Volume 262, number 1, 145–148 FEBS 08233 March 1990

A new superfamily of putative NTP-binding domains encoded by genomes of small DNA and RNA viruses

Alexander E. Gorbalenya, Eugene V. Koonin* and Yuri I. Wolf°

*Institute of Poliomyelitis and Viral Encephalitides, USSR Academy of Medical Sciences, Moscow Region, *Institute of Microbiology, USSR Academy of Sciences, Moscow,* and °*Institute of Bioorganic Chemistry, USSR Academy of Sciences, Novosibirsk, USSR*

Received 8 January 1990

Statistically significant similarity was revealed between amino acid sequences of NTP-binding pattern-containing domains which are among the most conserved protein segments in dissimilar groups of ss and dsDNA viruses (papova-, parvo-, geminiviruses and P4 bacteriophage), and RNA viruses (picorna-, como- and nepoviruses) with small genomes. Within the aligned domains of 100–120 amino acid residues, three highly conserved sequence segments have been identified, i.e. 'A' and 'B' motifs of the NTP-binding pattern, and a third, C-terminal motif 'C', not described previously. The sequence of the 'B' motif in the proteins of the new superfamily is unusually variable, with substitutions, in some of the members, of the Asp residue conserved in other NTP-binding proteins. The 'C' motif is characterized by an invariant Asn residue preceded by a stretch of hydrophobic residues. As the new superfamily included a well studied DNA and RNA helicase, T antigen of SV40, helicase function could be tentatively assigned also to the other related viral putative NTP-binding proteins. On the other hand, the possibility of different and/or multiple functions for some of these proteins is discussed.

RNA virus; DNA virus; NTP-binding sequence pattern; Helicase; Virus evolution

1. INTRODUCTION

Sequencing of a number of viral genomes allowed delineation of amino acid sequence segments of varying degrees of conservation. In several viral groups, (putative) NTPase domains containing the NTP-binding sequence pattern [1] are among the most conserved sequences. The conservation of this domain was described for early proteins of papovaviruses, NS1 proteins of parvoviruses, and replicative proteins of herpesviruses and positive strand RNA viruses ([2], and references therein). In parvoviruses the putative NTPase domain of approx. 130 amino acid residues is the most, and in fact, the only highly conserved sequence. Significant similarity has been observed between these sequences of parvoviruses and those of (putative) NTPase domains of papovavirus replicative proteins (3).

Putative NTPases of positive strand RNA viruses had been classified into three distinct families [4,5]. It had been shown that two of these families constituted subdivisions of larger superfamilies, each including also (putative) DNA and/or RNA helicases from eubacteria, eukaryotes and large DNA viruses [5–7]. Only for the third family of the NTP-binding pattern-containing proteins of positive strand RNA viruses, which encompassed replicative proteins of animal picornaviruses and plant como- and nepoviruses, relatives with known functions have not yet been identified. Here, we demonstrate significant sequence similarity between the (putative) NTP-binding domains of the latter three groups of RNA viruses, and those of small DNA viruses, i.e. papova-, parvo-, geminiviruses, and P4 bacteriophage. Based on the well-characterized helicase activity of SV40 T antigen and on circumstantional evidence for other viruses, we suggest that conservation of the helicase function might underlie the observed sequence conservation.

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

Abbreviations: PV1, poliovirus type 1; CVB3, Coxsackie virus type B3; HRV2,14,89, human rhinoviruses, respective types; EMCV, encephalomyocarditis virus; TMEV, Theiler murine encephalomyelitis virus; FMDV, foot-and-mouth disease virus type A10; HAV, hepatitis A virus; CPMV, cowpea mosaic virus; TBRV, tomato black ring virus; AAV, adeno-associated virus; MVM, minute virus of mice; BPV, bovine parvovirus; ADV, aleutian disease virus; MDV, mosquito densonucleosis virus; WDV, wheat dwarf virus; MSV, maize streak virus; CSMV, cassava striate mosaic virus; TGMV, tomato golden mosaic virus; CLV, cassava latent virus; BCTV, beet curly top virus; BKV, BK virus; PyV, human polyoma virus; BFDV, budgerigar fledgling disease virus; HPV 1a,6b,18, human papilloma viruses, respective types; BPV1, bovine papilloma virus type 1

2. MATERIALS AND METHODS

2.1. *Amino acid sequences*

Amino acid sequences compared were those of 2C proteins of picornaviruses: PV1, CVB3, HRV2,14,89, EMCV, TMEV, FMDV, and HAV; p58 of CPMV; p72 of TBRV; NS1 proteins of par-

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

voviruses; AAV, B19, MVM, BPV, ADV, and MDV; non-structural proteins of geminiviruses: WDV, MSV, TGMV, GLV, CSMV, and BCTV; T antigens of polyomaviruses: BKV, SV40, PyV, and BFDV; E1 proteins of papillomaviruses: HPV1a, 6b, 18, and BPV1; bacteriophage P4 α protein. The sequences were from the NBRF amino acid sequence Database, or from current literature ([2], and reference therein).

2.2. *Sequence analysis*

Amino acid sequences were compared by the program OPTAL as previously described [4] using the amino acid residue comparison matrix MDM78. The significance of the obtained alignment was assessed by a Monte Carlo procedure and expressed in standard deviation (SD) units. Amino acid sequences were aligned progressively in the order of decreasing similarity. The program OPTAL was written in FORTRAN77. Secondary structure prediction was by the modified Garnier method [8] and by the Chou-Fasman method [9] implemented as Pascal programs. The programs were run on IBM PC AT.

3. RESULTS AND DISCUSSION

3.1. *NTP-binding pattern-containing domains of several groups of small viruses are related*

Several observations suggested that the (putative) NTPase domains of diverse groups of small RNA and DNA viruses shared important common features and that a detailed sequence comparison might be worthwhile. Inspection of the published alignments of the NTP-binding pattern-containing domains of picorna/como/nepo, parvo- and papovaviruses revealed that: (i) in each of these groups, only relatively small domains of 120–150 amino acid residues were well conserved; (ii) unlike many other NTP-binding pattern-containing proteins, the 'A' and 'B' motifs were separated by spacers of rather uniform, and relatively short, length; (iii) in addition to the 'A' and 'B' motifs, all the three groups of proteins had a third conserved segment (hereafter 'C' motif) which resides between the 'B' motif and the C-terminus of the (putative) NTPase domain and consists of an Asn preceded by a run of hydrophobic residues. These notions held also for the NTP-binding pattern-containing proteins of geminiviruses where the presence of the pattern had not been noted previously. Here again, the sequence of this protein was best conserved among all viral gene products.

Comparison of the four alignments indeed showed a statistically convincing similarity of over 6 SD for each clustering step, including the last one when the RNA virus sequences were added to the alignment of DNA virus proteins. The close affinities between the (putative) NTPase domains of papova- vs parvo-, and picorna vs como-/nepoviruses were confirmed. Surprisingly, when compared separately, the proteins of parvo- and papillomaviruses displayed an even greater similarity than that between the two subdivisions of papovaviruses. Inspection of the alignment of four groups of viral proteins (fig.1) showed that 'A', 'B' and 'C' motifs were the only strongly conserved segments. Only four invariant residues were found, two Gly and Lys in the 'A' motif, and Asn in the 'C' motif. More conservation could be observed when accounted for homologous replacements ('consensus' in Fig.1). Secondary structure prediction showed, notwithstanding some exceptions due probably to the known imprecision of the predictive algorithms, that 'A', 'B' and 'C' motifs consist of hydrophobic β-strands flanking the (nearly) invariant residues, in accord with the published structural model of the polyomavirus T antigen [10].

Other proteins containing the NTP-binding pattern and a properly located Asn residue flanked by a hydrophobic stretch were probed for possible relatedness to the newly characterized group. A strong similarity was revealed to the putative NTPase domain of bacteriophage P4 primase (fig.1), justifying its inclusion. Moreover, an even more striking relationship (>7 SD) was observed between P4 primase and the putative NTPases of parvoviruses. The putative NTPase domain of *E. coli* La protease displayed a more modest but significant similarity (approx. 6 SD). However, the La sequence differed from the latter by a longer distance between the 'A' and 'B' motifs. As will be shown elsewhere (A.E.G. and E.V.K., submitted), the NTP-binding pattern-containing domain of La protease appears to belong to a different family of NTPases which might be related to the group of virus proteins described in this paper.

The small number of conserved segments which all were tightly packed within a relatively small domain contrasted the situation in several other (super)families of NTP-binding pattern-containing proteins characterized by much larger conserved domains [2,5–7]. Another distinctive feature of the new group was the unprecedented variability of the 'B' motif sequence (fig.1), whereas the 'A' motif was highly conserved.

Together, these observations suggested the NTP-binding pattern-containing domains of papova-, parvo-, gemini-, picorna-, como-, nepoviruses, and P4 bacteriophage to be regarded as a superfamily in the sense that their sequences are more closely related to each other than to those of other proteins containing the same pattern. The organization of conserved sequence segments in the proteins of this superfamily resembled somewhat that in another large group that brought together numerous bacterial, phage and eukaryotic NTP-binding proteins involved both in genome replication, recombination and repair, and in active transport ('UvrA-related' superfamily; [2,11], and A.E.G. and E.V.K., submitted).

3.2. *Functional and evolutionary implications*

The sequence, and presumably structural, similarity between the NTP-binding pattern-containing domains of the new superfamily suggested they might be similar also functionally. A plausible possibility is that their common activity might be that of a DNA and/or RNA helicase. SV40 T antigen acts as a DNA or RNA

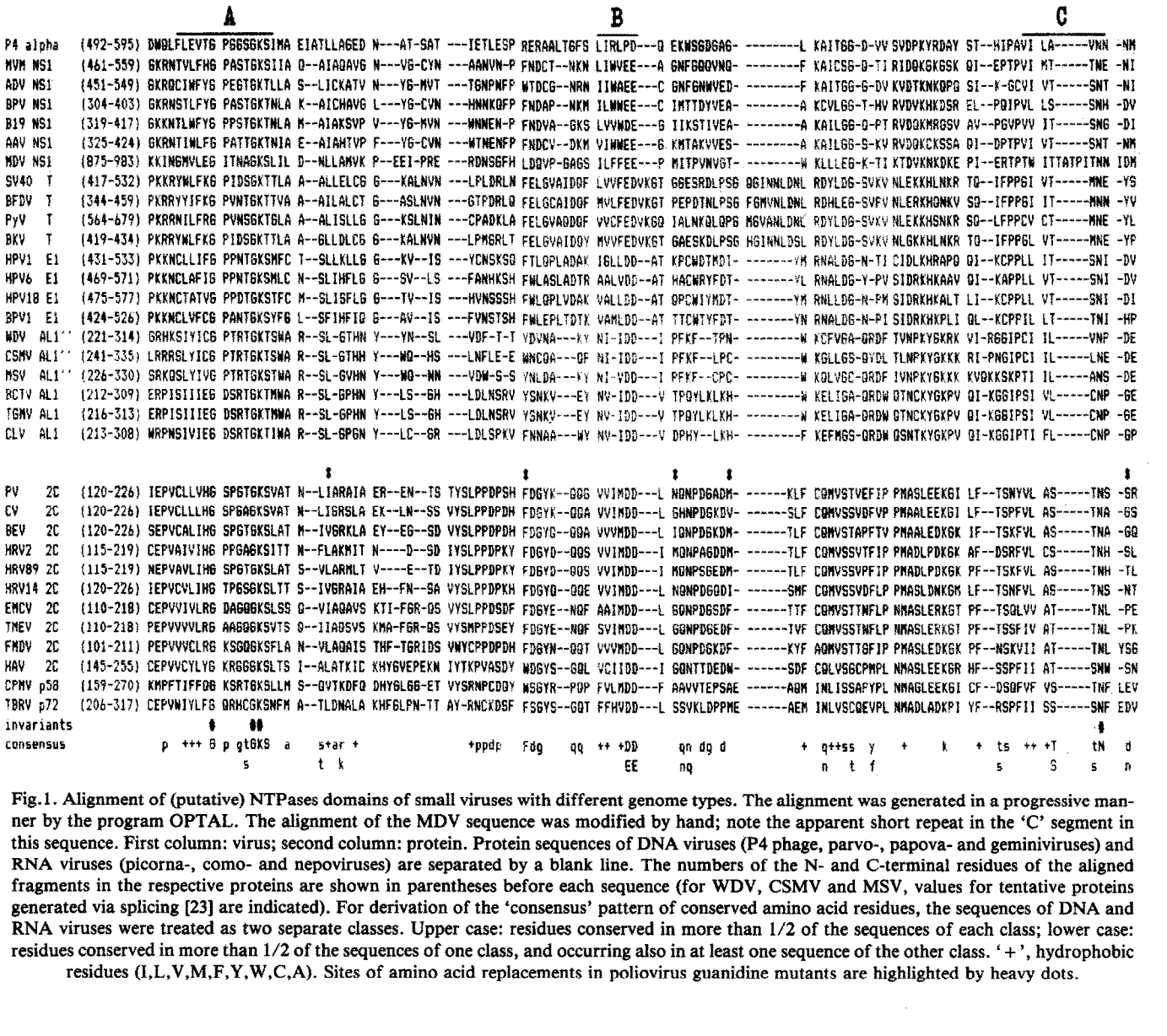

Fig.1. Alignment of (putative) NTPases domains of small viruses with different genome types. The alignment was generated in a progressive manner by the program OPTAL. The alignment of the MDV sequence was modified by hand; note the apparent short repeat in the 'C' segment in this sequence. First column: virus; second column: protein. Protein sequences of DNA viruses (P4 phage, parvo-, papova- and geminiviruses) and RNA viruses (picorna-, como- and nepoviruses) are separated by a blank line. The numbers of the N- and C-terminal residues of the aligned fragments in the respective proteins are shown in parentheses before each sequence (for WDV, CSMV and MSV, values for tentative proteins generated via splicing [23] are indicated). For derivation of the 'consensus' pattern of conserved amino acid residues, the sequences of DNA and RNA viruses were treated as two separate classes. Upper case: residues conserved in more than 1/2 of the sequences of each class; lower case: residues conserved in more than 1/2 of the sequences of one class, and occurring also in at least one sequence of the other class. '+', hydrophobic residues (I,L,V,M,F,Y,W,C,A). Sites of amino acid replacements in poliovirus guanidine mutants are highlighted by heavy dots.

helicase, dependent on the NTP substrate, and is involved in viral DNA replication, transcription, and perhaps translation regulation [12]. Moreover, the ATPase and nucleotide binding activities of T antigen had been mapped to the region encompassing the NTP-binding pattern [10]. Helicase activity of the P4 primase had not been studied but generally presence of such activity in a primase is most likely [13]. Functions of the putative NTPase domains of the two other families of small DNA viruses are less fully understood, but it is well established that they are involved in genome replication, in accord with possible helicase activity [14–16]. NTP-binding pattern-containing proteins of picorna- and comoviruses are constituents of viral replication complexes [17,18]. Genetic analysis showed that 2C proteins of poliovirus and foot-and-mouth disease virus are the targets of the action of guanidine, an antiviral agent inhibiting ssRNA synthesis [19]. Strikingly, all poliovirus mutations to guanidine resistance and dependence that have been mapped fell within the putative NTPase domain, mostly in the close proximity of one or another of the three highly conserved segments ([19,20]; fig.1). On the other hand, 2–4 amino acid insertions introduced by site-directed mutagenesis in this domain were all lethal [21]. It had been shown that EMCV replication complexes in vitro readily utilized non-hydrolyzable ATP analogs for dsRNA synthesis, whereas ssRNA yield was drastically reduced, suggesting an NTP-consuming unwinding step [22].

Thus the available experimental data appear to support the helicase hypothesis. The sequence conservation in the new superfamily was, however, restricted to relatively short sequences of the size typical of NTPase domains proper [10]. This suggests that whereas the mechanisms of NTP hydrolysis might be largely com-

mon, mechanisms of DNA (RNA) duplex unwinding probably differ. On the other hand, our observations do not exclude the possibility of more drastic functional differences between the (putative) NTP-binding proteins of the new superfamily. This possibility seems even more likely taking into account, first, the functional diversity of the 'UvrA-related' superfamily, and, second, the relationship between La protease and its homologs, and the group of viral proteins described here (see above).

Finally, the finding that highly conserved domains of small viruses with ssRNA, and linear or circular ssDNA and dsDNA genomes are apparently homologous may indicate that, despite fundamental differences in genome structure, all these viruses may have evolved from a common ancestor.

Acknowledgements: The authors are grateful to Professor V.I. Agol for constant interest, to Drs A.P. Donchenko and D.R. Davydov for help with computer programming, and to Dr B.N. Afanasiev for communicating his sequence data prior to publication.

REFERENCES

[1] Walker, J.E., Saraste, M., Runswick, M.J. and Gay, N.J. (1982) EMBO J. 1, 945–951.
[2] Gorbalenya, A.E. and Koonin, E.V. (1989) Nucleic Acids Res. 17, 8413–8440.
[3] Astell, C.R., Mol, C.D. and Anderson, W.F. (1987) J. Gen. Virol. 68, 885–893.
[4] Gorbalenya, A.E., Blinov, V.M., Donchenko, A.P. and Koonin, E.V. (1989) J. Mol. Evol. 28, 256–268.
[5] Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1989) Nucleic Acids Res. 17, 4713–4730.
[6] Gorbalenya, A.E., Koonin, E.V., Donchenko, A.P. and Blinov, V.M. (1988) FEBS Lett. 235, 16–24.
[7] Hodgman, T.C. (1988) Nature 233, 22–23.
[8] Gibrat, J. F., Garnier, J. and Robson, B. (1987) J. Mol. Biol. 198, 425–443.
[9] Chou, P.Y. and Fasman, G.D. (1978) Adv. Enzymol. 47, 45–148.
[10] Bradley, M.K., Smith, T.F., Lathrop, R.H., Livingston, D.M. and Webster, T.A. (1987) Proc. Natl. Acad. Sci. USA 84, 4026–4030.
[11] Higgins, C.F., Gallagher, M.P., Mimmack, M.L. and Pearse, S.R. (1988) BioEssays 8, 111–116.
[12] Scheffner, M., Knippers, R. and Stahl, H. (1989) Cell 57, 955–963.
[13] Nakai, H. and Richardson, C.C. (1988) J. Biol. Chem. 263, 9818–9830.
[14] Tullis, G.E., Labienic-Pintel, L., Clemens, K.E. and Pintel, D. (1988) J. Virol. 62, 2736–2744.
[15] Elmer, J.S., Brand, L., Sunter, G., Gardiner, W.E., Bisaro, D.M. and Rogers, S.G. (1988) Nucleic Acids Res. 16, 7043–7060.
[16] Lazarowitz, S., Pinder, A.J., Damsteegt, V.D. and Rogers, S.G. (1989) EMBO J. 8, 1023–1032.
[17] Kuhn, R.J. and Wimmer, E. (1987) in: The Molecular Biology of Positive Strand RNA Viruses (Rowlands, D.J., Mahy, B.W.J. and Mayo, M. eds), pp. 17–51, Academic Press, London.
[18] Van Kammen, A. and Eggen, H.I.L. (1986) BioEssays 5, 261–266.
[19] Pincus, S.E., Rohl, H. and Wimmer, E. (1987) Virology 157, 83–88.
[20] Baltera, R.F. and Tershak, D.R. (1989) J. Virol. 63, 4441–4444.
[21] Li, J.-P. and Baltimore, D. (1988) J. Virol. 62, 4016–4021.
[22] Dmitrieva, T.M., Eremeeva, T.P. and Agol, V.I. (1980) FEBS Lett. 115, 19–22.
[23] Schalk, H.-J., Matzeit, V., Schiller, B., Schell, J. and Gronenborn, B. (1989) EMBO J. 8, 359–364.