Volume 288, number 1,2, 201–205 FEBS 10058 August 1991

ADONIS 001457939100768X

Putative papain-related thiol proteases of positive-strand RNA viruses

Identification of rubi- and aphthovirus proteases and delineation of a novel conserved domain associated with proteases of rubi-, α- and coronaviruses

Alexander E. Gorbalenya[1], Eugene V. Koonin[2] and Michael M.-C. Lai[3]

[1]*Institute of Poliomyelitis and Viral Encephalitides, USSR Academy of Medical Sciences, 142782 Moscow Region, USSR,* [2]*Institute of Microbiology, USSR Academy of Sciences, 117811 Moscow, USSR* and [3]*Howard Hughes Medical Institute and Department of Microbiology, University of Southern California School of Medicine, Los Angeles, CA 90033, USA*

Received 13 June 1991

A computer-assisted comparative analysis of the amino acid sequences of (putative) thiol proteases encoded by the genomes of several diverse groups of positive-stranded RNA viruses and distantly related to the family of cellular papain-like proteases is presented. A high level of similarity was detected between the leader protease of foot-and-mouth-disease virus and the protease of murine hepatitis coronavirus which cleaves the N-terminal p28 protein from the polyprotein. Statistically significant alignment of a portion of the rubella virus polyprotein with cellular papain-like proteases was obtained, leading to tentative identification of the papain-like protease as the enzyme mediating processing of the non-structural proteins of this virus. Specific grouping between the sequences of the proteases of α-viruses, and poty- and bymoviruses was revealed. It was noted that papain-like proteases of positive-stranded RNA viruses are much more variable both in their sequences and in genomic locations than chymotrypsin-related proteases found in the same virus class. A novel conserved domain of unknown function has also been identified which flanks the papain-like proteases of α-, rubi- and coronaviruses.

Papain-like protease; RNA virus; Polyprotein processing; Sequence motif; Catalytic center

1. INTRODUCTION

Polyprotein processing is the strategy employed by a number of groups of positive-stranded RNA viruses for genome expression (for review see [1]). Processing of membrane proteins of enveloped viruses is usually mediated by cellular proteases, whereas processing of non-membrane proteins by virus-encoded proteases. A large superfamily of virus-encoded proteases related to chymotrypsin-like cellular serine proteases has been described [2–5]. Some of these viral proteases have the substitution of Cys for the principal catalytic Ser, not found in cellular enzymes, comprising a unique group of cysteine proteases.

Only very recently, the existence of 'classical' cysteine proteases related to papain-like cellular enzymes has been claimed for several positive-stranded RNA viruses. The essential Cys and His residues were identified in the potyvirus CI [6], α-virus nsP2 [7,8] and murine coronavirus 'leader' (L-pro) ([9], and Baker, et al., submitted) proteases by site-directed mutagenesis. The relative positions of these residues in the respective proteins and their amino acid contexts resemble those of the catalytic residues in the papain-like proteases [6–9]. Two other putative papain-like proteases were revealed in the polyproteins of coronaviruses by comparative sequence analysis. The putative 'main' protease (M-pro) is related to MHV L-pro and is conserved in both IBV and MHV polyproteins [10], and the putative SPL protease shares similarity to the protease from *Streptococcus* and is present in IBV polyprotein only [11].

Two proteases of picornaviruses, L-pro of FMDV and VPO, have not been characterized with respect to the type of the catalytic residues [1]. In addition, the presence of proteases could be suspected in several

Correspondence address: A.E. Gorbalenya, Inst. of Poliomyelitis and Viral Encephalitides, USSR Academy of Medical Sciences, Post-office Inst. of Poliomyelitis, 142782 Moscow Region, USSR.

Abbreviations: SPP, *Streptococcus pyogenes* peptidase A; CP1 and CP2, *Dictyostelium discoideum* cysteine proteinase 1 and 2; actin, *Actinidia chinensis actinidin*; papain and omega, *Carica papaya* proteinase I and III, respectively; aleur, *Hordeum vulgare* aleurain; Cat.H and L, *Rattus norvegicus* cathepsins H and L; Cat.B, *Homo sapiens* cathepsin B; SH-EP, *Vigna mungo* cysteine endopeptidase; bromel, *Ananas comosus* bromelain; derpt, *Dermatophagoides pteronyssinus* major mite fecal allergen; calp, *Mus musculus* calpain; SFV, Semliki forest virus; SNBV, Sindbis virus; VEEV, Venezuelan equine encephalomyelitis virus; ONNV, O'Nyong-Nyong virus; RRV, Ross River virus; MidV, Middelburg virus (alphaviruses); PVY, potato virus Y; PPV, plum pox virus; TEV, tobacco etch virus; TVMV, tobacco vein mottling virus (potyviruses); BaYMV, barley yellow mosaic virus (bymovirus); MHV, murine hepatitis virus; IBV, avian bronchitis virus (coronaviruses); FMDV A10, foot-and-mouth-disease virus A10 strain (aphthovirus); RuV, rubella virus (rubivirus); HC, helper component; M-pro, 'main' protease; L-pro, 'leader', or accessory protease (see text); SPL, 'Streptococcus-like' protease; CI, cylindrical inclusion (potyvirus protein).

other viruses (e.g. rubi-, tymo- and furoviruses) encoding very large proteins [1,12]. Here we identify a putative protease (M-pro) encoded by the rubella virus genome and show that this protease and aphthovirus L-pro belong to the papain-like protease group. A novel conserved domain associated with the (putative) papain-like proteases of α-, corona- and rubiviruses is described.

2. MATERIALS AND METHODS

2.1. *Amino acid sequences*

All sequences were from the SWISSPROT data bank Release 16, except for MHV [10], RuV [12], IBV [13], and BaYMV [14].

2.2. *Comparative sequence analysis*

Amino acid sequences were compared by the program OPTAL as previously described [15] using the amino acid residue comparison matrix MDM78. Program OPTAL, implementing the Sankoff algorithm, generates multiple sequence alignments in a stepwise manner and calculates adjusted alignment scores as the number of standard deviations (SD) over the mean of 25 random simulations. The program DotHelix [16], a module of the GENEBEE program package for biopolymer sequence analysis [17], was used to build up complete local similarity maps for pairs of amino acids sequences.

3. RESULTS AND DISCUSSION

The sequences of cellular and viral papain-like proteases are quite variable; the only reliable conserved region is a stretch of approximately 10 amino acid residues centering at the catalytic Cys ([18], and unpublished observations). The sequences of positive-stranded RNA viral proteins which could be suspected for protease activity were searched for segments resembling this conserved stretch. The pieces of FMDV and RuV polyproteins selected in this way were analyzed in detail.

3.1. *Aphthovirus and coronavirus leader proteases are related*

Pronounced similarity was found between the segments around the putative catalytic Cys of coronavirus proteases (particularly MHV L-pro), and a sequence located near the N terminus of FMDV L-pro and containing a Cys residue conserved in all sequenced FMDV strains ([19]; Fig. 1). When the entire polyproteins of MHV and FMDV (more than 6900 and 2300 amino acids, respectively) were compared by program DotHelix, these segments were found to be the most closely related, with their alignment score being about 8 SD above the random expectation (not shown). These observations allow us to predict the catalytic Cys residue of the L-pro of FMDV. It has been shown that the substitution of Ile for Thr in the vicinity of this residue (Fig. 1) abolished the protease activity, unlike four other mutations in the N-terminal half of L-pro [19]. Due to the weak sequence conservation around the catalytic His of the identified viral papain-like proteases, all three histidines, which are conserved in the sequenced FMDV strains, remain, for the meantime, candidates for the role of the catalytic residue.

3.2. *RuV polyprotein contains a protease-like domain*

A statistically significant alignment was obtained between a segment of the RuV non-structural polyprotein and cellular papain-like proteases, with sequence conservation around the (putative) catalytic Cys and His residues (Fig. 1). The sequence similarity between the rubivirus polyprotein fragment (residues from 1125 to 1320) and 12 eucaryotic proteases could be characterized by scores in the range between 4.5 and 12.2 SD. These values were obtained upon aligning the RuV sequence with the cellular ones with or without the omission of inserts present in some of the cellular proteases, respectively (not shown). This identifies the putative rubella virus protease and demonstrates the so far most pronounced similarity between cellular and viral papain-like proteases. It was noted, however, that not all of the sequence segments highly conserved in cellular enzymes are retained in the putative protease of RuV. In particular, of the six Cys residues conserved in the cellular proteases, only two (including the catalytic one) are found in the rubella virus protein, suggesting that the two characteristic disulfide bridges of the cellular proteases are not conserved in the viral counterpart. Among the viral proteases, RuV M-pro shares the most convincing similarity with IBV M-pro in the region around the putative catalytic Cys residue (Fig. 1).

3.3. *Papain-like proteases of α-viruses, and poty- and bymoviruses constitute a distinct group*

Our analysis revealed a previously unnoticed resemblance between the papain-like proteases of alphaviruses, on the one hand, and potyviruses and the closely related bymovirus, on the other hand (Fig. 1). The adjusted alignment score was 5.5 SD for approximately 150 amino acid residue domains of the two protease groups. The alignment showed well-conserved spacing of the catalytic residues and highlighted several additional invariant and conserved residues characteristic specifically for these two enzyme groups (Fig. 1). A notable common feature of the α-virus and poty-(bymo)virus proteases is their specificity towards pairs of small amino acid residues [6,7].

3.4. *A novel conserved domain associated with the papain-like proteases of rubi-, α- and coronaviruses*

Analysis of the sequences of viral polyproteins surrounding the putative papain-like proteases unexpectedly led to the discovery of a new conserved domain. This domain has been described previously as the most similar segment in the α-virus and rubivirus polyproteins [12]. Independently, a strongly conserved region adjacent to the putative papain-like protease(s) was identified upon comparison of the polyproteins of

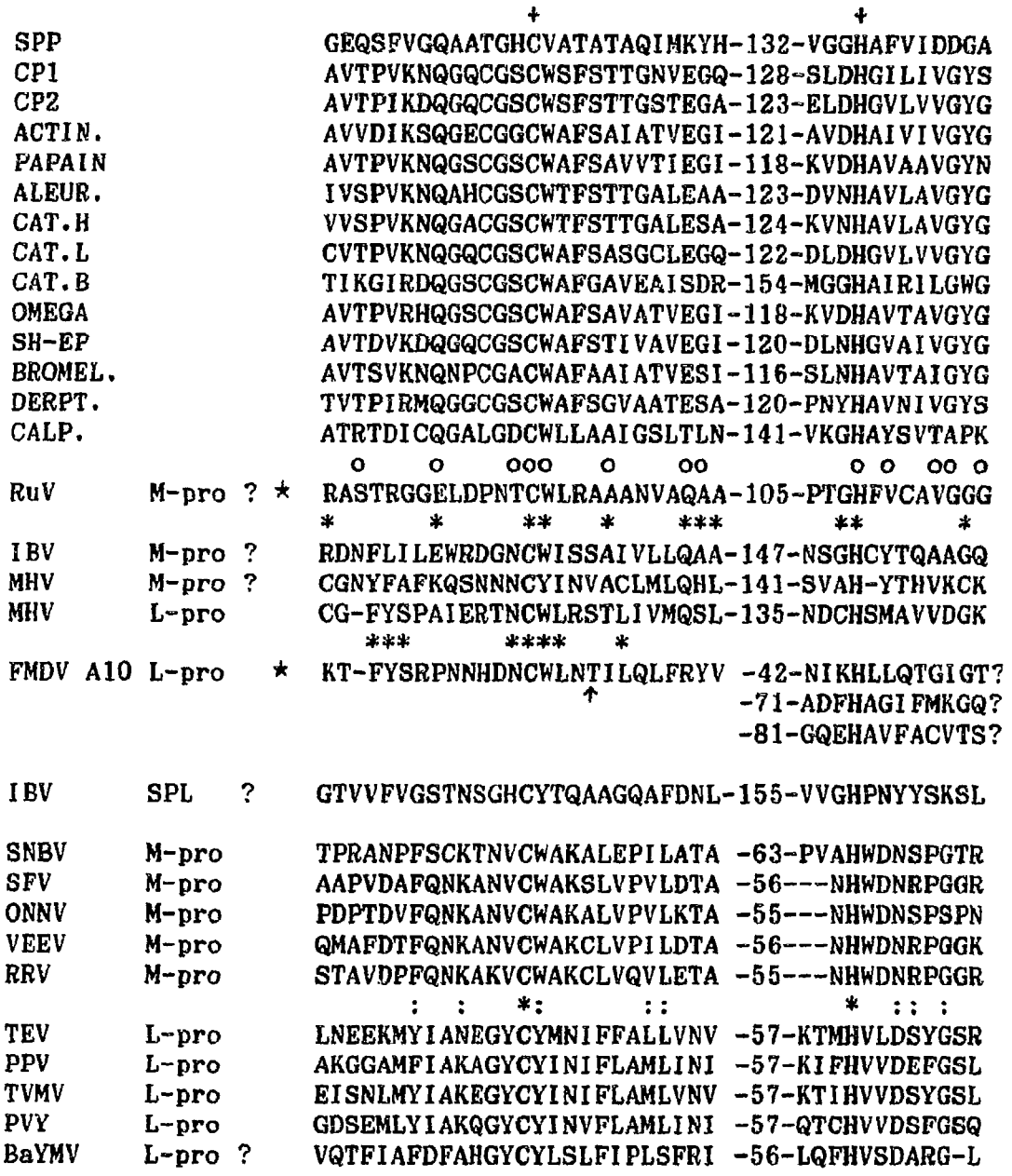

Fig. 1. Alignment of the segments around the catalytic Cys and His residues of cellular and (putative) positive-stranded RNA virus papain-like proteases. The numbers of amino acid residues separating the aligned segments are indicated. Plus, the (putative) catalytic residues; circles, residues conserved in the putative protease of RuV and most of the cellular papain-like proteases; asterisks, identities between the sequences around the putative catalytic residues of the proteases of RuV and IBV (M-pro), FMDV and MHV (L-pro), and the groups of the proteases of α-viruses and poty-/bymoviruses; colons, residues partially conserved in the latter two groups; question marks, the proteases which have been identified only by amino acid sequence comparison; bold asterisk, the proteases which have been added to the papain-like group in this study; arrow, the residue which was found to be replaced in the FMDV mutant lacking the protease activity.

the coronaviruses IBV and MHV [10]. Comparison of the two alignments obtained this way has suggested that all these domains constitute a single family (Fig. 2). Within this family, approximately the same level of similarity was observed between the sequences of all three groups of viruses. Screening of the Swissprot database provided no clue as to the possible function of this conserved domain (hereafter designated 'X' domain).

3.5. *Concluding overview of viral papain-like proteases*

Together with the previously reported data on proteases of α-, corona- and potyviruses [6–11,20], these observations delineated the set of (putative) positive-stranded RNA virus papain-like proteases. As a whole, the sequences of these proteases around the proposed catalytic Cys and His residues have relatively little in common, except for the notable CW(Y) dipeptide. The sequences around the catalytic His residues of cellular,

```
                              I                           II         III           IV

SNBV   (18-  115)AVVNAANPLGRPGEGVCRAI 27 VIHAVGPD 7 EALKLLQNAYHAV 11VAIPLLSTGIYA
SFV    (18-  115)AVVNAANARGTVGDGVCRAV 27 VIHAVAPN 7 EGDRELAAVYRAV 11VAIPLLSTGVFS
ONNV   (18-  115)CVVNAANPRGVPGDGVCKAV 27 VIHAVGPN 7 EGDRELASVYREV 11VAIPLLSTGVYS
RRV    (18-  115)AVVNAANAKGTVGDGVCRAV 27 VIHAVAPN 7 EGDRELAAAYRAV 11VAIPLLSTGVFS
MidV   (18-  113)VLVNQLGVNNKVCDGVCRAM 26 IVHAYCPN 7 VADADLAAVYRAV 10MAIPLLSTGTFA
VEEV   (18-  115)VIINAANSKGQPGGGVCGAL 27 IIHAVGPN 7 EGDKQLAEAYESI 11VAIPLLSTGIFS

RuV   (834-  940)VVVNAANEGLLAGSGVCGAI 35 IIHAVAPR 9 EGEALLERAYRSI 11VACPLLGAGVYG

IBV  (1038-1138)CIVNAANEHMTHGSGVAKAI 37 VNNVVGPR 5 LHEKLVA-AYKNV  7YVVPVLSLGIFG
MHV  (1358-1459)VIVNPANGRMAHGAGVAGAI 39 VLNIVGPD 6 ECYSLLERAYQHI  5VVTTLISAGIFS
CONSENSUS        ++vNaan    g GV gA+    v+  vaP   e e ll  aYr v    v  pll  G+fg
                                        i   g       d      k i                ya
```

Fig. 2. Alignment of the 'X' domains of α-, rubi- and coronaviruses. The alignment was generated using the OPTAL program, yielding the score of over 6 SD for each step. Only the four most conserved segments are shown. The numbering is given for α-virus nsP3 protein and for rubi- and coronaviruses non-structural polyproteins. Consensus: upper case, invariant residues; lower case, residues found in at least one sequence of each of the three virus groups (α-, rubi- and coronaviruses). The grouping of similar residues was as follows: I,L,V,M; F,Y,W; K,R; S,T; D,E,N,Q. Residues identical or similar in the sequences of different virus groups are highlighted by boldface.

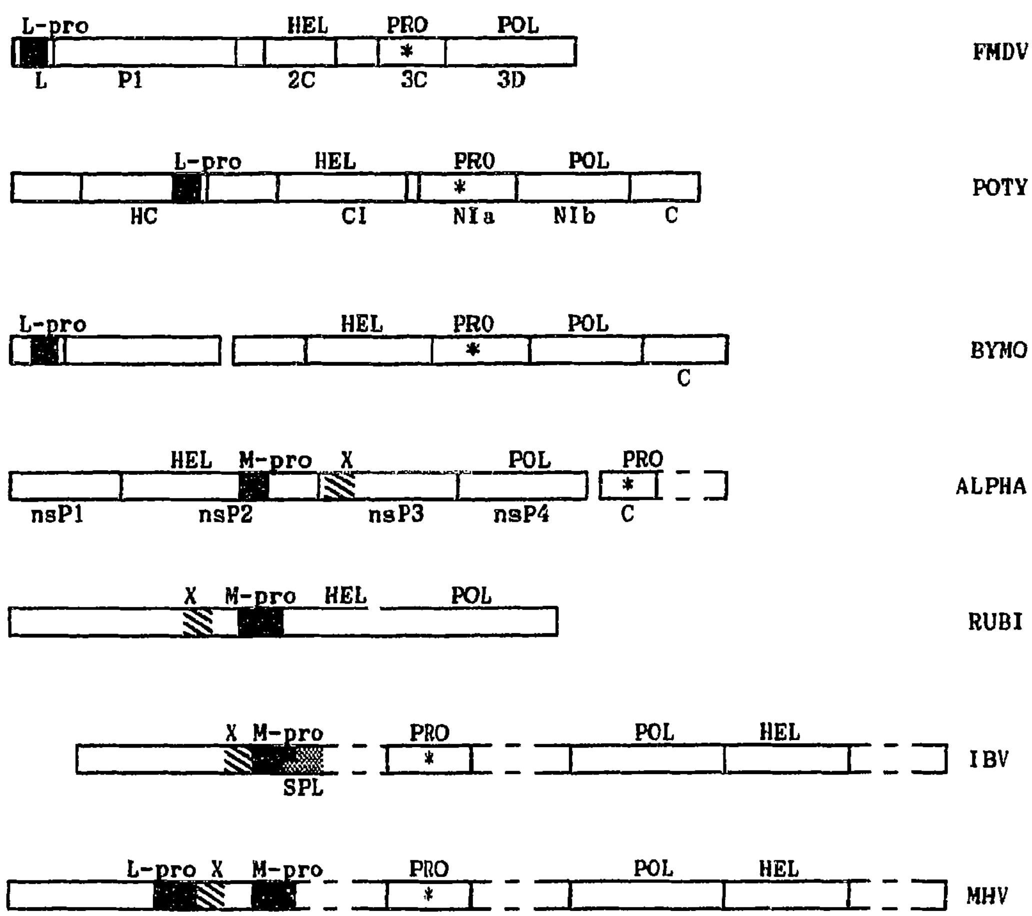

Fig. 3. Location of the papain-like proteases and 'X' domains in viral polyproteins. Only polymerase (POL), helicase (HEL), chymotrypsin-related protease (PRO) and papain-related protease (L-pro, M-pro, SPL) domains are indicated. The designations of specific viral proteins are shown where appropriate. Papain-like proteases and 'X' domains are highlighted by respective hatching. In the IBV polyprotein, M-pro and SPL share a common segment which comprises the C-terminal portion of the first of these putative proteases, and the N-terminal portion of the second one.

and of (putative) viral proteases are even more variable (Fig. 1). The lengths of the spacers separating the two catalytic residues in the (putative) viral proteases varied from quite long (i.e. comparable to the longest found in cell proteases) in the coronavirus enzymes to exceptionally short in the α-, poty- and bymovirus proteases (Fig. 1).

The location of the (putative) papain-like proteases and of 'X' domains in virus polyproteins is highly variable (Fig. 3). Nevertheless, certain regularities could be noticed, and two groups of proteases could be delineated, based on their roles in the processing of the viral polyproteins. The first group includes the proteases of poty-, bymo- and aphthoviruses. These are 'accessory' leader proteases mediating a single cleavage event at their own C termini, while most of the cleavages of the respective polyproteins are effected by chymotrypsin-related proteases [1]. Accordingly, these papain-like protease domains lie outside the arrays of domains directly involved in genome replication and expression, occupying the very N-terminal (FMDV L protein and the bymovirus putative protease) or near-terminal (potyvirus HC protein) positions in the polyproteins (Fig. 3).

The second group encompasses the proteases of α-, and probably of rubiviruses, which appear to be the 'main', and possibly the only enzymes responsible for the processing of non-structural polyproteins. These proteases constitute parts of the arrays of the domains mediating viral RNA replication and expression, which include the RNA polymerase and the (putative) helicase (Fig. 3). It is interesting that the proteases of the second, but not of the first group are associated with the 'X' domain; thus, it is tempting to speculate that this domain might be involved in the regulation of the polyprotein processing.

Acknowledgements: A.E.G. is most grateful to Prof. V.I. Agol for support and encouragement. M.M.C.L. is an Investigator of the Howard Hughes Medical Institute.

REFERENCES

[1] Krausslich, H.-G. and Wimmer, E. (1988) Annu. Rev. Biochem. 57, 701-756.
[2] Bazan, J.F. and Fletterick, R.J. (1988) Proc. Nat. Acad. Sci. USA 85, 7872-7876.
[3] Gorbalenya, A.E., Blinov, V.M., Donchenko, A.P. and Koonin, E.V. (1989) FEBS Lett. 243, 103-114.
[4] Gorbalenya, A.E., Donchenko, A.P., Koonin, E.V. and Blinov, V.M. (1989) Nucleic Acids Res. 17, 3889-3897.
[5] Bazan, J.F. and Fletterick, R.J. (1989) Viroiogy 171, 637-639.
[6] Oh, C.-S. and Carrington, J.C. (1989) Virology 173, 692-699.
[7] Hardy, W.R. and Strauss, J.H. (1989) J. Virol. 63, 4653-4664.
[8] Strauss, E.G., de Groot, R., Shirako, Y., Hardy, W.R. and Strauss, J.H. (1990) VIIIth International Congress of Virology, Berlin (Abstract W2-005).
[9] Baker, S.C., La Monica, N., Shieh, C.-K. and Lai, M.M.C. (1990) in: Pathogenesis and Molecular Biology of Coronaviruses (Cavanagh, D. and Brown, T.D.K., eds.), Plenum, New York, in press.
[10] Lee, H.-J., Shieh, C.-K., Gorbalenya, A.E., Koonin, E.V., La Monica, N., Tuler, J., Bagdzhadzhayan, A. and Lai, M.M.C. (1991) Virology, 190, 567-582.
[11] Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1989) Nucleic Acids Res. 17, 4847-4861.
[12] Dominguez, G., Wang, C.-Y. and Frey, T.K. (1990) Virology 177, 225-238.
[13] Boursnell, 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.
[14] Kashiwazaki, S., Minobe, Y. and Hibino, H. (1991) J. Gen. Virol. 72, 995-999.
[15] Gorbalenya, A.E., Blinov, V.M., Donchenko, A.P. and Koonin, E.V. (1989) J. Mol. Evol. 28, 256-268.
[16] Leontovich, A.M., Brodsky, L.I. and Gorbalenya, A.E. (1990) Biopolimery i Kletka 6, 14-21.
[17] Brodsky, L.I., Drachev, A.L., Tatuzov, R.L. and Chumakov, K.M. (1991) Biopolimery i Kletka 7, 10-14.
[18] Sebti, S.M., Mignano, J.E., Jani, J.P., Srimatkandada, S. and Lazo, J.S. (1989) Biochemistry 28, 6544-6548.
[19] Strebel, K. and Beck, E.J. (1986) J. Virol. 58, 893-899.
[20] Ding, M. and Schlesinger, M.J. (1989) Virology 171, 280-284.