovirus type 1a; HRV2, human rhinovirus type 2; HRV14, human rhinovirus type 14; CVB3, coxsackie virus type B3; ECHO, echovirus type 9; BEV, bovine enterovirus; TMEV, Theiler murine encephalomyelitis virus; EMCV, encephalomyocarditis virus; FMDV, foot-and-mouth disease virus type A10; HAV, hepatitis A virus (picornaviruses); CPMV, cowpea mosaic virus (comovirus); TBRV, tomato black ring virus (nepovirus); TVMV, tobacco vein mottling virus; TEV, tobacco etch virus (potyviruses); SBMV, southern bean mosaic virus (sobemovirus); SGPA, *Streptomyces griseus* protease A; SGPB, *Streptomyces griseus* protease B; CHT, chymotryptin; TRP, trypsin; ELA, elastase (chymotrypsin-like proteases)

Reported in part at International Symposium on Fundamental and Applied Aspects of the Molecular Biology of Picornaviruses, Moscow, May 1988

1. INTRODUCTION

Cysteine and serine proteases are usually regarded as unrelated enzyme classes [1,2]. Specifically, 3C proteases ($3C^{Pro}$) involved in polyprotein processing of picornaviruses and similar enzymes of three groups of plant viruses (como-, nepo- and potyviruses), for some of which principal catalytic residues have been identified as Cys by inhibitor studies [3,4], were traditionally compared to cysteine proteases such as cathepsins and papain [5–8]. Of the few residues conserved in all aligned sequences of $3C^{Pro}$, only two, Cys and His, both near the C-terminus, were considered as possible catalytic ones, based on the analogy with cellular

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50

cysteine proteases; notably, however, in the latter the catalytic Cys is located near the N-terminus (cf. [5]). The functional importance of these two residues has been subsequently confirmed by site-directed mutagenesis [9], though direct test of the hypothesis has not been reported. Due to the different location of the (putative) catalytic Cys residues and to the lack of overall sequence similarity, it was suggested that 3C-like proteases were not evolutionarily related to other cysteine proteases [5], their formal analogy being explained by convergence. The considerable similarity between the regions of 3C proteases around the putative catalytic Cys to those surrounding the catalytic Ser of chymotrypsin-like proteases noticed by us [10,11] was also attributed to convergence [8]. Hence, the general consensus that 3C-like proteases constitute an entirely independent enzyme family. However, two very recent observations encouraged re-evaluation of this concept. First, it has been shown that the His residue implicated in catalysis is not conserved in the putative protease of a nepovirus [12]. Second, we have tentatively identified, in a sobemovirus, a serine protease significantly similar to $3C^{Pro}$ [13].

Using an algorithm for stepwise multiple sequence alignment, we here present a new version of the complete sequence alignment of 3C-like proteases. Previously not detected conserved His and Asp (Glu) residues have been revealed, which, together with the Cys (Ser) residue identified earlier, might constitute a catalytic triad similar to that of chymotrypsin-like serine proteases. Moreover, a significant overall similarity at the primary and secondary structure levels between 3C-like and chymotrypsin-like proteases was revealed, allowing tentative identification of other functionally important sites of the former. We hypothesize that 3C-like and chymotrypsin-like proteases provide a previously unprecedented case of structural and evolutionary relatedness between proteases of different classes.

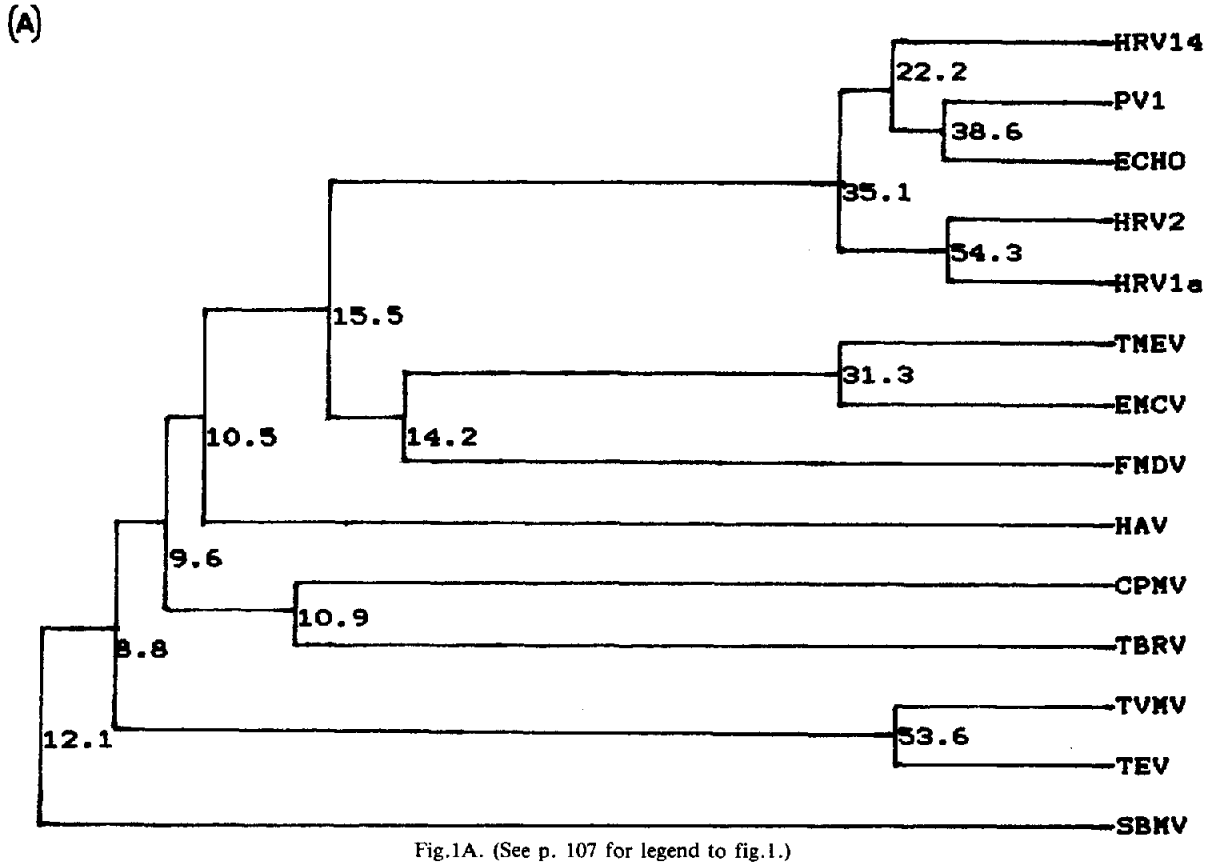
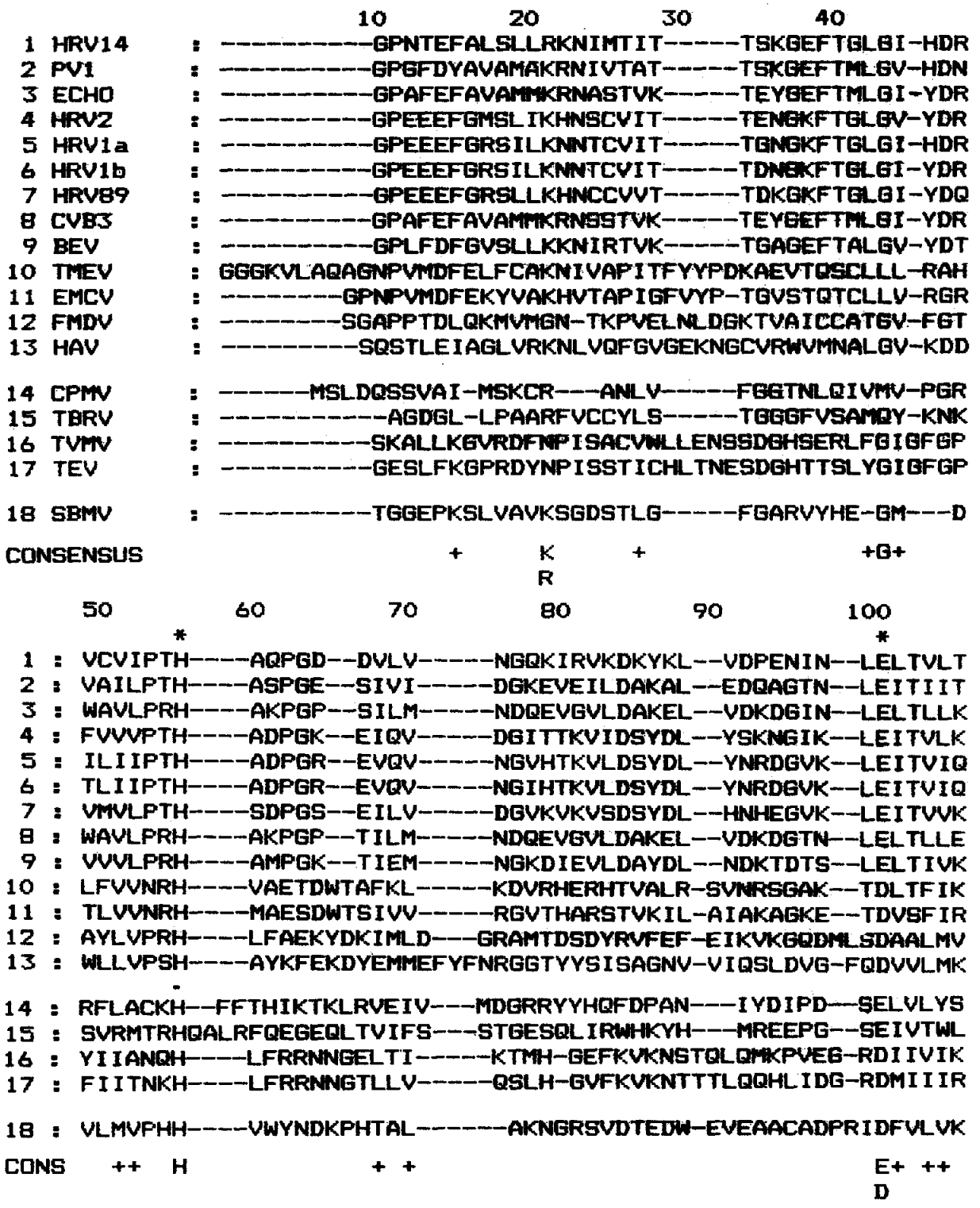
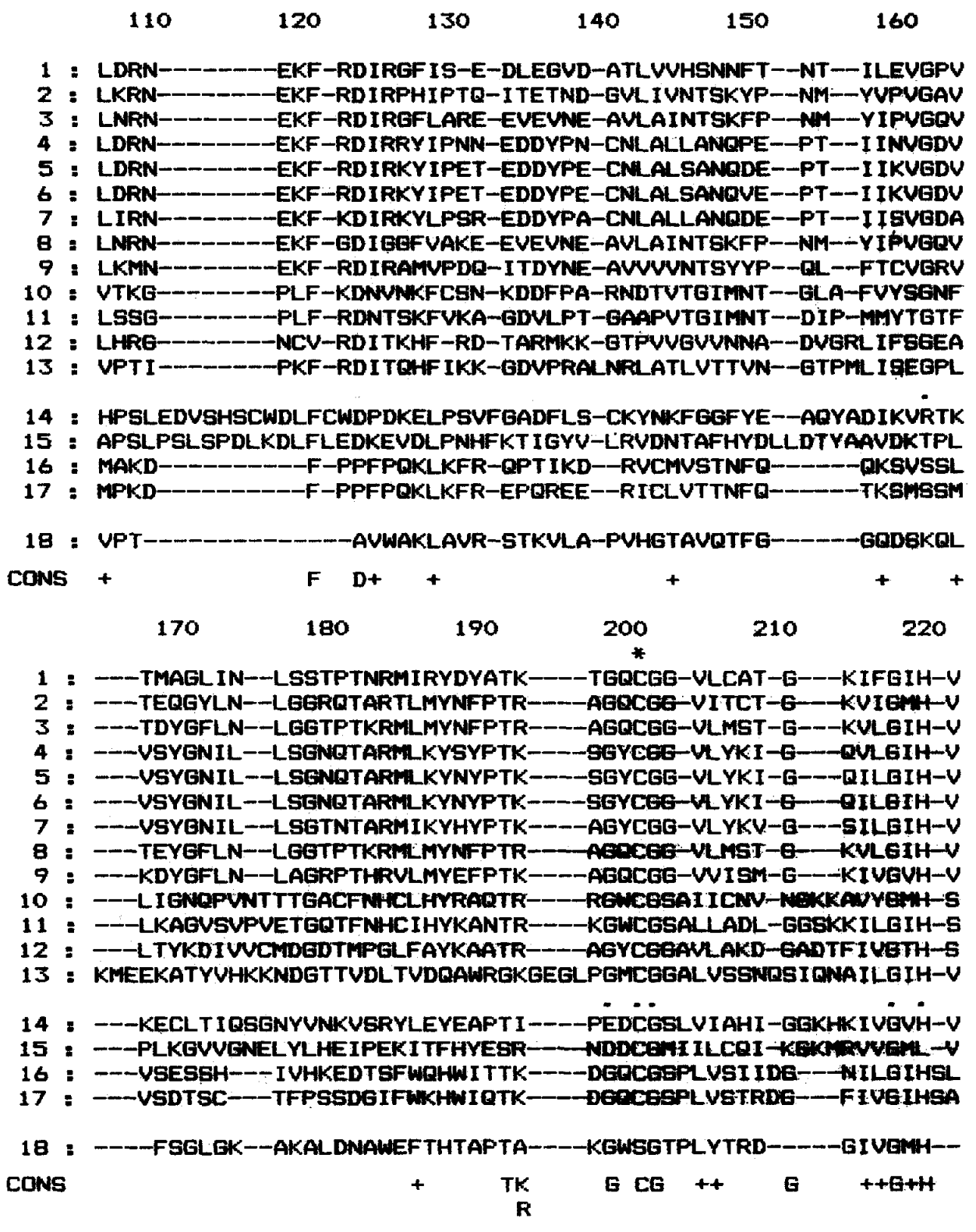


Fig.1A. (See p. 107 for legend to fig.1.)

(B)


10 20 30 40
1 HRV14 : ----------GPNTEFALSLLRKNIMTIT-----TSKGEFTGLGI-HDR
2 PV1 : ----------GPGFDYAVAMAKRNIVTAT-----TSKGEFTMLGV-HDN
3 ECHO : ----------GPAFEFAVAMMKRNASTVK-----TEYGEFTMLGI-YDR
4 HRV2 : ----------GPEEEEFGMSLIKHNSCVIT-----TENGKFTGLGV-YDR
5 HRV1a : ----------GPEEEEFGRSILKNNTCVIT-----TGNGKFTGLGI-HDR
6 HRV1b : ----------GPEEEEFGRSILKNNTCVIT-----TDNGKFTGLGI-YDR
7 HRV89 : ----------GPEEEEFGRSLLKHNCCVVT-----TDKGKFTGLGI-YDQ
8 CVB3 : ----------GPAFEFAVAMMKRNGSTVK-----TEYGEFTMLGI-YDR
9 BEV : ----------GPLFDFGVSLLKKNIRTVK-----TGAGEFTALGV-YDT
10 TMEV : GGGKVLAQAGNPVMDFELFCAKNIVAPITFYYPDKAEVTQSCLLL-RAH
11 EMCV : ---------GPNPVMDFEKYVAKHVTAPIGFVYP-TGVSTQTCLLV-RGR
12 FMDV : ---------SGAPPTDLQKMVMGN-TKPVELNLDGKTVAICCATGV-FGT
13 HAV : ----------SQSTLEIAGLVRKNLVQFGVGEKNGCVRWVMNALGV-KDD
14 CPMV : ------MSLDQSSVAI-MSKCR---ANLV-----FGGTNLQIVMV-PGR
15 TBRV : ----------AGDGL-LPAARFVCCYLS-----TGGGFVSAMQY-KNK
16 TVMV : ----------SKALLKGVRDFNPISACVWLLENSSDGHSERLFGIGFGP
17 TEV : ----------GESLFKGPRDYNPISSTICHLTNESDGHTTSLYGIGFGP
18 SBMV : ----------TGGEPKSLVAVKSGDSTLG-----FGARVYHE-GM---D
CONSENSUS + K + +G+
R
50 60 70 80 90 100
*
1 : VCVIPTH----AQPGD--DVLV-----NGQKIRVKDKYKL--VDPENIN--LELTVLT
2 : VAILPTH----ASPGE--SIVI-----DGKEVEILDAKAL--EDQAGTN--LEITIIT
3 : WAVLPRH----AKPGP--SILM-----NDQEVGVLDAKEL--VDKDGIN--LELTLLK
4 : FVVVPTH----ADPGK--EIQV-----DGITTKVIDSYDL--YSKNGIK--LEITVLK
5 : ILIIPTH----ADPGR--EVQV-----NGVHTKVLDSYDL--YNRDGVK--LEITVIQ
6 : TLIIPTH----ADPGR--EVQV-----NGIHTKVLDSYDL--YNRDGVK--LEITVIQ
7 : VMVLPTH----SDPGS--EILV-----DGVKVKVSDSYDL--HNHEGVK--LEITVVK
8 : WAVLPRH----AKPGP--TILM-----NDQEVGVLDAKEL--VDKDGTN--LELTLLE
9 : VVVLPRH----AMPGK--TIEM-----NGKDIEVLDAYDL--NDKTDTS--LELTIVK
10 : LFVVNRH----VAETDWTAFKL-----KDVRHERHTVALR-SVNRSGAK--TDLTFIK
11 : TLVVNRH----MAESDWTSIVV-----RGVTHARSTVKIL-AIAKAGKE--TDVSFIR
12 : AYLVPRH----LFAEKYDKIMLD---GRAMTDSDYRVFEF-EIKVKGQDMLSDAALMV
13 : WLLVPSH----AYKFEKDYEMMEFYFNRGGTYYSISAGNV-VIQSLDVG-FQDVVLMK
14 : RFLACKH--FFTHIKTKLRVEIV---MDGRRYYHQFDPAN---IYDIPD--SELVLYS
15 : SVRMTRHQALRFQEGEQLTVIFS---STGESQLIRWHKYH---MREEPG--SEIVTWL
16 : YIIANQH----LFRRNNGELTI-----KTMH-GEFKVKNSTQLQMKPVEG-RDIIVIK
17 : FIITNKH----LFRRNNGTLLV-----QSLH-GVFKVKNTTTLQQHLIDG-RDMIIIR
18 : VLMVPHH----VWYNDKPHTAL------AKNGRSVDTEDW-EVEAACADPRIDFVLVK
CONS ++ H + + E+ ++
D


110 120 130 140 150 160
1 : LDRN--------EKF-RDIRGFIS-E-DLEGVD-ATLVVHSNNFT--NT--ILEVGPV
2 : LKRN--------EKF-RDIRPHIPTQ-ITETND-GVLIVNTSKYP--NM--YVPVGAV
3 : LNRN--------EKF-RDIRGFLARE-EVEVNE-AVLAINTSKFP--NM--YIPVGQV
4 : LDRN--------EKF-RDIRRYIPNN-EDDYPN-CNLALLANQPE--PT--IINVGDV
5 : LDRN--------EKF-RDIRKYIPET-EDDYPE-CNLALSANQDE--PT--IIKVGDV
6 : LDRN--------EKF-RDIRKYIPET-EDDYPE-CNLALSANQVE--PT--IIKVGDV
7 : LIRN--------EKF-KDIRKYLPSR-EDDYPA-CNLALLANQDE--PT--IISVGDA
8 : LNRN--------EKF-GDIGGFVAKE-EVEVNE-AVLAINTSKFP--NM--YIPVGQV
9 : LKMN--------EKF-RDIRAMVPDQ-ITDYNE-AVVVVNTSYYP--QL--FTCVGRV
10 : VTKG--------PLF-KDNVNKFCSN-KDDFPA-RNDTVTGIMNT--GLA-FVYSGNF
11 : LSSG--------PLF-RDNTSKFVKA-GDVLPT-GAAPVTGIMNT--DIP-MMYTGTF
12 : LHRG--------NCV-RDITKHF-RD-TARMKK-GTPVVGVVNNA--DVGRLIFSGEA
13 : VPTI--------PKF-RDITQHFIKK-GDVPRALNRLATLVTTVN--GTPMLISEGPL
14 : HPSLEDVSHSCWDLFCWDPDKELPSVFGADFLS-CKYNKFGGFYE--AQYADIKVRTK
15 : APSLPSLSPDLKDLFLEDKEVDLPNHFKTIGYV-LRVDNTAFHYDLLDTYAAVDKTPL
16 : MAKD----------F-PPFPQKLKFR-QPTIKD--RVCMVSTNFQ------QKSVSSL
17 : MPKD----------F-PPFPQKLKFR-EPQREE--RICLVTTNFQ------TKSMSSM
18 : VPT--------------AVWAKLAVR-STKVLA-PVHGTAVQTFG------GQDSKQL
CONS + F D+ + + + +
170 180 190 200 210 220
1 : ---TMAGLIN--LSSTPTNRMIRYDYATK-----TGQCGG-VLCAT-G---KIFGIH-V
2 : ---TEQGYLN--LGGRQTARTLMYNFPTR-----AGQCGG-VITCT-G---KVIGMH-V
3 : ---TDYGFLN--LGGTPTKRMLMYNFPTR-----AGQCGG-VLMST-G---KVLGIH-V
4 : ---VSYGNIL--LSGNQTARMLKYSYPTK-----SGYCGG-VLYKI-G---QVLGIH-V
5 : ---VSYGNIL--LSGNQTARMLKYNYPTK-----SGYCGG-VLYKI-G---QILGIH-V
6 : ---VSYGNIL--LSGNQTARMLKYNYPTK-----SGYCGG-VLYKI-G---QILGIH-V
7 : ---VSYGNIL--LSGTNTARMIKYHYPTK-----AGYCGG-VLYKV-G---SILGIH-V
8 : ---TEYGFLN--LGGTPTKRMLMYNFPTR-----AGQCGG-VLMST-G---KVLGIH-V
9 : ---KDYGFLN--LAGRPTHRVLMYEFPTK-----AGQCGG-VVISM-G---KIVGVH-V
10 : ---LIGNQPVNTTTGACFNHCLHYRAQTR-----RGWCGSAIICNV-NGKKAVYGMH-S
11 : ---LKAGVSVPVETGQTFNHCIHYKANTR-----KGWCGSALLADL-GGSKKILGIH-S
12 : ---LTYKDIVVCMDGDTMPGLFAYKAATR-----AGYCGGAVLAKD-GADTFIVGTH-S
13 : KMEEKATYVHKKNDGTTVDLTVDQAWRGKGEGLPGMCGGALVSSNQSIQNAILGIH-V
14 : ---KECLTIQSGNYVNKVSRYLEYEAPTI-----PEDCGSLVIAHI-GGKHKIVGVH-V
15 : ---PLKGVVGNELYLHEIPEKITFHYESR-----NDDCGMIILCQI-KGKMRVVGML-V
16 : ---VSESSH---IVHKEDTSFWQHWITTK-----DGQCGSPLVSIIDG---NILGIHSL
17 : ---VSDTSC---TFPSSDGIFWKHWIQTK-----DGQCGSPLVSTRDG---FIVGIHSA
18 : ----FSGLGK--AKALDNAWEFTHTAPTA-----KGWSGTPLYTRD-----GIVGMH--
CONS + TK G CG ++ G ++G+H
R

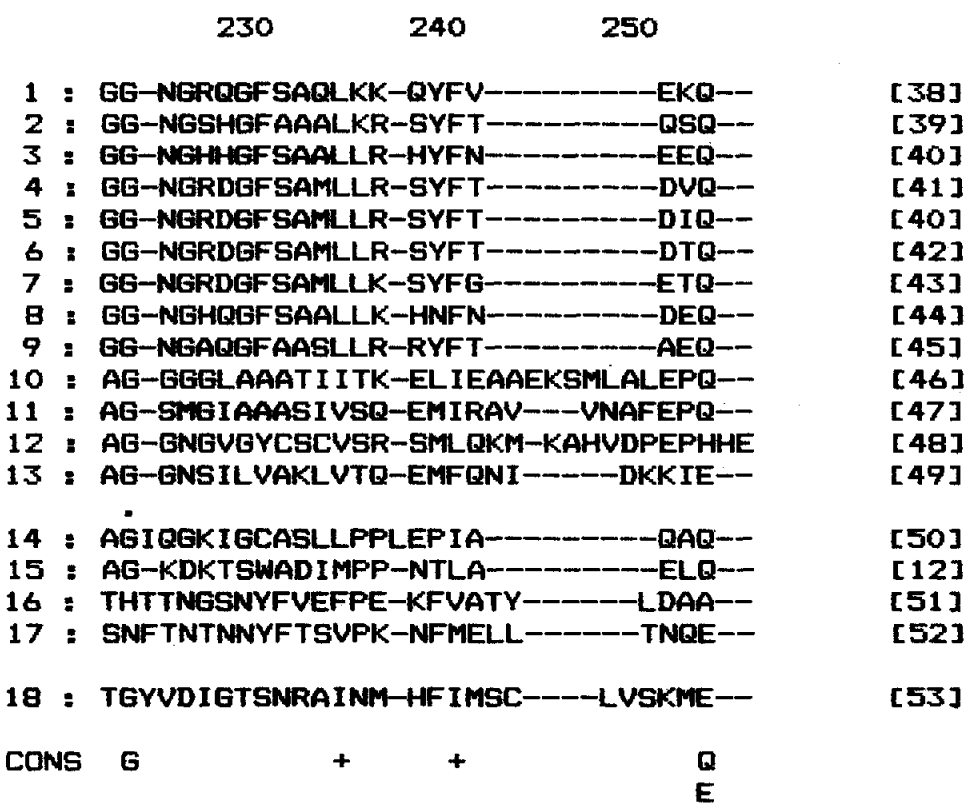


Fig.1. Alignment of amino acid sequences of 3C-like proteases. (A) A dendrogram schematically depicting the course of alignment in the order of decreasing similarity. Branch lengths are in approximate inverse proportion to the degree of sequence similarity observed at each step. AS values in SD units are indicated for each step. (B) The resulting protease alignment (sequences of $3C^{Pro}$ of HRV1a, HRV89, CVB3 and BEV published recently and those not included in A were added by hand, based on their unambiguous alignment with $3C^{Pro}$ of other entero- and rhinoviruses). The aligned sequences are numbered arbitrarily, beginning from the first position of the alignment. Between residues 81 and 120, the alignment of CPMV and TBRV proteases with those of other viruses was uncertain and was corrected by the HELIX program comparing multiple pre-aligned sequences in a diagonal plot and revealing conserved regions (in preparation). Below the aligned sequences the derived consensus (CONS) is shown. A residue (or two homologous residues) was included in the consensus if it occurred in at least 14 out of 18 sequences. Residues belonging to one of the following groups were scored as homologous: D, E, N, Q; S, T; K, R; V, L, I, M; F, Y, W. +, hydrophobic residues (V, L, I, M, F); *, putative catalytic residues. Dots: residues invariant in picornaviral $3C^{Pro}$. Sequences were from the references indicated at the end of the alignment. The proteases of potyviruses (NI_a proteins) have terminal extensions [51,52]. Cleavage sites flanking the putative protease of SBMV are discussed in [13].

2. SEQUENCE ALIGNMENT OF 3C-LIKE PROTEASES

Amino acid sequences of 3C-like proteases were aligned by the OPTAL program which performs stepwise optimal alignment of multiple amino acid sequences and its statistical assessment by a Monte Carlo procedure [14,15]. Alignments were statistically characterized by alignment scores (AS) as follows: $AS = S^o - S^r/\sigma$ where S^o is the score calculated for an alignment of two sequences, or a group of sequences, by use of the MDM78 amino acid residue comparison matrix, S^r is the mean score for alignments of 25 random permutations of the same sequences, and σ is the standard deviation. The alignment of 3C-like proteases, together with the AS values obtained at each step, is shown in fig.1. The alignment of all the sequences was highly significant (fig.1A), confirming that 3C-like proteases most probably constitute a monophyletic protein family [16].

The general premise underlying any functional

implications of sequence comparisons is that common functions (primarily catalytic) should be performed by conserved amino acid residues [16]. Upon alignment of picornaviral $3C^{Pro}$, 9 invariant residues (5 Gly, two His, one Cys and one Asp) were revealed (fig.1B). Of special interest was the conservation of His^{56} (numbering of the alignment shown in fig.1B) in the relatively variable N-terminal half of $3C^{Pro}$. Addition of the sequences of more distantly related plant viral (putative) proteases reduced this number to only 3, namely His^{56}, Gly^{203} and Gly^{219}, with the putative catalytic Cys^{202} replaced by Ser in the SBMV protein [13]. On the other hand, His^{221} and Asp^{125}, which are conserved in picornaviral proteases and were previously tentatively implicated in catalysis [5–8], are substituted by non-homologous residues in nepovirus and in poty- and sobemovirus proteins, respectively.

3. COMPARISON OF $3C^{Pro}$ AND CHYMOTRYPSIN-LIKE PROTEASES

The mutual orientation of the conserved His^{56} and $Cys(Ser)^{202}$ residues in 3C-like proteases resembles that of the respective catalytic residues in chymotrypsin-like serine proteases (His^{57} and Ser^{195}, according to the chymotrypsin numbering system [17]) but not in cellular cysteine proteases. The similarity of sequence stretches surrounding the (putative) catalytic Cys residues of $3C^{Pro}$ to those around the catalytic Ser of chymotrypsin-like proteases has been noticed and discussed previously [11]. These observations prompted a further, more detailed comparison between the two enzyme families.

Secondary and tertiary structures of chymotrypsin-like proteases are better conserved than amino acid sequences [17]. Thus it seemed important to compare them with 3C-like proteases at these levels of organization. Since X-ray data for 3C-like enzymes are not available, only secondary structure predictions could be used to this end. Secondary structures of 3C-like proteases were predicted by the ALBEAL program based on the algorithm of Finkelstein and Ptitsyn [18,19]. To improve prediction quality, α-helix and β-strand potentials were averaged according to the amino acid sequence alignment shown in fig.1. This type of analysis was restricted to picornaviral $3C^{Pro}$ for which the sequence alignment was most reliable. Comparison of the resulting secondary structure profile with those determined for 5 chymotrypsin-like proteases by the same approach and by X-ray crystallography revealed reasonable similarity (fig.2). Obviously, $3C^{Pro}$ belong to the class of proteins of which chymotrypsin-like enzymes are typical representatives [20]. The latter are known to comprise 12 β-strands (A to L in fig.2C). 11 of 12 strands and a C-terminal α-helix could be identified in the predicted profile (fig.2A,C) though the strength of prediction varied considerably. Curiously, the secondary structure of $3C^{Pro}$ appeared to be predicted somewhat better than that of chymotrypsin-like enzymes, with stronger β-strand prediction and counterparts available for all 12 strands revealed in the latter by X-ray analysis (fig.2B,C). It must be emphasized that spacing of the (predicted) β-strands was very similar in the proteases of the two families. Similarity in the positions of deletions and insertions which are usually associated with loops in protein structure [21] is also notable (fig.2C). A specific element of $3C^{Pro}$ which is absent in chymotrypsin-like proteases is the strongly predicted N-terminal α-helix.

A salient feature of chymotrypsin-like proteases is that they consist of two topologically similar domains comprising 6 strands each [22]. This symmetrical organization is clearly seen in the secondary structure profile of $3C^{Pro}$, better in fact than in the chymotrypsin-like proteases themselves (cf. fig.2A and B). Moreover, each domain is composed of two half-domains [22], and these could also be discerned in the $3C^{Pro}$ profile.

These observations encouraged aligning amino acid sequences of $3C^{Pro}$ and chymotrypsin-like proteases by superposition of consecutive β-strands and C-terminal α-helices (fig.3). The most striking feature of the alignment was the equivalent location of the catalytic His and Ser residues of chymotrypsin-like proteases and the respective conserved residues of $3C^{Pro}$, i.e. His adjacent to the C-terminus of strand C and Cys(Ser) in the loop preceding strand J (fig.3). Moreover, Glu(Asp) at the N-terminus of strand F in $3C^{Pro}$, which was conserved also in plant viral proteases (fig.1B), appeared to match the third catalytic residue of chymotrypsin-like proteases, Asp^{102} (fig.3). Also notable was the coincidence or homologous replacement of a number of addi-

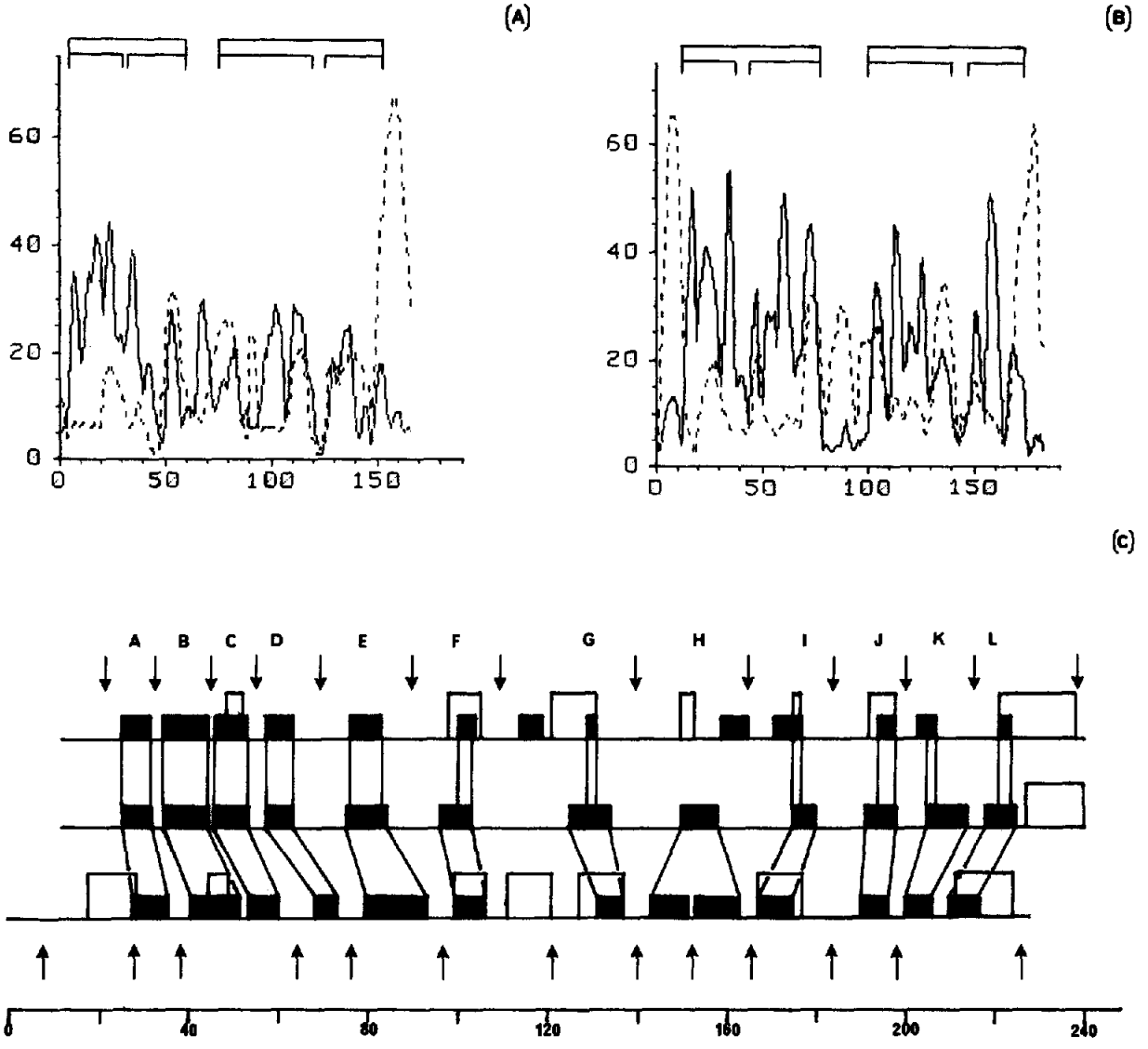


Fig.2. Comparison of secondary structures of $3C^{Pro}$ and chymotrypsin-like proteases. (A) An averaged plot of secondary structure probabilities for chymotrypsin-like proteases. For each position (X axis) containing no gaps in sequence alignment [21], average β- and α-potentials (Y axis) were calculated as follows: P = [(SGPA + SGPB)/2 + (CHY + TRP + ELA)/3]/2 where the protease abbreviations stand for respective β- and α-potentials calculated by the ALBEAL program. Solid line: β-strand probability; broken line: α-helix probability. Horizontal brackets delineate domains and half-domains (see text). (B) An analogous plot for $3C^{Pro}$. Here, P = [(PV1 + HRV14 + HRV2 + HRV1b + ECHO)/5 + (EMCV + TMEV)/2 + FMDV + HAV]/4. Designations as in A. (C) A schematic linear representation of secondary structures. Filled rectangles, β-strands; empty rectangles, α-helices. Upper row: average predicted profile for chymotrypsin-like proteases; middle row: average profile derived from X-ray data for chymotrypsin-like proteases; bottom row: average predicted profile for $3C^{Pro}$. β-strands are designated A to L according to [13]. Arrows indicate regions where positions containing gaps (presumably indicating deletions and insertions) were omitted from the profile calculations (panels A and B).

tional amino acid residues, mainly hydrophobic, as should be expected of β-strands. Quantitative evaluation of the alignment by use of the MDM78 matrix demonstrated that the level of similarity between $3C^{Pro}$ and eukaryotic chymotrypsin-like proteases was not lower than that between the lat-

Fig.3. Sequence alignment of $3C^{Pro}$ and chymotrypsin-like proteases based on secondary structure superposition. 12 consecutive β-strands designated as in fig.2, C-terminal helices and some adjacent conserved regions were aligned. The number of residues in each secondary structure element is shown. For $3C^{Pro}$ the data are from the prediction shown in fig.2, and for chymotrypsin-like proteases from X-ray analysis. Strands I and L which were predicted ambiguously in $3C^{Pro}$ are shown in parentheses. For some long β-strands and for the C-terminal helices only partial sequences are included. Numbers stand for lengths of spacers and terminal extensions. Amino acid residues having at least one identical or homologous (see legend to fig.2B) counterpart in the other sequence set are designated by capitals. Colons: positions occupied by identical or homologous residues in at least 1/2 of the sequences of each of the sets; asterisks: putative catalytic residues.

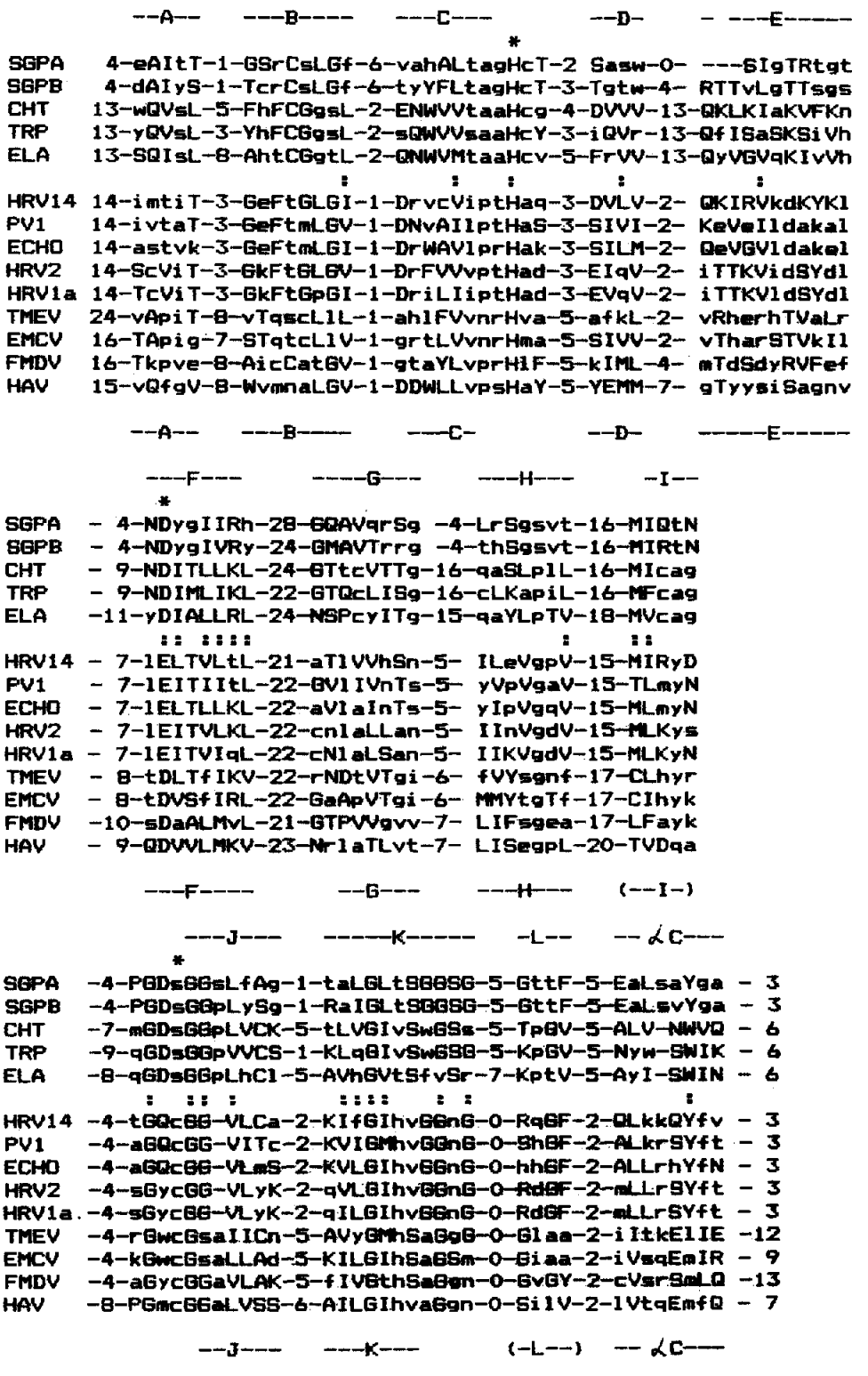
ter and prokaryotic proteases (not shown). These observations made us hypothesize that His^{56}, $Glu(Asp)^{102}$ and Cys^{202} of $3C^{Pro}$ might constitute a catalytic triad similar to that of chymotrypsin-like proteases [17]. In chymotrypsin-like proteases, substitution of Glu for Asp^{102} has not been described (cf. [2,21]). However, it is possible to speculate that the presence of Cys in the place of Ser^{195} might confer additional flexibility to the catalytic center, permitting both Glu and Asp as members of the catalytic triad.

4. A CHYMOTRYPSIN-LIKE STRUCTURAL FOLD IN $3C^{Pro}$

The above observations strongly suggest that $3C^{Pro}$ should be similar to chymotrypsin-like proteases also at the level of the tertiary structure. We hypothesize that the $3C^{Pro}$ molecule consists of two twisted antiparallel β-barrels connected by a long loop. The hydrophobic core of each barrel is constituted by 6 β-strands. Secondary structure predictions for non-picornaviral 3C-like proteases (not shown) and their sequence similarity to $3C^{Pro}$ (fig.1B) suggest that these enzymes might form an analogous fold. The proposed β-sheet topology of poliovirus $3C^{Pro}$ is shown in fig.4. This arrangement of β-strands is compatible with recently reported data on site-directed and random mutagenesis of this protease. Specifically, substitution of Val or Ala for Gly^{51} (hereafter in this section the poliovirus numbering is used), presumably disrupting a β-turn, was lethal, whereas substitution of Asp (a residue frequently occurring in β-turns [23]) in the same position resulted in a viable virus [24]. Substitutions in strands E and F which could cause local deformations of the β-sheet exerted relatively mild effects on viral reproduction [24–26], and a substitution of Ser for Cys^{153} in strand J appeared to be without effect on the activity of $3C^{Pro}$ expressed in *E. coli* [9]. Moreover, the processing defects inflicted by substitutions in strands E and F were similar (namely, impairment of the cleavage at the C-terminus of $3C^{Pro}$ itself [24,25]), in accord with our proposal that these strands might interact with each other in native $3C^{Pro}$.

Based on the analogy with chymotrypsin-like proteases, we predict that the three putative catalytic residues, two of which, His and Asp(Glu), reside in the N-terminal domain of $3C^{Pro}$, and the 3rd, Cys(Ser), in the C-terminal one, should be juxtaposed in the interdomain cleft. This suggests that the mechanism of peptide bond cleavage catalysis by $3C^{Pro}$ may be similar to that described for chymotrypsin-like proteases involving formation of a three-residue charge-transfer system [27,28]. The involvement of Cys (in all 3C-like enzymes except the putative protease of SBMV) and Glu (in some 3C-like proteases) in such a system is a novel theme expanding the existing ideas of proteolysis mechanisms. Despite the similarity in the positioning of the (putative) catalytic residues in $3C^{Pro}$ and chymotrypsin-like enzymes, Cys and Ser are not as easily interchangeable in the triad as could be imagined. Thus Cys^{147} to Ser substitution in poliovirus $3C^{Pro}$ completely abolished its protease activity [9]. In nature, however, such substitutions appear to work as exemplified by the putative protease of SBMV; presumably this is gratified by some compensatory substitution(s).

Substrate-binding pockets of chymotrypsin-like proteases are formed by three non-contiguous segments [29–31]. Two conserved sites in the strands K and L were implicated in supporting the 'walls' of the cavity, while a more variable site in the loop preceding the catalytic Ser is thought to constitute its base, being the main determinant of cleavage specificity. It is tempting to speculate that equivalent segments of 3C-like proteases are also involved in substrate binding. This is especially


--A-- ---B---- ---C--- --D- - ---E-----
SGPA 4-eAItT-1-GSrCsLGf-6-vahALtagHcT-2 Sasw-O- ---GIgTRtgt
SGPB 4-dAIyS-1-TcrCsLGf-6-tyYFLtagHcT-3-Tgtw-4- RTTvLgTTsgs
CHT 13-wQVsL-5-FhFCGgsL-2-ENWVVtaaHcg-4-DVVV-13-QKLKIaKVFKn
TRP 13-yQVsL-3-YhFCGgsL-2-sQWVVsaaHcY-3-iQVr-13-QfISaSKSiVh
ELA 13-SQIsL-8-AhtCGgtL-2-QNWVMtaaHcv-5-FrVV-13-QyVGVqKIvVh
HRV14 14-imtiT-3-GeFtGLGI-1-DrvcViptHaq-3-DVLV-2- QKIRVkdKYKl
PV1 14-ivtaT-3-GeFtmLGV-1-DNvAIlptHaS-3-SIVI-2- KeVeIldakal
ECHO 14-astvk-3-GeFtmLGI-1-DrWAVlprHak-3-SILM-2- QeVGVldakel
HRV2 14-ScViT-3-GkFtGLGV-1-DrFVVvptHad-3-EIqV-2- iTTKVidSYdl
HRV1a 14-TcViT-3-GkFtGpGI-1-DriLIiptHad-3-EVqV-2- iTTKVldSYdl
TMEV 24-vApiT-8-vTqscLlL-1-ahlFVvnrHva-5-afkL-2- vRherhTVaLr
EMCV 16-TApig-7-STqtcLlV-1-grtLVvnrHma-5-SIVV-2- vTharSTVkIl
FMDV 16-Tkpve-8-AicCatGV-1-gtaYLvprHlF-5-kIML-4- mTdSdyRVFef
HAV 15-vQfgV-8-WvmnaLGV-1-DDWLLvpsHaY-5-YEMM-7- gTyysiSagnv
--A-- ---B----- ---C- --D- -----E-----
---F--- ----G--- ---H--- -I--
SGPA - 4-NDygIIRh-28-GQAVqrSg -4-LrSgsvt-16-MIQtN
SGPB - 4-NDygIVRy-24-GMAVTrrg -4-thSgsvt-16-MIRtN
CHT - 9-NDITLLKL-24-GTtcVTTg-16-qaSLplL-16-MIcag
TRP - 9-NDIMLIKL-22-GTQcLISg-16-cLKapiL-16-MFcag
ELA -11-yDIALLRL-24-NSPcyITg-15-qaYLpTV-18-MVcag
HRV14 - 7-lELTVLtL-21-aTlVVhSn-5- ILeVgpV-15-MIRyD
PV1 - 7-lEITIItL-22-GVlIVnTs-5- yVpVgaV-15-TLmyN
ECHO - 7-lELTLLKL-22-aVlaInTs-5- yIpVgqV-15-MLmyN
HRV2 - 7-lEITVLKL-22-cnlaLLan-5- IInVgdV-15-MLKys
HRV1a - 7-lEITVIqL-22-cNlaLSan-5- IIKVgdV-15-MLKyN
TMEV - 8-tDLTfIKV-22-rNDtVTgi-6- fVYsgnf-17-CLhyr
EMCV - 8-tDVSfIRL-22-GaApVTgi-6- MMYtgTf-17-CIhyk
FMDV -10-sDaALMvL-21-GTPVVgvv-7- LIFsgea-17-LFayk
HAV - 9-QDVVLMKV-23-NrlaTLvt-7- LISegpL-20-TVDqa
---F---- --G--- ---H--- (--I-)
---J--- -----K----- -L-- -- ∠C---
SGPA -4-PGDsGGsLfAg-1-taLGLtSGGSG-5-GttF-5-EaLsaYga - 3
SGPB -4-PGDsGGpLySg-1-RaIGLtSGGSG-5-GttF-5-EaLsvYga - 3
CHT -7-mGDsGGpLVCK-5-tLVGIvSwGSs-5-TpGV-5-ALV-NWVQ - 6
TRP -9-qGDsGGpVVCS-1-KLqGIvSwGSG-5-KpGV-5-Nyw-SWIK - 6
ELA -8-qGDsGGpLhCl-5-AVhGVtSfvSr-7-KptV-5-AyI-SWIN - 6
HRV14 -4-tGQcGG-VLCa-2-KIfGIhvGGnG-O-RqGF-2-QLkkQYfv - 3
PV1 -4-aGQcGG-VITc-2-KVIGMhvGGnG-O-ShGF-2-ALkrSYft - 3
ECHO -4-aGQcGG-VLmS-2-KVLGIhvGGnG-O-hhGF-2-ALLrhYfN - 3
HRV2 -4-sGycGG-VLyK-2-qVLGIhvGGnG-O-RdGF-2-mLLrSYft - 3
HRV1a -4-sGycGG-VLyK-2-qILGIhvGGnG-O-RdGF-2-mLLrSYft - 3
TMEV -4-rGwcGsaIICn-5-AVyGMhSaGgG-O-Glaa-2-iItkElIE -12
EMCV -4-kGwcGsaLLAd-5-KILGIhSaGSm-O-Giaa-2-iVsqEmIR - 9
FMDV -4-aGycGGaVLAK-5-fIVGthSaGgn-O-GvGY-2-cVsrSmLQ -13
HAV -8-PGmcGGaLVSS-6-AILGIhvaGgn-O-SilV-2-lVtqEmfQ - 7
--J--- ---K--- (-L--) -- ∠C---

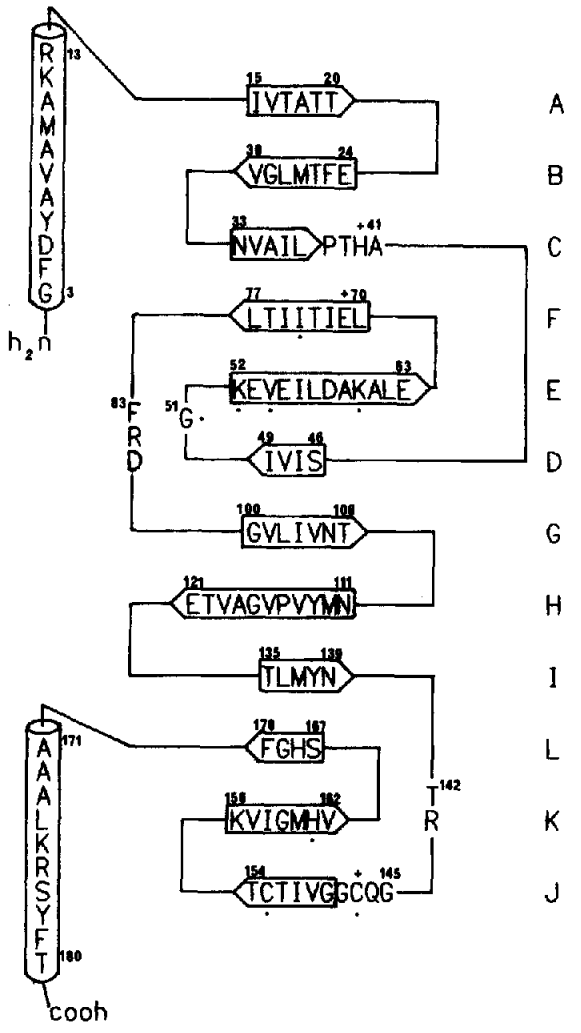


Fig.4. Proposed β-sheet topology for $3C^{Pro}$ of PV1. Arrow-headed rectangles: β-strands; cylinders: α-helices. To delineate secondary structure elements, predictions for $3C^{Pro}$ of PV were used, with some corrections based on average predictions. The numbering is for $3C^{Pro}$ of PV and does not correspond to that in fig.1B. In the right part of the figure strands are designated as in figs 2C and 3. Residues outside secondary structure elements which are (partially) conserved in other $3C^{Pro}$ are also shown. Plus signs: putative catalytic residues; dots: residues subjected to mutagenesis (see text).

plausible as strand K is highly conserved throughout the family, including nearly invariant His and Gly residues, and strand L also contains a partially conserved Gly residue, similarly to the substrate-binding site of chymotrypsin-like proteases (figs 1B and 3). The importance of His^{161} for substrate binding might explain the inactivation of poliovirus $3C^{Pro}$ upon substitution of Gly for this residue [9]. As for the putative specificity site, in 3C-like proteases it contains a partially conserved dipeptide $ThrArg^{143}$(Lys) (positions 193–194 in fig.1B). This is interesting in view of the conserved, despite the high divergence of enzymes themselves, cleavage specificity of 3C-like proteases which act primarily at Q, E/G,S dipeptides [7]. On the other hand, variations observed in this segment of the proteases may account for different requirements to the residues flanking cleavage sites revealed upon site-directed mutagenesis of picornaviral and potyviral polyproteins [32,33]. Certainly, other regions of $3C^{Pro}$ might also contribute to their specificity, as emphasized by the above-mentioned effects of mutations in strands E and F on cleavage at specific sites.

Along with these similarities, considerable structural and functional differences seem to exist between chymotrypsin-like and 3C-like proteases. Both types of enzymes are generated via proteolytic processing of precursors, which involves liberation of the N-terminus in cellular proteases, and of both termini in 3C-like proteases [2,7]. However, chymotrypsin-like protease precursors have only very low activity, activation achieved through formation of an electrostatic bridge between the new N-terminal residue (which is always Ile or Val) and the invariant Asp residue adjacent to the catalytic Ser. In 3C-like proteases, which are cleaved from viral polyproteins autocatalytically [34], this mechanism is not operational. Accordingly, the above residue pair is not conserved in this family, the position near the (putative) catalytic Cys becoming variable (fig.1B). Another notable difference is the absence, in 3C-like enzymes, of the system of disulphide bonds which are conserved in chymotrypsin-like proteases, making their structure rigid [17]. A highly conserved site in $3C^{Pro}$ is the sequence PheArg(Lys)Asp^{85} (positions 122–125 in fig.1B). Previously it was suggested that Asp^{85} could be involved in catalysis [8]. However, this is unlikely as in our model this sequence lies in the loop connecting the two domains (fig.4). Possibly it may function by binding some ligand other than the substrate.

5. EVOLUTIONARY IMPLICATIONS

The significant structural similarity between 3C-

like and chymotrypsin-like proteases strongly favors their divergent rather than convergent evolutionary origin. Moreover, as the proteases of the two families probably share the two-barrel organization and the positions of the (putative) catalytic residues, it is logical to suggest that: (i) their common ancestor has already been a protease, and (ii) the divergence of the families succeeded the initial duplication leading to the two-domain structure (cf. [22]). It is not clear at present what was the nature of the hypothetical ancestor protease. However, we have argued in previous papers that $3C^{Pro}$ have some features which could be expected in a primordial protease [11,35]. Interestingly, it has been very recently proposed, starting from quite different observations, that Cys might be the predecessor of Ser in the catalytic sites of enzymes of several classes [36]. That enzymes possibly similar to the ancestral forms are found in positive strand RNA viruses, is intriguing in view of the ideas relating their genomes to primordial genetic systems [35,37].

Acknowledgements: The authors are grateful to Dr A.V. Finkelstein for help with secondary structure prediction, to Drs L.I. Brodsky, K.M. Chumakov and A.L. Drachev for help with computer programming, and to Professor V.I. Agol for useful discussions. Thanks are also due to Dr C. Fritsch for sending a preprint of his work.

REFERENCES

[1] Antonov, V.K. (1983) The Chemistry of Proteolysis, Nauka, Moscow, in Russian.
[2] Neurath, H. (1984) Science 224, 350–357.
[3] Pelham, H.R.B. (1978) Eur. J. Biochem. 85, 457–462.
[4] Gorbalenya, A.E. and Svitkin, Yu.V. (1983) Biokhimia 48, 385–395.
[5] Argos, P., Kamer, G., Nicklin, M.J.H. and Wimmer, E. (1984) Nucleic Acids Res. 12, 7251–7267.
[6] Nicklin, M.J.H., Toyoda, H., Murray, M.G. and Wimmer, E. (1986) Bio/Technology 4, 33–42.
[7] Palmenderg, A.C. (1987) J. Cell. Biochem. 33, 191–198.
[8] Wellink, J.E. and Van Kammen, A. (1988) Arch. Virol. 98, 1–26.
[9] Ivanoff, L.A., Towatari, T., Ray, J., Korant, B.D. and Petteway, S.R. (1986) Proc. Natl. Acad. Sci. USA 83, 5392–5396.
[10] Blinov, V.M., Gorbalenya, A.E. and Donchenko, A.P. (1984) Dokl. Akad. Nauk SSSR 279, 502–505.
[11] Gorbalenya, A.E., Blinov, V.M. and Donchenko, A.P. (1986) FEBS Lett. 194, 253–257.
[12] Greif, C., Hemmer, O. and Fritsch, C. (1988) J. Gen. Virol. 69, 1517–1529.
[13] Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1988) FEBS Lett. 236, 287–290.
[14] Pozdnyakov, V.I. and Pankov, Yu.A. (1981) Int. J. Peptide Protein Res. 17, 284–291.
[15] Gorbalenya, A.E., Blinov, V.M., Donchenko, A.P. and Koonin, E.V. (1988) J. Mol. Evol. 28, in press.
[16] Doolittle, R.F. (1986) Of URFs and ORFs: A Primer on How to Analyze Derived Amino Acid Sequences, University Science Books, Mill Valley, CA.
[17] Greer, J.J. (1981) J. Mol. Biol. 153, 1027–1042.
[18] Finkelstein, A.V. (1975) Dokl. Akad. Nauk SSSR 223, 744–747.
[19] Ptitsyn, O.B. and Finkelstein, A.V. (1983) Biopolymers 22, 15–25.
[20] Richardson, J.J. (1981) Adv. Protein Chem. 34, 167–339.
[21] Craik, C.S., Rutter, W.J. and Fletterick, R. (1983) Science 220, 1125–1129.
[22] McLachlan, A.D. (1979) J. Mol. Biol. 128, 49–79.
[23] Schulz, G.E. and Schirmer, R.H. (1979) Principles of Protein Structure, Springer, New York.
[24] Dewalt, P.G. and Semler, B. (1987) J. Virol. 61, 2162–2170.
[25] Kean, K.M., Agut, H., Fichot, O., Wimmer, E. and Girard, M. (1988) Virology 163, 330–340.
[26] Dewalt, P.G. and Semler, B. (1988) Abstracts of the 1988 ICN-UCI International Conference on Virology, p.7.
[27] Blow, D.M., Birktoft, J.J. and Hartley, B.S. (1969) Nature 221, 337–340.
[28] Sprang, S., Standing, T., Fletterick, R.J., Stroud, R.M., Finer-Moore, J., Xuong, N.-H., Hamlin, R., Rutter, W.J. and Craik, C.S. (1987) Science 237, 905–909.
[29] Craik, C.S., Largman, C., Fletcher, T., Roszniak, S., Barr, P., Fletterick, R. and Rutter, W.J. (1985) Science 228, 291–297.
[30] Graf, L., Craik, C.S., Patthy, A., Roszniak, S., Fletterick, R.J. and Rutter, W.J. (1987) Biochemistry 26, 2616–2623.
[31] Delbaere, L.T.J. and Brayer, G.D. (1985) J. Mol. Biol. 183, 89–103.
[32] Parks, G.D. and Palmenberg, A.C. (1987) J. Virol. 61, 3680–3687.
[33] Dougherty, W.G., Carrington, J.C., Cary, S.M. and Purks, T. (1988) EMBO J. 7, 1281–1287.
[34] Palmenberg, A.C. and Rueckert, R.R. (1982) J. Virol. 41, 244–249.
[35] Gorbalenya, A.E., Donchenko, A.P. and Blinov, V.M. (1986) Mol. Genet. 1, 36–41.
[36] Brenner, S. (1988) Nature 334, 528–530.
[37] Eigen, M. and Schuster, P. (1979) The Hypercycle: Principle of Natural Self-Organization, Springer, Berlin.
[38] Stanway, G., Hughes, P., Mountford, R.C., Minor, P.D. and Almond, J.W. (1984) Nucleic Acids Res. 11, 5629–5643.
[39] Racaniello, V.C. and Baltimore, D. (1981) Proc. Natl. Acad. Sci. USA 78, 4887–4891.
[40] Werner, G., Rosenwirth, B., Bauer, E., Seifert, J.-M., Werner, J.-F. and Besemer, J. (1986) J. Virol. 57, 1084–1093.
[41] Skern, T., Sommergruber, W., Blaas, D., Gruendler, P., Fraundorfer, F., Pieler, C., Fogy, I. and Kuechler, E. (1985) Nucleic Acids Res. 12, 7859–7875.

[42] Hughes, P.J., North, C., Jellis, C., Minor, P.D. and Stanway, G. (1988) J. Gen. Virol. 69, 49–58.
[43] Duechler, M., Skern, T., Sommergruber, W., Neubauer, C., Gruendler, P., Fogy, I., Blaas, D. and Kuechler, E. (1987) Proc. Natl. Acad. Sci. USA 84, 2605–2609.
[44] Lindberg, A.M., Stalhandske, P.O.K. and Pettersson, U. (1987) Virology 156, 50–63.
[45] Earle, J.A.P., Skuce, R.A., Fleming, C.S., Hoey, E.M. and Martin, S.J. (1988) J. Gen. Virol. 69, 253–263.
[47] Pevear, D.C., Calenoff, M., Rozhon, E. and Lipton, H.L. (1987) J. Virol. 61, 1507–1516.
[47] 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.
[48] Carrol, A.R., Rowlands, D.J. and Clarke, B.E. (1984) Nucleic Acids Res. 12, 2461–2472.
[49] Najarian, R., Caput, D., Gee, W., Potter, S.J., Renard, A., Merryweather, J., Van Nest, G. and Dina, D. (1985) Proc. Natl. Acad. Sci. USA 82, 2627–2631.
[50] Lomonosoff, G.P. and Shanks, M. (1983) EMBO J. 2, 2253–2258.
[51] Domier, L.L., Franklin, K.M., Shahabuddin, M., Hellmann, G.M., Overmeyer, J.H., Hiremath, S.T., Siaw, M.F.E., Lomonosoff, G.P., Shaw, J.G. and Rhoads, R.E. (1986) Nucleic Acids Res. 14, 5417–5430.
[52] Allison, R., Johnston, R.E. and Dougherty, W.G. (1986) Virology 154, 9–20.
[53] Wu, S., Rinehart, C.A. and Kaesberg, P. (1987) Virology 161, 73–80.