J Mol Evol (1989) 28:256–268

Journal of
Molecular Evolution
© Springer-Verlag New York Inc. 1989

An NTP-Binding Motif Is the Most Conserved Sequence in a Highly Diverged Monophyletic Group of Proteins Involved in Positive Strand RNA Viral Replication

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

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

Summary. NTP-motif, a consensus sequence previously shown to be characteristic of numerous NTP-utilizing enzymes, was identified in nonstructural proteins of several groups of positive-strand RNA viruses. These groups include picorna-, alpha-, and coronaviruses infecting animals and como-, poty-, tobamo-, tricorna-, hordei-, and furoviruses of plants, totalling 21 viruses. It has been demonstrated that the viral NTP-motif-containing proteins constitute three distinct families, the sequences within each family being similar to each other at a statistically highly significant level. A lower, but still valid similarity has also been revealed between the families. An overall alignment has been generated, which includes several highly conserved sequence stretches. The two most prominent of the latter contain the socalled "A" and "B" sites of the NTP-motif, with four of the five invariant amino acid residues observed within these sequences. These observations, taken together with the results of comparative analysis of the positions occupied by respective proteins (domains) in viral multidomain proteins, suggest that all the NTP-motif-containing proteins of positive-strand RNA viruses are homologous, constituting a highly diverged monophyletic group. In this group the "A" and "B" sites of the NTP-motif are the most conserved sequences and, by inference, should play the principal role in the functioning of the proteins. A hypothesis is proposed that all these proteins possess NTP-binding capacity and possibly NTPase activity, performing some NTP-dependent function in viral RNA replication. The importance of phylogenetic analysis for the assessment of the significance of the occurrence of the NTP-motif (and of sequence motifs of this sort in general) in proteins is emphasized.

Key words: Evolution — Multiple sequence alignment — NTP binding — Phylogenetic analysis — Positive-strand RNA viruses

Introduction

Structural (sequence) motifs thought to be identifiers of certain protein activities are among the main tools used in the functional and evolutionary interpretation of protein sequence data (Doolittle 1986a,b; Hodgman 1986). Because these motifs are short sequence stretches, and usually include amino acid residues frequent in proteins (i.e., Gly, Ala, Ser, and some others), the presence of such a motif in a protein sequence is, as a rule, in itself not statistically significant. Thus, it is important to work out some additional criteria for evaluation of such observations.

One of the most widespread sequence motifs is implicated in an activity crucial to the function of a great variety of proteins, namely purine nucleotide binding followed in most cases by hydrolysis of the β–γ phosphate bond. This motif was first recognized by Walker and coworkers in several ATP- and GTP-utilizing enzymes (Walker et al. 1982; Gay and Walker 1983). It consists of two separate units designated "A" and "B" sites; the "B" site is located in the polypeptide chains C-proximally relative to the "A" site. For the "A" site the following consensus sequence was proposed: GXXXXGK-(T)

Offprint requests to: A.E. Gorbalenya

XXXXXXI/V, and the "B" consensus was R/K-XXXGXXXL***D, where X stands for any amino acid residue, and * for a hydrophobic residue (Walker et al. 1982). Results of subsequent analyses of a variety of NTP-utilizing proteins (reviewed by Halliday 1984; Möller and Amons 1985; Doolittle 1986a) suggest much more liberal consensus formulas, namely G/AXXXXGKT/S for the "A" site, and an Asp residue preceded by five residues, three of which are hydrophobic, for the "B" site. Hereafter we accept these loose definitions for the "A" and "B" consensus sequences; taken together, they are designated "NTP-motif." In fact, in recent studies, protein sequences were searched for the "A" consensus alone as the "B" consensus in its loosest form is obviously too degenerate to be unequivocally recognized, except in a family of diverged proteins (see below).

For adenylate kinase, *Escherichia coli* Tu factor, p21ras oncoprotein, SV40 T antigen, and some other proteins, there is experimental evidence that the NTP-motif, or at least a larger segment of a protein encompassing it, is involved in NTP binding and/or cleavage (Clertant and Seif 1984; Jurnak 1985; La Cour et al. 1985; Fry et al. 1986). More specifically, the "A" site has been implicated directly in the binding of the pyrophosphate moiety of NTP, whereas the Asp residue of the "B" site appears to interact with the magnesium cation complexed with the same phosphate groups (Möller and Amons 1985; Bradley et al. 1987). All these observations make it an attractive idea to search sequences of functionally uncharacterized proteins for the presence of the NTP-motif to the end of prediction of NTPase activity, or at least NTP-binding capacity.

Following this line, we screened the protein sequences of positive-strand RNA viruses (the largest class of viruses, whose single-stranded genomic RNA also serves as the mRNA for the synthesis of viral proteins) and identified the "A" consensus in nonstructural proteins of several viral families (Gorbalenya et al. 1985). In some of these proteins the presence of this consensus has been independently noticed by other workers too (Argos and Leberman 1985; Doolittle 1986a; Dever et al. 1987; Domier et al. 1987). Also, the NTP-binding capacity of one of these proteins, p126 of tobacco mosaic virus, has been demonstrated experimentally quite independently (Evans et al. 1985). We proposed that such a capacity should be characteristic of all the NTP-motif-containing proteins of positive-strand RNA viruses. On the other hand, the validity of such predictions in general has been disputed (Argos and Leberman 1985; Doolittle 1986a). Moreover, Doolittle (1986a) identified the "A" consensus in several proteins reported to be devoid of NTP-binding properties.

In the present study we undertook a more systematic investigation of the primary structures of the proteins of positive-strand RNA viruses containing the "A" consensus of the NTP-motif. We demonstrate that the NTP-motif-containing proteins of positive-strand RNA viruses (including 8 proteins identified in the previous paper and 13 proteins of viruses whose genomes have been sequenced since then) constitute three monophyletic families that can be brought together into a higher rank taxon. In these homologous proteins the "A" and "B" sites of the NTP-motif are the most strictly conserved sequences and, by inference, should be of principal functional importance, presumably constituting parts of NTPase catalytic centers.

Methods

Protein sequences were extracted from the current literature (for references see Table 1). The initial screening of the sequences of positive-strand RNA viral proteins for the presence of the "A" consensus sequence of the NTP-motif was performed by use of the program SRCH designed to screen protein sequences for defined amino acid residue strings (motifs). The selected sequences were further analyzed manually for the presence of candidate "B" sequences C-proximal to the "A" site. Sequences of the viral proteins containing the NTP-motif were compared by use of the programs DIAGON (Staden 1982) and OPTAL (Pozdnyakov and Pankov 1981); the latter program was modified and adopted for multiple sequence alignment as described below. All the programs were written in FORTRAN and run on an ES-1060 computer.

The program OPTAL, based on the original algorithm of Sankoff (1972), performs optimal alignment of pairs of protein sequences or stepwise alignment of multiple sequences. According to the Sankoff algorithm, a series of cumulative similarity matrices for the compared sequences is created. In the present work, for the calculation of the elements of these matrices, weights of amino acid residue pairs were taken from the mutation rate scoring matrix MDM78 (Staden 1982). To accelerate calculations, only those elements of the matrices enclosed within a diagonal window, the width of which was chosen to be equal to ⅕ of the length of the compared sequences, were computed; it has been shown that this window width is sufficient in most cases for generation of the optimal alignment (Pozdnyakov and Pankov 1981). At the first step of the optimal alignment generation, a series of locally optimal alignments with $q = 0, 1, 2, \ldots, q_{max}$ gaps was obtained. In practice, for sequence lengths of up to 250 residues dealt with in this work, q_{max} was chosen to be equal to 15. This exceeds the gap number most frequently observed in related protein sequences (about four gaps per 100 residues; see Doolittle 1981) and should guarantee the generation of the optimal alignment. For selection of the best of the alignments with a given q value and concomitant assessment of its statistical significance, the following Monte Carlo procedure was employed. Twenty-five pairs of "random" sequences were generated by scrambling the real compared sequences, and the above alignment procedure was simulated for each pair. The mean score, Sq^{rand} and the standard deviation, σ_q, were calculated separately for alignments with different gap numbers. For each of the locally optimal real alignments, the deviation of the observed score from the mean value for the randomized sequences was calculated in SD units:

Table 1. NTP-motif-containing protein sequences from positive-strand RNA viruses

	Virus							
No.	Full name	Short name	NTP-motif containing protein	Family (group) of viruses	Protein family	Coordinates and size of the protein (amino acid residue numbers)	Coordinates and size of the evolutionary conserved regions (amino acid residue numbers)	References
1	Cucumber mosaic virus	CMV	RNA 1 pr	Tricornaviridae	1	1–991 991	708–843–976 136 133	Rezaian et al. 1985
2	Brome mosaic virus	BMV	RNA 1 pr	Tricornaviridae		1–961 961	680–816–946 137 130	Ahlquist et al. 1984
3	Alfalfa mosaic virus	AlMV	RNA 1 pr	Tricornaviridae		1–1126 1126	836–972–1098 137 127	Cornelissen et al. 1983
4	Tobacco mosaic virus	TMV	p126	Tobamovirus		1–1116 1116	828–969–1086 142 114	Goelet et al. 1982
5	Sindbis virus	SNBV	nsP2	Alphaviridae		541–1347 807	721–856–969 136 113	Strauss et al. 1984
6	Semliki forest virus	SFV	nsP2	Alphaviridae		534–1332 799	718–850–963 133 113	Takkinen 1986
7	Barley stripe mosaic virus	BSMV	p58 (pr of RNAβ ORF 2)	Hordeivirus		1–528 528	264–385–490 122 105	Gustafson and Armour 1986
8	Beet necrotic yellow vein virus	BNYVV	p43 (pr of RNA 2 ORF 3)	Furoviridae		1–384 384	118–244–354 127 110	Bouzoubaa et al. 1986
9	Beet necrotic yellow vein virus	BNYVV	p237 (RNA 1 pr)	Furoviridae		1–2109?	937–1069–1183 133 114	Bouzoubaa et al. 1987
10	Infectious bronchitis virus	IBV	F2	Coronaviridae		1–2652?	1168–1334–1470 167 136	Boursnell et al. 1987
11	Coxsackie virus type B3	CV	2C	Picornaviridae	2	1101–1429 329	1224–1347 124	Lindberg et al. 1987
12	Poliovirus type 1	PV	2C	Picornaviridae		1128–1456 329	1246–1369 124	Racaniello and Baltimore 1981
13	Human rhinovirus type 2	RV2	2C	Picornaviridae		1087–1409 323	1201–1322 122	Skern et al. 1985
14	Human rhinovirus type 14	RV14	2C	Picornaviridae		1100–1429 330	1219–1342 124	Stanway et al. 1984
15	Encephalomyocarditis virus	EMCV	2C	Picornaviridae		1193–1517 325	1302–1427 126	Palmenberg et al. 1984
16	Theiler murine encephalomyelitis virus	TMEV	2C	Picornaviridae		1192–1517 326	1302–1427 126	Pevear et al. 1987
17	Foot-and-mouth disease virus type A10	FMDV	2C	Picornaviridae		1108–1425 318	1207–1334 128	Carrol et al. 1984
18	Hepatitis A virus	HAV	2C	Picornaviridae		1081–1415 335	1213–1340 128	Najarian et al. 1985
19	Cowpea mosaic virus	CPMV	p58	Comovirus		327–919 593	484–613 130	Lomonosoff and Shanks 1983
20	Tobacco vein mottling virus	TVMV	CI	Potyvirus	3	1113–1747 635	1192–1394 203	Domier et al. 1986
21	Tobacco etch virus	TEV	CI	Potyvirus		1164–1795 632	1242–1144 203	Allison et al. 1986

For picorna- and potyviruses, the numbering of complete polyproteins is indicated; for alphaviruses, the numbering of the nonstructural polyproteins is indicated; for CPMV the numbering of the polyprotein encoded by RNA B is indicated; and for BNYVV RNA 1 the numbering of the entire high-molecular weight product is indicated. For the proteins of the 1st family, the coordinates and lengths of the N-terminal and the C-terminal subdomains of the conserved domain (see text) are indicated separately. In the potyviruses, the entire sequences of CI proteins are conserved; only those segments that have counterparts in the proteins of other families are indicated

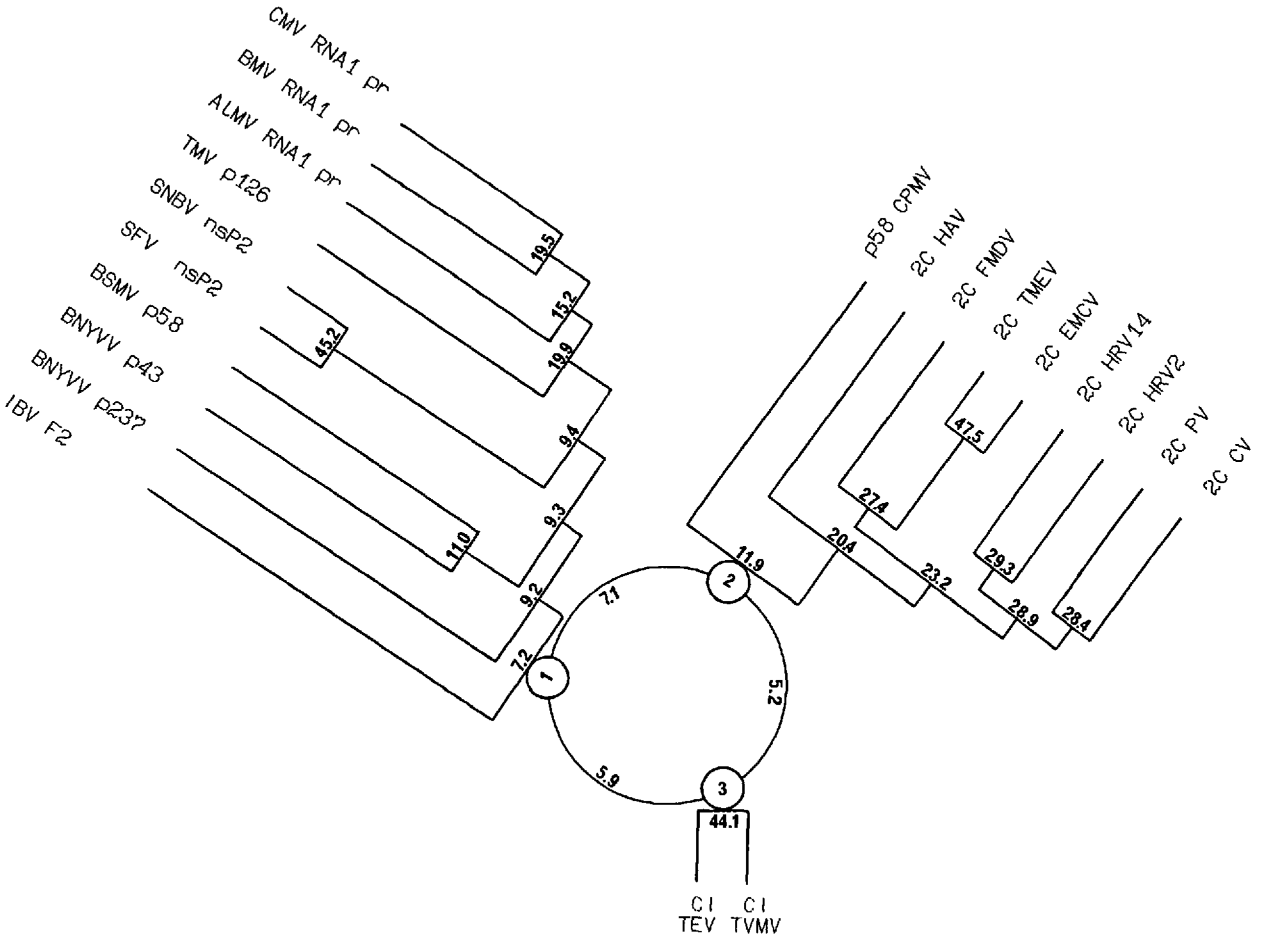

Fig. 1. Sequence similarity between the evolutionarily conserved segments of the NTP-motif-containing proteins of positive-strand RNA viruses. The three dendrograms depict the order of alignment of sequences within each family of viral proteins. The lengths of the branches are in approximate inverse proportion to the degree of sequence similarity observed at each step of the alignment (with some exceptions). In the nodes of the dendrograms D values for each step expressed in SD units (see Methods) are shown. The values obtained upon alignment of prealigned sets of sequences of the three families with each other are shown in the center of the scheme. The dendrograms were designed to visualize the procedure of multiple sequence alignment in the order of decreasing similarity between proteins; they cannot be automatically regarded as evolutionary trees. For abbreviations of viruses see Table 1. The values refer to the conserved domains, as indicated in Table 1, and, for the proteins of the 1st family, to the N-terminal subdomains.

$$Dq = Sq^{real} - Sq^{rand}/\sigma_q.$$

The alignment with the maximal D value was considered optimal for the selected window width.

For multiple sequence alignment, a generalization of this procedure was employed. To align two sets of m and n prealigned sequences, cumulative similarity matrices were created as before, but for the calculation of their elements, values $W_{ij} = \Sigma w_{ij}$, i.e., the combined weights of all possible pairs of residues (m·n total) in the ith position of the set n and the jth position of the set m are used instead of the weights of individual pairs of residues. A weight of 10 was ascribed to a pair of two gaps, and a weight of zero to a pair of gaps with any residue. The procedures of alignment generation and the choice of the optimal alignment were as described, but, for the generation of "random" sequence sets, "columns" of residues occupying each position in the real sets of aligned sequences were jumbled.

Results

Three Families of Viral NTP-Motif-Containing Proteins

Preliminary comparative analysis of the amino acid sequences of the NTP-motif-containing proteins of positive-strand RNA viruses by use of the programs DIAGON and OPTAL (see Methods) revealed three distinct families and some additional proteins in whose close relatives the motif was not conserved. Within each family, all the proteins contained stretches at least 120 residues long that were similar to each other at a statistically highly significant level. For most pairs, the observed alignment scores ex-

←

in this table. Question marks indicate that the real size of the respective proteins is not known; the large proteins presented in the table may in fact be processed. In those cases where sequences of several serotypes (strains) of a single virus species were available (specifically, for several picornaviruses and TMV), only one sequence was included. An exception is rhinovirus serotypes 14 and 2 with sequences that are substantially different. pr = product

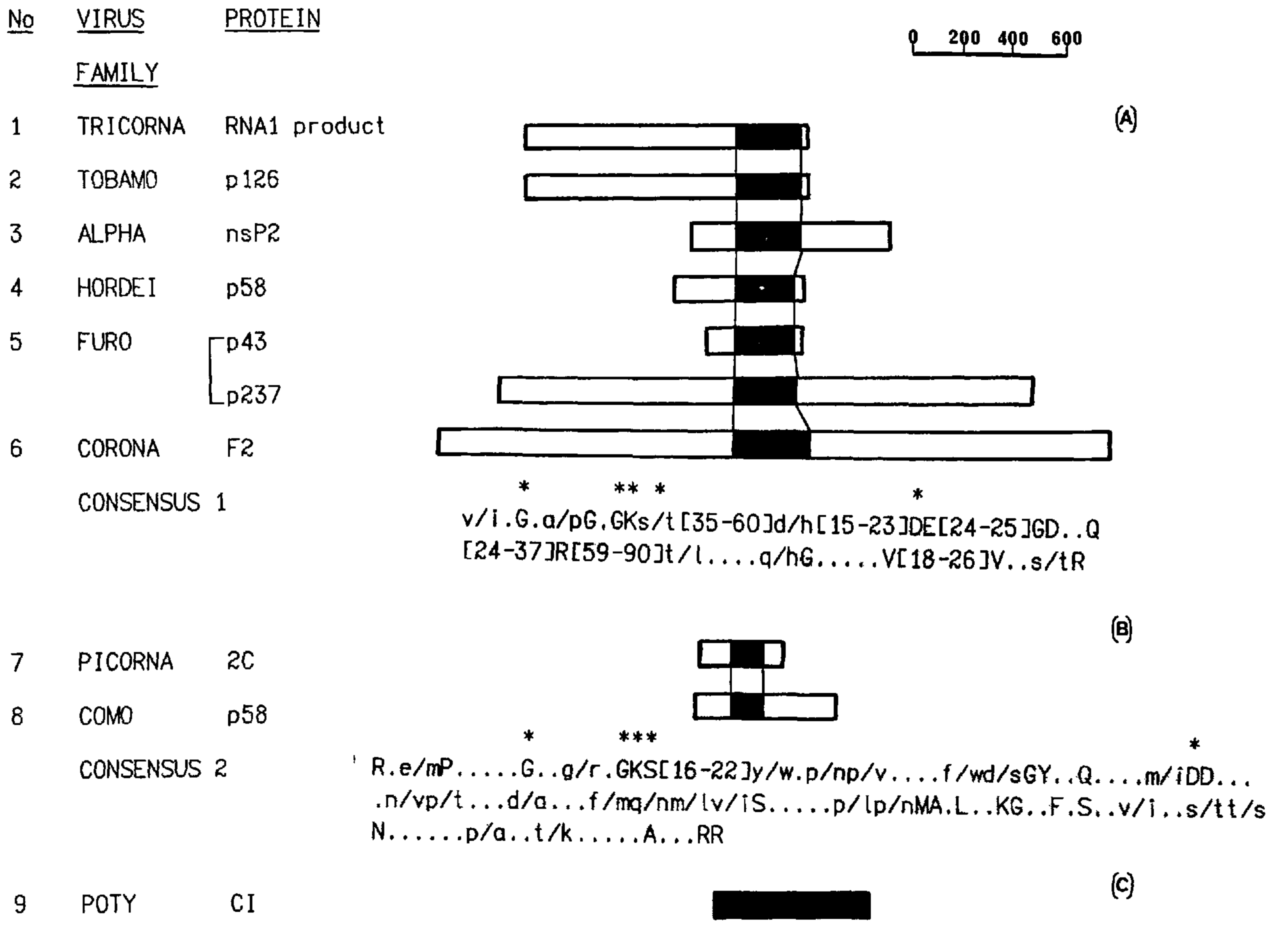
No
VIRUS FAMILY
PROTEIN
0 200 400 600
(A)
1 TRICORNA RNA1 product
2 TOBAMO p126
3 ALPHA nsP2
4 HORDEI p58
5 FURO p43
p237
6 CORONA F2
CONSENSUS 1
* ** * *
v/i.G.a/pG.GKs/t[35-60]d/h[15-23]DE[24-25]GD..Q
[24-37]R[59-90]t/l....q/hG.....V[18-26]V..s/tR
(B)
7 PICORNA 2C
8 COMO p58
CONSENSUS 2
* *** *
R.e/mP.....G..g/r.GKS[16-22]y/w.p/np/v....f/wd/sGY..Q....m/iDD...
.n/vp/t...d/a...f/mq/nm/lv/iS.....p/lp/nMA.L..KG..F.S..v/i..s/tt/s
N......p/a..t/k.....A...RR
(C)
9 POTY CI

Fig. 2. Evolutionarily conserved segments of the NTP-motif-containing proteins (domains) of positive-strand RNA viruses. The NTP-motif-containing proteins of the 1st (A), 2nd (B), and 3rd (C) families are schematically represented by rectangles drawn to scale shown in the upper right part of the figure. The NTP-motif-containing segments displaying statistically significant similarity within each family (see Table 1) are shown in black; they were aligned by the "A" sites of the NTP-motif (see text). The tricorna- and tobamovirus proteins are multidomain proteins, with the N-terminal domains similar to each other and to alphavirus nsP1 protein (not indicated; Ahlquist et al. 1985). In the bottom of A and B the respective patterns of evolutionarily conserved amino acid residues are shown (designated "consensus 1" and "consensus 2"). Invariant residues are capitalized. Dots stand for variable residues (or gaps) within conserved residue clusters; the lengths of variable regions between these clusters are indicated by bracketed numbers. Asterisks denote the amino acid residues constituting the "A" site and the proposed Mg^{2+}-binding D residue of the "B" site.

ceeded the mean scores for randomized sequences by at least 5 SD (see Methods). Such a level of sequence similarity between proteins is usually regarded as serious evidence for their monophyletic origin (Doolittle 1981, 1986a,b; Dayhoff et al. 1983). To obtain optimal group alignments for each family, the sequences were aligned in order of decreasing similarity (Fig. 1). As is evident from the figure, the significance of the multiple sequence alignments was quite high for each of the families. The families were numbered 1st, 2nd, and 3rd in order of decreasing sequence divergence between the presently recognized members. For part of the proteins constituting the 1st and the 2nd families, analogous sequence comparisons (but with no reference to the NTP-motif) were performed previously by other workers and meaningful similarities have also been observed (Argos et al. 1984; Cornelissen and Bol 1984; Franssen et al. 1984; Ahlquist et al. 1985; Bouzoubaa et al. 1987; reviewed by Goldbach 1986). Very recently, Domier et al. (1987) compared the sequences of the proteins of the 2nd and the 3rd families and noticed the presence of the "A" consensus of the NTP-motif.

The 1st family includes the NTP-motif-containing proteins (domains) of alpha-, tobamo-, tricorna-, furo-, and coronaviruses as well as the putative product of the hordeivirus RNA β open reading frame (ORF) 2, totalling 10 proteins (Table 1). These proteins vary greatly in their size (Fig. 2A), genomic positions of the respective coding sequences, and modes of expression. For the proteins of this family, a statistically significant similarity was observed within a fragment of about 250 amino acid residues (Fig. 2A). This fragment contains 21 highly conserved residues (of which 14 are invariant) divided between seven clusters of unequal size (or individual residues). The first and third conserved clusters encompass the "A" and "B" sites of the NTP-motif, respectively. The N-proximal five clusters of conserved amino acid residues in these proteins are separated from the sixth and seventh clusters by a variable region of about 60–90 residues. In fact, the conserved domain appears to be further divided into two subdomains, the N-terminal one containing the NTP-motif, and the C-terminal one of totally unknown function. Two specific points are worth noting. First and most remarkably, two segments of the furovirus genome encode two NTP-motif-containing proteins only distantly related (as compared to other members of the family) to each other; this is demonstrated by direct pairwise comparison of their sequences (data not shown). Second, the inclusion of the coronavirus NTP-motif-containing domain within the 1st family is tentative, as the level of its sequence similarity to the other proteins of this family is not much higher than that between different families (see below). Also, the distance between the "A" and "B" sites of the NTP-motif is much longer in the coronavirus protein than those in the other proteins of this family. Nevertheless, all the amino acid residues invariant in the latter are conserved in the coronavirus protein also (see below), justifying its inclusion in this family.

The 2nd family of NTP-motif-containing proteins includes picornaviral proteins 2C and comoviral protein p58, totalling nine proteins (Table 1). These proteins are much more uniform in their size (Fig. 2B), genomic positions of the respective genes, and mode of expression than those of the 1st family. The region of the most prominent similarity spans the central domain of about 130 amino acid residues; this domain contains 45 conserved residues (23 invariant), more or less evenly distributed (Fig. 2B, consensus 2). The "A" and "B" sites of the NTP-motif are located near the N-terminus and in the middle of the conserved domain, respectively.

The 3rd family includes CI proteins of two potyviruses (Table 1 and Fig. 2C). These proteins are very similar to each other, having more than 50% identical amino acid residues. Thus, derivation of a consensus, like those derived for the other two families, made little sense. The "A" and "B" sites of the NTP-motif are located in the N-terminal parts of CI proteins.

Comparison of Protein Sequences between the Three Families

Comparison of the prealigned sequences of the three families of NTP-motif-containing proteins by the multiple alignment version of OPTAL yielded highly significant alignment scores for all three possible pairs (Fig. 1). However, the final alignment of the sequences of the three families generated by this program (not shown) was not quite satisfactory because the "B" sites of the NTP-motif, as well as some other clusters of residues that seemed good candidates for the conserved regions, did not coincide (although it must be pointed out that the "A" sites did match). Presumably, this might be due to different lengths of spacers separating these regions in the proteins of the three families. Thus, an overall alignment has been generated by manual fitting of the computer alignments of the three sets of sequences so as to maximize residue coincidence conserved within individual families (Fig. 3). In this alignment five amino acid residues are strictly invariant, four additional residues are common to the 2nd and 3rd families, and three residues are conserved in the 1st and 3rd families. In addition, several positions in all, or nearly all, the sequences are occupied by functionally related residues (Fig. 3, consensus). All in all, a certain degree of conservation was observed in about 40% of the positions of the alignment (highlighted in Fig. 3 and further characterized in the legend to this figure).

Strikingly, four of the five invariant residues are located within the "A" and "B" sites of the NTP-motif (Fig. 3). These sites and short sequence stretches surrounding them also contain a considerable number of additional coincidences and similar sequence replacements between proteins of different families. Thus, the "A" and "B" consensus sequences and short adjacent segments are the most similar portions of the NTP-motif-containing proteins of positive-strand RNA viruses.

Of additional interest is a comparison of the positions of the NTP-motif-containing proteins (domains) in viral multidomain proteins; this approach is illustrated in Fig. 4. Only two stretches of similar amino acid sequences are common to all viruses analyzed in this study: (1) the conserved region of the RNA polymerase (Kamer and Argos 1984; Morozov and Rupasov 1985; Koonin et al. 1987, 1988), and (2) the NTP-motif-containing domain (this paper). Viruses, with proteins that constitute the 2nd and 3rd families characterized above, possess an additional protein sequence of significant similarity, i.e., the proteases of picorna-, como-, and potyviruses (Argos et al. 1984; Franssen et al. 1984; Carrington and Dougherty 1987; Domier et al. 1987). In all viruses with nonsegmented genomes (with the probable exception of coronaviruses), in CPMV B polyprotein, and in the furovirus RNA 1 product (p237), the proteins (domains) containing similar sequence stretches are positioned in the same order within multidomain proteins, namely N-NTP-motif-containing domain–(protease)–polymerase-C (Fig. 4). In coronaviruses the polymerase has not yet been identified. However, the results of our preliminary analysis indicate that the polymerase do-

```
                        10        20        30        40

 1 CMV(1)    : ISQVDGVAGCGKTTAIKSMFNPST---DIIVTANKKSAQDV-RYAL-
 2 BMV(1)    : ISMVDGVAGCGKTTAIKDAFRMGE---DLIVTANRKSAEDV-RMAL-
 3 AlMV(1)   : VTIVDGVAGCGKTTNIKQIARSSGRDVDLILTSNRSSADEL-KETI-
 4 TMV       : VVLVDGVPGCGKTKEILSRVNFDE---DLILVPGKQAAEMI-RRRA-
 5 SNBV      : TIGVIGTPGSGKSAIIKSTVTAR----DLVTSGKKENCREI-EADV-
 6 SFV       : VVGVFGVPGSGKSAIIKSLVTKH----DLVTSGKKENCQEI-VNDV-
 7 BSMV(β)   : TGIISGVPGSGKSTIVRTLLK------GEFPAVCALANPAL-MNDY-
 8 BNYVV(2)  : VGIVLGAPGVGKSTSIKNLLD--KFGAKHKMVLCLPFSQLL-EGVF-
 9 BNYVV(1)  : LEYVKGGPGTGKSFLIRSLADPIR---DLVVAPFIKLRSDY-QNQR-
10 IBV       : RTTVQGPPGSGKSHFAIGLAVYFSSARVVFTACSHAAVDALCEKAFK

11 CV B3     : CLLLHGSPGAGKS--VATNLI------------GRSLAEKL-N-S--
12 PV1       : CLLVHGSPGTGKS--VATNLI------------ARAIAERE-N-T--
13 HRV2      : AIVIHGPPGAGKS--ITTNFL------------AKMITN---D-S--
14 HRV14     : CVLIHGTPGSGKS--LTTSIV------------GRAIAEHF-N-S--
15 EMCV      : VIVLRGDAGQGKS--LSSQVI------------AQAVSKTI-F-G--
16 TMEV      : VVVLRGAAGQGKS--VTSQII------------AQSVSKMA-F-G--
17 FMDV A10  : VVCLRGKSGQGKS--FLANVL------------AQAISTHF-T-G--
18 HAV       : VCYLYGKRGGGKS--LTSIAL------------ATKICKHY-GVE--
19 CPMV(B)   : TIFFQGKSRTGKS--LLMSQV------------TKDFQDHY-GLG--

20 TVMV      : DIILMGAVGSGKSTGLPTNLCKFG--GVLLLEPTRPLAENV-TKQMR
21 TEV       : DFLVRGAVGSGKSTGLPYHLSKRG--RVLMLEPTRPLTDNM-HKQLR
cons            * G  g GKs  *   *
                            t

       50        60        70        80        90        100

 1 : ---FKST-DSKEACAFV------RTADS----ILLN-DCP-TVS--RVLVDEVVL
 2 : ---FPDTYNSKVALDVV------RTADS----AIMH-GVP-SCH--RLLVDEAGL
 3 : ------DCSPLTKLHYI------RTCDS----YLMS-ASAVKAQ--RLIFDECFL
 4 : ----NSSGIIVATKDNV------KTVDS----FMMNFGKSTRCQFKRLFIDEGLM
 5 : ------L-RLRGMQITS------KTVDS----VMLN-GCHKAVE--VLYVDEAFA
 6 : ------K-KHRGKGTSR------ENSDS----ILLN-GCRRAVD--ILYVDEAFA
 7 : -----------SGIEGV------YGLDD----LLLS-AVP-ITS-DLLIIDEYTL
 8 : -----------AGRLDT------FLVDD----LFCR-SVE-YGKYNTMLVDEVTR
 9 : ----------VGDELLS------WDFHT----PHKALDVT-GKQ--IIFVDEFTA
10 : FLKVDDCTRIVPQRTTVDCFSKFKANDTGKKYIFSTINALPEVSCDILLVDEVSM

11 : ---------------SVYSLPPDPDHFDGY------------KQQAVVIMDDLCH
12 : ---------------STYSLPPDPSHFDGY------------KQQGVVIMDDLNQ
13 : ---------------DIYSLPPDPKYFDGY------------DQQSVVIMDDIMQ
14 : ---------------AVYSLPPDPKHFDGY------------QQQEVVIMDDLNQ
15 : ------------R-QSVYSLPPDSDFFDGY------------ENQFAAIMDDLGQ
16 : ------------R-QSVYSMPPDSEYFDGY------------ENQFSVIMDDLGQ
17 : ------------RIDSVWYCPPDPDHFDGY------------NQQTVVVMDDLGQ
18 : ------------PEKNIYTKPVASDYWDGY------------SGQLVCIIDDIGQ
19 : ------------G-ETVYSRNPCDQYWSGY------------RRQPFVLMDDFAA

20 : GSPFFASPTLRMRNLSTFGSSPITVMTTGFALHFFANNVKEFDRYQFIIFDEFHV
21 : SEPFNCFPTLRMRGKSTFGSSPITVMTSGFALHHFARNIAEVKTYDFVIIDECHV
cons                             g                  ***De
                                                       d

            110       120       130       140       150

 1 : -LHFGQLCAVMS----------------------------------KLHAVRALC
 2 : -LHYGQLLVVAA----------------------------------LSKCSQVLA
 3 : -QHAGLVYAAAT----------------------------------LAGCSEVIG
 4 : -LHTGCVNFLVA----------------------------------MSLCEIAYV
 5 : -CHAGALLALIA----------------------------------IVRPRKKVV
 6 : -CHSGTLLALIA----------------------------------LVKPRSKVV
 7 : -AESAEILLLQR----------------------------------RLRASMVLL
 8 : -VHMCEILVLAG----------------------------------HLGVKNVIC
 9 : -YDW-RLLAVLA----------------------------------YRNHAHTIY
10 : -LTNYELSFING----------------------------------KINYQYVVY

11 : -NPDGKDVSLFC----------------------------------QMVSSVDFV
12 : -NPDGADMKLFC----------------------------------QMVSTVEFI
13 : -NPAGDDMTLFC----------------------------------QMVSSVTFI
14 : -NPDGQDISMFC----------------------------------QMVSSVDFL
15 : -NPDGSDFTTFC----------------------------------QMVSTTNFL
16 : -NPDGEDFTVFC----------------------------------QMVSSTNFL
17 : -NPDGKDFKYFA----------------------------------QMVSTTGFI
18 : -NTTDEDWSDFC----------------------------------QLVSGCPMR
19 : VVTEPSAEAQMI----------------------------------NLISSAPYP

20 : LDSNAIAFRNLCHEYSYNGKIIKVSATPPGRECDLTTQYPVELLIEEQLSLRDFV
21 : NDASAIAFRNLLFEHEFEGKVLKVSATPPGREVEFTTQFPVKLKIEEALSFQEFV
cons                                                    s
```

Fig. 3. Continued on next page.

```
        160       170       180       190       200
  1 : F--GDSEQIAFSS-RDASFDM--RFSK----LIPDETSDADTT---FRS
  2 : F--GDTEQISFKS-RDAGFKL--LHGN----LQYDRRDVVHKT---YRC
  3 : F--GDTEQIPFVS-RNPSFVF--RHHK----LT-GKVERKLIT---WRS
  4 : Y--GDTQQIPYIN-RVSGFPYPAHFAK----LEVDEVETRRTT---LRC
  5 : L-CGDPMQCGFFN-MMQ-LKV--HFNH-PEKDICTKTFYKYIS---RRC
  6 : L-CGDPKQCGFFN-MMQ-LKV--NFNH----NICTEVCHKSIS---RRC
  7 : V--GDVAQ-GKAT-TASSIEY----------LTLPVIYRSETT---YRL
  8 : F--GDPAQ-GLNY-KAGSAVN----------YNFPIIAECYAS---RRF
  9 : L-VGDEQQTGIQEGRGEGISI-LNKID----LSKVSTHVPIMN---FRN
 10 : V--GDPAQLPAPRTLLNGSLSPKDYNV-VTNLMVCVKPDIFLAK-CYRC

 11 : PPMAALEEKGILFTSP--FVLAST-NA-GSINA-PTVSDSRAL--ARRF
 12 : PPMASLEEKGILFTSN--YVLAST-NS-SRISP-PTVAHSDAL--ARRF
 13 : PPMADLPDKGKAFDSR--FVLCST-NH-SLLTP-PTITSLPAM--NRRF
 14 : PPMASLDNKGMLFTSN--FVLAST-NS-NTLSP-PTILNPEAL--VRRF
 15 : PNMASLERKGTPFTSQ--LVVATT-NL-PEFRP-VTIAHYPAV--ERRI
 16 : PNMAHLERKGTPFTSS--FIVATT-NL-PKFRP-VTVAHYPAV--DRRI
 17 : PPMASLEDKGKPFNSK--VIIATT-NLYSGFTP-RTMVCPDAL--NRRF
 18 : LNMASLEEKGRHFSSP--FIIATS-NW-SNPSP-KTVYVKEAI--DRRL
 19 : LNMAGLEEKGICFDSQ--FVFVST-NF-LEVSPEAKVRDDEAFK-NRRH

 20 : DAQGTDAHADVVKKGDNILVYYVASYNEVDQLSKMLNERGFLVTKVDGRT
 21 : SLQGTGANADVISCGDNILVYYVASYNDVDSLGKLLVQKGYKVSKIDGRT
cons     g                       n                     R
         a
```

Fig. 3. An overall alignment of the evolutionarily conserved segments of the viral NTP-motif-containing proteins of the three families. Only partial sequences of the conserved regions (Table 1) were aligned; they encompass the N-terminal subdomains of the proteins of the 1st family, the sequences of the 2nd family without the five N-terminal amino acid residues, and complete sequences of the 3rd family. The residue numbers shown above the alignment are arbitrary; the numbering begins from the first residues of the aligned stretches and includes gaps. The sets of sequences of the three families aligned by the program OPTAL are separated by blank lines. Dots denote conservative positions. These are defined here as positions occupied by similar amino acid residues in at least 50% of the sequences of each of any two of the three families. The upper, middle, and lower rows of dots indicate the conservative positions of the 1st, 2nd, and 3rd families, respectively. Thus, if a given position in the alignment contains dots, say, in the upper and lower rows, this indicates the conservation of residues (in the above sense) between the 1st and 3rd families, and so on. Similar residues are defined as those belonging to one of the following groups: A, V, I, L, M, and F (hydrophobic); F, Y, and W (aromatic); G and A (small); S and T (hydroxy-); K, R, and H (basic); D, E, N, and Q (acidic and their derivatives); C and P have no similar residues. The pattern of highly conserved residues is shown under the aligned sequences, designated "cons" for consensus. Uppercase letters correspond to invariant residues, and lowercase letters to those conserved in two out of three families; in the latter case, where a similar residue was conserved in the 3rd family, it was also indicated. * = a hydrophobic residue. The "A" (positions 6–13 in the alignment) and "B" (positions 93–98) sites of the NTP-motif are denoted by horizontal bars above and below the alignment. For viruses with segmented genomes, the specific designations of the RNA segments encoding the NTP-motif-containing proteins are given in parentheses.

main also resides in F2, but its position relative to the NTP-motif-containing one is reversed as compared to the "canonical" array described above (unpublished observations). Anyway, this single exception certainly does not invalidate the general trend for the specific positioning of these domains in viral multidomain proteins.

Comparative analysis of the amino acid sequences of all positive-strand RNA virus RNA polymerases provides a strong case for their monophyletic origin (Koonin et al. 1987, 1988). We believe that the sequence similarity between the NTP-motif-containing proteins, together with their similar localization in viral multidomain proteins, indicate that they also constitute a monophyletic group.

Search for Other NTP-Motif-Containing and Related Viral Proteins

In the course of the present study we screened all the available protein sequences of positive-stand RNA viruses for the presence of the NTP-motif. Also, some additional searches have been made: (1) domains occupying positions similar to those of the NTP-motif-containing ones in viral multidomain proteins were searched for the possible presence of degenerate forms of the motif; and (2) partially sequenced proteins were tested for similarity to the NTP-motif-containing proteins.

The "A" consensus sequence has been found in the C-terminal part of AlMV RNA polymerase, in the capsid protein of yellow fever virus (a flavivirus), in NS1 proteins of four flaviviruses, and in the F1 polyprotein of IBV; also, a second "A" sequence (besides the one included in our alignment) is present in the furovirus p237. In the first three instances the consensus sequence was not conserved in the relatives of the respective proteins, suggesting that its occurrence was most likely fortuitous. In the last two cases, the absence of other coronavirus and furovirus protein sequences precluded this type of analysis, leaving the significance of these observa-

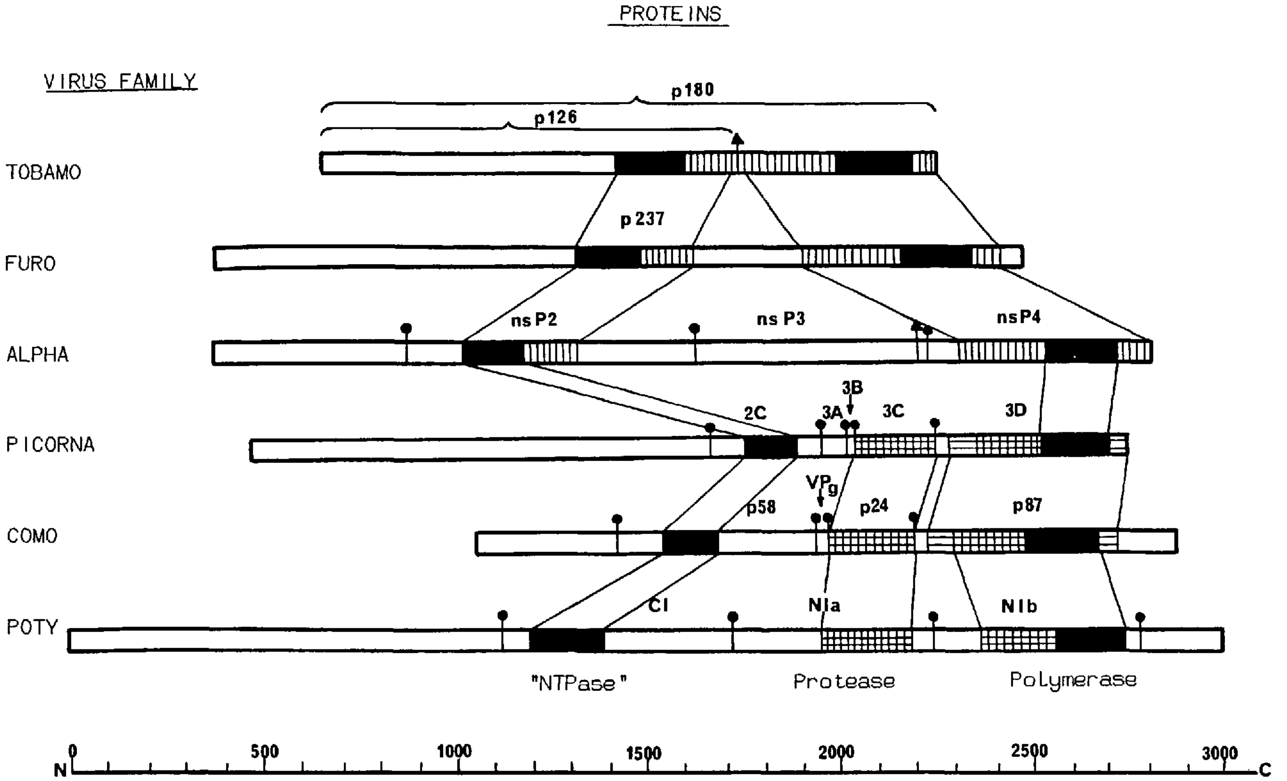

Fig. 4. The NTP-motif-containing proteins occupy similar positions in multidomain proteins of positive-strand RNA viruses. The whole polyproteins of picorna- and potyviruses, the nonstructural polyproteins of alpha- and comoviruses, the high-molecular weight protein encoded by furovirus RNA 1, and the multidomain protein encoded by the 5′ ORF of tobamovirus RNA are schematically represented by rectangles drawn to scale shown in the bottom of the figure. Regions of sequence similarity are highlighted as follows: ■, regions present in all viruses; ▦, regions observed in picorna-, como-, and potyviruses; ▤, regions observed in picorna- and comoviruses; ▥, regions observed in tobamo-, furo-, and alphaviruses (counterparts are also observed in tricorna-, hordei-, and coronavirus proteins). Similar sequence stretches are also joined by sloped lines. The NTP-motif-containing protein ("NTPase"), the RNA-dependent RNA polymerase (polymerase), and the protease (the latter identified in picorna-, como-, and potyviruses) are designated. Nominations of specific proteins are given above each rectangle. Other designations are: ●, sites of proteolytic processing; ▲, leaky termination codons [of two alphaviruses included, nsP4 is expressed only in SNBV via a leaky termination codon (Takkinen 1986)]. All the information is given only for the NTP-motif-containing proteins (domains), the polymerases, and the parts of multidomain proteins enclosed between.

tions uncertain. However, it should be noted that the segments of F1 and of p237 encompassing the consensus sequence bear no significant sequence similarity to the viral NTP-motif-containing domains described above (unpublished observations).

Flavivirus protein NS3, which occupies a position similar to that of alphavirus nsP2 in the polyproteins of these viruses, contains an "A" consensus sequence with a single deviation and a "B" sequence strikingly similar to those of the 1st and 3rd families of viral NTP-motif-containing proteins. Comparison of the three available sequences of NS3, those of yellow fever, West Nile, and dengue 2 flaviviruses (Rice et al. 1985; Castle et al. 1986; Yaegashi et al. 1986), demonstrated strict conservation of these sequences. A more detailed analysis that we recently performed revealed statistically significant similarity between the putative NTP-binding domains of NS3 and those of potyviral proteins (unpublished observations). It seems quite plausible that NS3 may have some degree of evolutionary and functional relatedness to the group of viral proteins described in this paper.

A striking similarity has been detected between the C-terminal sequence of the protein p120 encoded by BSMV RNA α [for which only a partial sequence has been reported (Rupasov et al. 1986)] and the C-terminal subdomain of the 1st family of viral NTP-motif-containing proteins. Although the N-terminal part of the p120 sequence is not yet known, in all other proteins of this family, invariably the two subdomains are observed together. Thus, the hordeivirus genome, like the furovirus genome, probably encodes two NTP-motif-containing proteins in two genomic segments (Gorbalenya et al. 1987).

In all other complete protein sequences of positive-strand RNA viruses reported, namely those of black beetle virus (a nodavirus), carnation mottle virus, and RNA bacteriophages, the consensus sequences of the NTP-motif have not been observed.

Discussion

The NTP-motif was first introduced by Walker et al. (1982) and was subsequently employed for localization of putative catalytic sites and for prediction of NTP-binding capacity in numerous proteins. However, for reasons mentioned in the Introduction, the validity of the whole approach remained rather uncertain. In the present study we demonstrate that in a highly diverged group including similar proteins of positive-strand RNA viruses, the consensus sequences of the NTP-motif constitute the most strictly conserved stretches, encompassing four of the five invariant amino acid residues. Moreover, the NTP-motif-containing domain is one of the two most conserved domains revealed upon an overall comparison of the sequences of this class of virus proteins. This strongly suggests that this protein domain possesses NTP-binding capacity and possibly NTPase activity, presumably supplying some NTP-dependent function(s) that is of vital importance for viral reproduction. This hypothesis is in agreement with the available experimental data implicating these proteins in viral RNA replication and with the reported NTP-binding capacity of TMV p126 (Evans et al. 1985), although direct testing is certainly warranted. In fact, there is experimental evidence that clearly, though indirectly, demonstrates the importance of the NTP-motif in viral RNA replication. Recently several poliovirus mutants resistant to or dependent on guanidine, a potent inhibitor of RNA replication of some picornaviruses, have been thoroughly studied (Pincus et al. 1986, 1987). They all mapped to the 2C protein, with the amino acid replacements located in the proximity of the "A" and "B" sites of the NTP-motif, or near the conserved Asn residue in the 183rd position of the segments of 2C aligned in this paper (Fig. 3).

Dever et al. (1987) have recently proposed a consensus for GTP-binding domains that includes, in addition to the "A" and "B" sites of the NTP-motif, a third highly conserved sequence element thought to determine the specificity for guanosine. They identified this sequence in the 2C protein of one serotype of FMDV (but not of the other picornaviruses) and suggested that this protein should possess specific GTP-binding capacity, as opposed to other picornaviral 2C proteins. However, it would be unprecedented for proteins so closely related to have different specificities for nucleotides. In our opinion, it is much more likely that, within groups of highly similar NTP-motif-containing proteins such as picornaviral 2C, the substrate specificities and other principal properties should be identical. On the other hand, when considering more distantly related proteins, such as those belonging to the three distinct families described above, one cannot exclude the possibility that such proteins might differ significantly in their activities and functions in viral reproduction.

It seems premature to discuss at length the possible significance of the present observations for understanding the evolution of positive-strand RNA viruses. Two trends, however are obvious. First, NTP-motif-containing proteins are nearly ubiquitous among eukaryotic positive-strand RNA viruses. The existing classification of these viruses (Matthews 1982) includes about 30 families (groups), or somewhat more, taking recent developments into consideration. For 13 of these, complete, or nearly complete genomic sequences are available. Proteins containing the typical NTP-motif were observed in nine families (Table 1); in addition, viruses of one family (Flaviviridae) probably possess a functionally related protein with a deviant motif. It is tempting to speculate that an NTP-dependent function supported by the amino acid residues constituting the NTP-motif may be indispensable for positive-strand viral RNA replication; in some cases this function may be supplied by cellular proteins. In this context it is compelling that the RNA replicase of single-stranded RNA bacteriophages contains the translation elongation factor Tu, an NTP-motif-containing GTPase, as one of its subunits (reviewed by Blumenthal 1979). Second, it appears that the sequence diversity of the NTP-motif-containing proteins as revealed here does not precisely reflect the "phenotypic" diversity of viruses that forms the basis for the existing classification. Of the nine virus families (groups) having NTP-motif-containing proteins, six contribute members to the 1st family of proteins (see above), two to the 2nd, and one to the 3rd family. Thus, the 1st family covers a very broad range of viral groups differing greatly in their genomic strategies and biological properties. It is anticipated that sequencing of genomes of new viral groups will add new members to this family.

The NTP-motif (or the "A" consensus alone) has been identified in an extremely large class of NTP-binding proteins, mostly NTPases (although it should be noted that the presence of this motif is not an absolute prerequisite for NTP-binding capacity). The NTP-motif-containing proteins include a large group of GTPases, namely the RAS family, G proteins, transducins, and some of translation initiation and elongation factors (Dever et al. 1987 and references therein). Also belonging to this class are numerous proteins involved in bacterial DNA synthesis, recombination, and repair, and in membrane transport (Doolittle et al. 1986; Finch et al. 1986a,b; Higgins et al. 1986; Husain et al. 1986; Yin et al. 1986; Gilchrist and Denhardt 1987), proteins implicated in multidrug resistance in mammalian cells (Chen et al. 1986; Gros et al. 1986), and several NTP-utilizing enzymes of DNA viruses (Gorbalenya et al. 1985; Anton and Lane 1986; Doolittle 1986a; Astell et al. 1987, and references therein). From this incomplete list it is obvious that the presence of the NTP-motif brings together numerous proteins with extremely diverse functions. It must be emphasized that many NTP-motif-containing proteins do not bear statistically significant similarity to each other (cf. Argos and Leberman 1985; Doolittle 1986a) and the existence of distinct monophyletic groups of such proteins (excluding very closely related, such as, for example, different RAS species) is not obvious a priori. Nevertheless, the NTP-motif-containing proteins of positive-strand RNA viruses do constitute such a family, whereas the GTPases probably constitute another.

Although widespread, NTP-motif-containing proteins are not strictly ubiquitous in all biological species. Specifically, this motif could not be found upon screening of the protein sequences of two large viral classes, negative-strand RNA viruses and retroid viruses (unpublished observations). Thus, the presence of proteins of this class in the majority of eukaryotic positive-strand RNA viruses appears to be a nontrivial observation, given their small genome size.

As for the value of the NTP-motif as a predictor of protein function, we believe that searching amino acid sequences for this motif (and conceivably for other sequence motifs of this kind) may be a very powerful methodology, if accompanied by phylogenetic analysis.

Addendum

During preparation and reviewing of this manuscript, important relevant information became available. Genome sequences of viruses of three more groups that encode NTP-motif-containing proteins were determined. These are tobacco rattle virus [a tobravirus (Hamilton et al. 1987)], white clover mosaic virus and potato virus X [two potexviruses (Forster et al. 1988; Krayev et al. 1988)], and tomato black ring virus [a nepovirus (C. Fritsch, personal communication)]. The presumptive NTP-binding domain of the tobravirus is closely related to that of TMV and clearly belongs to the 1st family of viral NTP-motif-containing proteins described above. The genomes of potexviruses each encode two NTP-motif-containing proteins. These proteins also belong to the 1st family, but their inclusion in the alignment further loosens the consensus. Interestingly, in some positions of the potexvirus proteins, residues otherwise invariant in the 1st family are replaced by those characteristic of the 2nd family. The nepovirus NTP-motif-containing protein belongs to the 2nd family. Thus, the new data appear to confirm our prediction that sequencing of genomes of viruses belonging to new groups should add members mainly to the 1st family of NTP-motif-containing proteins. Also, the genome sequence of southern bean mosaic virus, a sobemovirus, has been determined (Wu et al. 1987). The authors claimed that it encoded a presumptive NTP-binding domain. However, a more detailed analysis indicates that this domain probably fulfills an entirely different function, namely the protease one, with its sequence being strikingly similar to those of picornaviral proteases (Gorbalenya et al. 1988a). Thus, sobemoviruses may lack an NTP-motif-containing protein, which is similar to other positive-strand RNA viruses of small genome size (namely nodaviruses, carnation mottle virus, and RNA phages).

Comparison of the sequences of the 1st family of viral NTP-motif-containing proteins with those of several bacterial helicases revealed highly significant similarity, suggesting an RNA helicase function for these proteins (Gorbalenya and Koonin 1988; Gorbalenya et al. 1988b,c; Hodgman 1988).

Acknowledgments. The authors are deeply grateful to Professor V.I. Agol for constant interest and encouragement, to Dr. K.M. Chumakov for help with some of the computer programs, to Dr. S.Y. Morozov for useful discussions, and to Drs. C. Fritsch and S.Y. Morozov for communicating their sequence data prior to publication.

References

Ahlquist P, Dasgupta R, Kaesberg P (1984) Nucleotide sequence of the brome mosaic virus genome and its implications for viral replication. J Mol Biol 172:369–383

Ahlquist P, Strauss EG, Rice CM, Strauss JH, Haseloff J, Zimmern D (1985) Sindbis virus proteins nsP1 and nsP2 contain homology to nonstructural proteins from several RNA plant viruses. J Virol 53:536–542

Allison R, Johnston RE, Dougherty WG (1986) The nucleotide sequence of the coding region of tobacco etch virus genomic RNA: evidence for the synthesis of a single polyprotein. Virology 154:9–20

Anton IA, Lane DP (1986) Non-structural protein 1 of parvoviruses: homology to purine nucleotide using proteins and early proteins of papovaviruses. Nucleic Acids Res 14:7613

Argos P, Leberman R (1985) Homologies and anomalies in primary structural patterns of nucleotide binding proteins. Eur J Biochem 152:651–656

Argos P, Kamer G, Nicklin MJH, Wimmer E (1984) Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families. Nucleic Acids Res 12:7251–7267

Astell CR, Mol CL, Anderson WF (1987) Structural and functional homology of parvovirus and papovavirus polypeptides. J Gen Virol 68:885–893

Blumenthal T (1979) Q_β RNA replicase and protein synthesis elongation factors EF-Tu and EF-Ts. Methods Enzymol 60: 628–638

Boursnell MEG, Brown TDK, Foulds IJ, Green PF, Tomley FM, Binns MM (1987) Completion of the sequence of the genome of the coronavirus avian infectious bronchitis virus. J Gen Virol 68:57–77

Bouzoubaa S, Ziegler V, Beck D, Guilley H, Richards K, Jonard G (1986) Nucleotide sequence of beet necrotic yellow vein virus RNA-2. J Gen Virol 67:1689–1700

Bouzoubaa S, Quillet L, Guilley H, Jonard G, Richards K (1987) Nucleotide sequence of beet necrotic yellow vein virus RNA-1. J Gen Virol 68:615–626

Bradley MK, Smith TF, Lathrop FH, Livingston DM (1987) Consensus topography in the ATP binding site of the SV40 and polyomavirus large tumour antigens. Proc Natl Acad Sci USA 84:4026–4030

Carrington JC, Dougherty WG (1987) Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. J Virol 61:2540–2548

Carroll AR, Rowlands DJ, Clarke BE (1984) The complete nucleotide sequence of the RNA coding for the primary translation product of foot and mouth disease virus. Nucleic Acids Res 12:2461–2472

Castle E, Leidner U, Nowak T, Wengler G, Wengler G (1986) Primary structure of the West Nile flavivirus genome region coding for all nonstructural proteins. Virology 149:10–26

Chen C, Chin JE, Ueda K, Clark DP, Pastan I, Gettesman MM, Roninson IB (1986) Internal duplication and homology with bacterial transport proteins in the mdr1 (P-glycoprotein) gene from multidrug-resistant human cells. Cell 47:381–389

Clertant P, Seif I (1984) A common function for polyomavirus large-T and papillomavirus E1 proteins. Nature 311:276–279

Cornelissen BJC, Bol JF (1984) Homology between the proteins encoded by tobacco mosaic virus and two tricornaviruses. Plant Mol Biol 3:379–384

Cornelissen BJC, Brederode FT, Moormann RJM, Bol JF (1983) Complete nucleotide sequence of alfalfa mosaic virus RNA 1. Nucleic Acids Res 11:1253–1265

Dayhoff MO, Barker WC, Hunt LT (1983) Establishing homologies in protein sequences. Methods Enzymol 91:524–549

Dever TE, Glynias MJ, Merrick WC (1987) GTP-binding domains: three consensus sequence elements with distinct spacing. Proc Natl Acad Sci USA 84:1814–1818

Domier LL, Franklin KM, Shahabuddin M, Hellmann GM, Overmeyer JH, Hiremath ST, Siaw MFE, Lomonosoff GP, Shaw JG, Rhoads RE (1986) The nucleotide sequence of tobacco vein mottling virus RNA. Nucleic Acids Res 14: 5417–5430

Domier LL, Shaw JG, Rhoads RE (1987) Potyviral proteins share amino acid sequence homology with picorna-, como- and caulimoviral proteins. Virology 158:20–27

Doolittle RF (1981) Similar amino acid sequences: chance or common ancestry? Science 214:149–159

Doolittle RF (1986a) Protein sequence data banks: the continuing search for related sequences. In: Inouye M (ed) Protein engineering. Academic Press, New York, pp 15–27

Doolittle RF (1986b) Of URFs and ORFs. A primer on how to analyze derived amino acid sequences. Univ. Science Books, Mill Valley CA

Doolittle RF, Johnson MS, Husain I, Van Houten B, Thomas DC, Sancar A (1986) Domainal evolution of a prokaryotic DNA repair protein and its relationship to active transport proteins. Nature 323:451–453

Evans RK, Haley BE, Roth DA (1985) Photoaffinity labeling of a viral induced protein from tobacco. J Biol Chem 260: 7800–7804

Finch PW, Storey A, Chapman KE, Brown K, Hickson ID, Emmerson PT (1986a) Complete nucleotide sequence of the *Escherichia coli* recB gene. Nucleic Acid. Res 14:8573–8582

Finch PW, Storey A, Brown K, Hickson ID, Emmerson PT (1986b) Complete nucleotide sequence of recD, the structural gene for the α subunit of exonuclease V of *Escherichia coli.* Nucleic Acids Res 14:8583–8594

Forster RLS, Bevan MW, Harbison S-A, Gardner RC (1988) The complete nucleotide sequence of the potexvirus white clover mosaic virus. Nucleic Acids Res 16:290–303

Franssen H, Leunissen J, Goldbach R, Lomonosoff GP, Zimmern D (1984) Homologous sequences in non-structural proteins from cowpea mosaic virus and picornaviruses. EMBO J 3:855–861

Fry DC, Kuby SA, Mildvan AS (1986) ATP-binding site of adenylate kinase: mechanistic implications of its homology with ras-encoded p21, F_1-ATPase, and other nucleotide-binding proteins. Proc Natl Acad Sci USA 83:907–911

Gay NJ, Walker JE (1983) Homology between human bladder cacrinoma oncogene product and mitochondrial ATP-synthase. Nature 301:262–264

Gilchrist CA, Denhardt DT (1987) *Escherichia coli* rep gene: sequence of the gene, the encoded helicase, and its homology with uvrD. Nucleic Acids Res 15:465–475

Goelet P, Lomonosoff GP, Butler PJG, Akam ME, Gait MJ, Karn J (1982) Nucleotide sequence of tobacco mosaic virus RNA. Proc Natl Acad Sci USA 79:5818–5822

Goldbach R (1986) Molecular evolution of plant RNA viruses. Annu Rev Phytopathol 24:289–310

Gorbalenya AE, Blinov VM, Koonin EV (1985) Prediction of nucleotide-binding properties of virus-specific proteins from their primary structure. Molek Genetika No.11:30–36 [in Russian]

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1987) Two segments of barley stripe mosaic virus genomic RNA encode two homologous proteins which probably possess NTPase activity. Molek Biol 21:1566–1572 [in Russian]

Gorbalenya AE, Koonin EV (1988) One more conserved sequence motif in helicases. Nucleic Acids Res 16 (in press)

Gorbalenya AE, Koonin EV, Blinov VM, Donchenko AP (1988a) Sobemovirus genome appears to encode a serine protease related to cysteine proteases of picornaviruses. FEBS Letters (in press)

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1988b) A conserved NTP-motif in putative helicases. Nature 333:22

Gorbalenya AE, Koonin EV, Donchenko AP, Blinov VM (1988c) A novel superfamily of nucleoside triphosphate-binding motif containing proteins which are probably involved in duplex unwinding in DNA and RNA replication and recombination. FEBS Letters (in press)

Gros P, Croop J, Housman D (1986) Mammalian multidrug resistance gene: complete cDNA sequence indicates strong homology to bacterial transport proteins. Cell 47:371–380

Gustafson G, Armour SL (1986) The complete nucleotide sequence of RNAβ from the type strain of barley stripe mosaic virus. Nucleic Acids Res 14:3895–3909

Halliday K (1984) Regional homology in GTP-binding protooncogene and elongation factors. J Cyclic Nucleotide Protein Phosphorylation Res 9:435–448

Hamilton WDO, Boccara M, Robinson DJ, Baulcombe DC

(1987) The complete nucleotide sequence of tobacco rattle virus RNA-1. J Gen Virol 68:2563–2575

Higgins CF, Hiles ID, Salmond GPC, Gill DR, Downie JA, Evans IJ, Holland IB, Gray L, Buckel SD,Bell AW, Hermodson MA (1986) A family of related ATP-binding subunits coupled to many distinct biological processes in bacteria. Nature 323:448–450

Hodgman TC (1986) The elucidation of protein function from its amino acid sequence. CABIOS 2:181–187

Hodgman TC (1988) A new superfamily of replicative proteins. Nature 333:22–23, 578

Husain I, van Houten B, Thomas DC, Sancar A (1986) Sequences of *Escherichia coli* uvrA gene and protein reveal two potential ATP binding sites. J Biol Chem 261:4895–4901

Jurnak F (1985) Structure of the GDP domain of EF-Tu and location of the amino acids homologous to ras oncogene proteins. Science 230:32–36

Kamer G, Argos P (1984) Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res 12:7269–7282

Koonin EV, Gorbalenya AE, Chumakov KM, Donchenko AP, Blinov VM (1987) Evolution of RNA-dependent RNA polymerases of positive strand RNA viruses. Molek Genetika No. 7:27–39 [in Russian]

Koonin EV, Chumakov KM, Yushmanov SV, Gorbalenya AE (1988) Evolution of RNA-dependent RNA polymerases of positive strand RNA viruses: a comparison of phylogenetic trees generated by different methods. Molek Genetika No. 3: 16–19 [in Russian]

Krayev AS, Morozov SY, Lukasheva LI, Rozanov MN, Chernov BK, Simonova ML, Golova YB, Belzhelarskaya SN, Pozmogova GE, Skryabin KG, Atabekov JG (1988) The complete nucleotide sequence and genomic organization of potato X virus. Dokl Akad Nauk SSSR 300:711–716 [in Russian]

La Cour TFM, Nyborg J, Thirup S, Clark BFC (1985) Structural details of the binding of guanosine diphosphate to elongation factor Tu from *Escherichia coli* as studied by X-ray crystallography. EMBO J 4:2385–2388

Lindberg AM, Stalhandske POK, Pettersson U (1987) Genome of coxsackievirus B3. Virology 156:50–63

Lomonosoff GP, Shanks M (1983) The nucleotide sequence of cowpea mosaic B RNA. EMBO J 2:2253–2258

Matthews REF (1982) Classification and nomenclature of viruses. Intervirology 17:1–199

Möller W, Amons R (1985) Phosphate-binding sequences in nucleotide-binding proteins. FEBS Lett 186:1–7

Morozov SY, Rupasov VV (1985) On the possibility of a common origin of the genes encoding the RNA polymerases of bacterial, plant and animal positive strand RNA viruses. Biol Nauki No. 10:19–23 [in Russian]

Najarian R, Caput D, Gee W, Potter SJ, Renard A, Merryweather J, Van Nest G, Dina D (1985) Primary structure and gene organization of human hepatitis A virus. Proc Natl Acad Sci USA 82:2627–2631

Palmenberg AG, Kirby EM, Janda MR, Drake NL, Duke GM, Potratz KF, Collett MS (1984) The nucleotide abd deduced amino acid sequences of the encephalomyocarditis viral polyprotein coding region. Nucleic Acids Res 12:2969–2985

Pevear DC, Calenoff M, Rozhon E, Lipton HL (1987) Analysis of the complete nucleotide sequence of the picornavirus Theiler's murine encephalomyelitis virus (TMEV) indicates that it is closely related to cardioviruses. J Virol 61:1507–1516

Pincus SV, Diamond DC, Emini EA, Wimmer E (1986) Guanidine-selected mutants of poliovirus: mapping of point mutations to polypeptide 2C. J Virol 57:638–646

Pincus SE, Rohl H, Wimmer E (1987) Guanidine-dependent mutants of poliovirus: identification of three classes with different growth requirements. Virology 157:83–88

Pozdnyakov VI, Pankov YA (1981) Accelerated method for comparing amino acid sequences with allowance for possible gaps. Plotting optimum correspondence paths. Int J Pept Protein Res 17:284–291

Racaniello V, Baltimore D (1981) Molecular cloning of poliovirus cDNA and determination of the complete nucleotide sequence of the viral genome. Proc Natl Acad Sci USA 78: 4887–4891

Rezaian MA, Williams RHV, Symons R (1985) Nucleotide sequence of cucumber mosaic virus RNA 1. Eur J Biochem 150:331–339

Rice CM, Lenches EM, Eddy SR, Shin SJ, Sheets R, Strauss JH (1985) Nucleotide sequence of yellow fever virus implications for flavivirus gene expression and evolution. Science 229:726–733

Rupasov VV, Afanasiev BN, Adyshev DA, Kozlov YV (1986) Nucleotide sequence of 3'-terminal regions of barley stripe mosaic virus RNAs 1 and 3. Dokl Akad Nauk SSSR 288: 1237–1241 [in Russian]

Sankoff D (1972) Matching sequences under deletion/insertion constraints. Proc Natl Acad Sci USA 69:4–6

Skern T, Sommergruber W, Blaas D, Gruendler P, Fraundorfer F, Pieler C, Fogy I, Kuechler E (1985) Human rhinovirus 2: complete nucleotide sequence and proteolytic processing signals in the capsid protein region. Nucleic Acids Res 13: 2111–2126

Staden R (1982) An interactive graphics programme for comparing and aligning nucleic acid and amino acid sequences. Nucleic Acids Res 10:2951–2961

Stanway G, Hughes PJ, Mountford RC, Minor PD, Almond JW (1984) The complete sequence of a common cold virus: human rhinovirus 14. Nucleic Acids Res 12:7859–7875

Strauss EG, Rice CM, Strauss JH (1984) Complete nucleotide sequence of the genomic RNA of Sindbis virus. Virology 133: 92–110

Takkinen A (1986) Complete nucleotide sequence of the nonstructural protein genes of Semliki Forest virus. Nucleic Acids Res 14:5667–5682

Walker JE, Saraste M, Runswick MJ, Gay NJ (1982) Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO J 2:945–951

Wu S, Rinehart CA, Kaesberg P (1987) Sequence and organization of southern bean mosaic virus genomic RNA. Virology 161:73–80

Yaegashi T, Vakharia VN, Page K, Sasaguri Y, Feighny R, Padmanabhan R (1986) Partial sequence analysis of cloned dengue virus type 2 genome. Gene 46:257–267

Yin K-C, Blinkova A, Walker JR (1986) Nucleotide sequence of the *Escherichia coli* replication gene dnaZX. Nucleic Acids Res 14:6541–6549

Received February 5, 1988/Revised April 10, 1988