Tick-borne encephalitis virus genome

The nucleotide sequence coding for virion structural proteins

A.G. Pletnev, V.F. Yamshchikov and V.M. Blinov*

Novosibirsk Institute of Bioorganic Chemistry, Siberian Division of the USSR Academy of Sciences, Novosibirsk 90 and **All-Union Research Institute of Molecular Biology, Koltsovo, Novosibirsk District, USSR*

Received 18 March 1986

RNA of a flavivirus, tick-borne encephalitis virus (TBEV; strain Sofjin), was subjected to reverse transcription and the DNA copy was transformed into double-stranded DNA by the action of *E. coli* DNA-polymerase I (Klenow fragment). This DNA was annealed with plasmid pBR322. The recombinant plasmids were cloned in *E. coli* K802. The nucleotide sequence of the inserts of the clones, coding for region structural proteins C, M, E and nonstructural protein NS1, was determined by the Maxam-Gilbert method. The genes of structural proteins form a compact cluster. Homology has been studied of the TBEV sequences found with the structures of proteins and RNAs of other flaviviruses, yellow fever virus and West Nile virus, and a high degree of homology was found.

Encephalitis virus *cDNA cloning* *Nucleotide sequence* *Viral protein* *Protein structure*

1. INTRODUCTION

Tick-borne encephalitis is a significant public health problem in some districts of the USSR and of other countries in Europe and Asia. The virus of this disease belongs to the flaviviruses of the Togaviridae family. Its virion is a nucleocapsid covered by a lipoprotein envelope. The envelope consists of two proteins: protein M and glycoprotein E having molecular masses of some 7.5–8 and 53–60 kDa, respectively. The nucleocapsid is a complex of protein C (14 kDa) and mRNA (4×10^6 kDa) [1]. The tick-borne encephalitis virus (TBEV) causes generalized infection in humans and other animals leading to fever and encephalitis.

Direct studies of the TBEV genome are complicated by its high degree of pathogenicity and by the infectivity of its RNA. Genetic engineering techniques help to overcome these difficulties. We have reported cloning of DNA copies of TBEV genome fragments in *E. coli*, the nucleotide sequences of some of these copies and the localization on the genome of the genes of proteins C and E [2,3]. This paper gives an account of studies on the part of the genome coding for proteins C, M and E. A high degree of homology exists between the amino acid sequences of the structural proteins of TBEV, yellow fever virus (YFV) [4] and West Nile virus (WNV) [5].

2. MATERIALS AND METHODS

TBEV (strain Sofjin) was isolated and purified as described in [6] by Dr S.G. Rubin (Institute of Polyomyelitis and Viral Encephalitides, Moscow). TBEV RNA was obtained by phenol deproteinization and purified by ultracentrifugation in a 5–20% sucrose gradient. The preparation of cDNAs, their cloning and the selection of clones of recombinant plasmids have been described [2]. Cross-hybridization of cDNA inserts gave the map of these inserts [3]. Sequencing was performed by the Maxam-Gilbert method [7] with modifications described in [8].

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50

3. RESULTS AND DISCUSSION

The part of the genome coding for the structural proteins of TBEV (strain Sofjin) has been identified [3] by comparison of the nucleotide sequence with the known N-terminal amino acid sequences of proteins C and E of TBEV (strain Neudörfl) [9] and with the amino acid sequences of the structural proteins of some other flaviviruses [10]. The sequences of four tryptic peptides of TBEV (strain Sofjin) protein E [11] are in accordance with the nucleotide sequence presented in [3]. Since then our efforts have been aimed at the elucidation of the nucleotide sequence over that part of the genome covered by recombinant plasmids 2, 4 and 10 (fig.1). The nucleotide sequence found and the amino acid sequences of proteins C, M, E and NS1 which follow from this nucleotide sequence are presented in figs 2 and 3.

The structures of the genes coding for proteins C and M of WNV [4] and the complete sequence of YFV RNA [5] have been published. Both viruses are mosquito-born flaviviruses and are distant relatives of TBEV. We looked for homologous amino acid sequences of proteins from these three viruses and nucleotide sequences of their genomes using the computer programs described in [12,13].

The genes of structural proteins C, M and E of these flaviviruses form a cluster shifted to the 5′-termini of viral RNAs. This cluster of TBEV RNA is translated in vitro into a 90 kDa polycistronic protein [14] which is subsequently

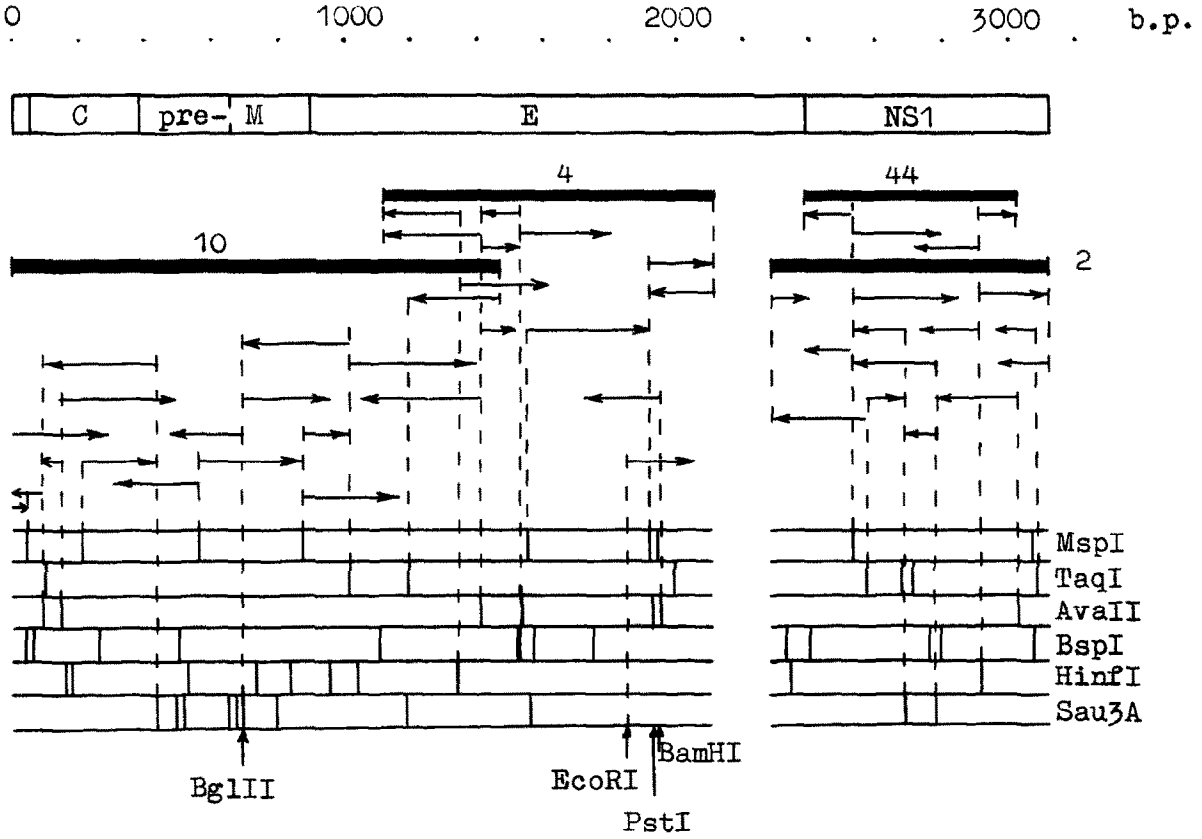

Fig.1. Scheme of the sequencing and the restriction nuclease sites of the inserts of recombinant plasmids of clones 2, 4 and 10. The numbers of clones are indicated above the corresponding inserts. The inserts were removed by hydrolysis with *Pst*I and isolated by electrophoresis in 1% agarose gel. The subfragments obtained by action of restriction nucleases are shown as boxes. The arrows show the direction from the ^{32}P label; the solid part corresponds to the sequence elucidated; the broken line denotes the sequence which was not elucidated in a given experiment. C, pre-M, E and NS1, genes of the corresponding proteins of TBEV.

Fig.2. Nucleotide sequence of the DNAs of the recombinant plasmid inserts. (A) Clones 4 and 10. (B) Clone 2. The sequence of the chain homologous to that of TBEV RNA is presented. The methionine codon corresponding to the start of polycistronic protein synthesis is boxed.

cleaved into proteins C and E by membrane proteases [15]. The protein C is coded by the 5′-terminal part of the cluster. Its synthesis starts with methionine (ATG codon, positions 49–51, fig.2) which is cleaved off in the course of maturation. The molecular mass of protein C, according to the nucleotide sequence, is 10580 Da. The amino acid composition following from the structure of the gene is in good accord with the data of direct analysis [9]. The high content of Lys and Arg (25.5%) may be necessary for the interaction with RNA in the course of nucleocapsid folding. The maximum homology of the amino acid sequences from the C proteins of different flaviviruses is as follows: TBEV/YFV, 32.4%; TBEV/WNV, 26.8%; YFV/WNV, 28.7%. Noteworthy is the high degree of homology between the N-termini of the C proteins from two strains of TBEV: Neudörfl, the European strain, and Sofjin, the strain obtained from a patient in the Far East. The C-terminal amino acid sequence of protein C of TBEV is unknown. The presence of a site of cleavage for cellular proteases -Lys-Arg-Arg-Ser- (residues 107–110, fig.3) in the polypeptide chain suggests that cleavage at this site takes place in the course of maturation of TBEV and YFV.

The genes of protein M follow those of protein C in the genomes. No protein M is found in cells infected by flaviviruses, being present only in mature virions. This polypeptide is formed via a glycosylated precursor of M, or gp19, NV2 [15–17]. The pre-M protein of WNV has the N-terminal sequence Val-Thr-Leu- which is also present at positions 131–133 in polycistronic proteins of all three flaviviruses. Comparison of the amino acid sequences of the pre-M proteins of TBEV, WNV and YFV reveals a strong homology (31–37%). It is remarkable that the Cys residues in all pre-M proteins occupy the same positions. Clusters of noncharged amino acids are present at the N- and C-termini of pre-M. These clusters may play the role of 'anchors' holding glycosylated pre-M on the membrane during maturation of M. The potential glycosylation site of TBEV pre-M is the sequence Arg-Arg-Ser (positions 224–226). Protein M of TBEV most probably covers the sequence between Ser 226 and Ser 302; this sequence has a strong homology with the proteins M of WNV and YFV. However, the N-terminal sequence of TBEV E protein has been reported by Boege et al. [9] as Ser-Val-Leu-Ile-. According to

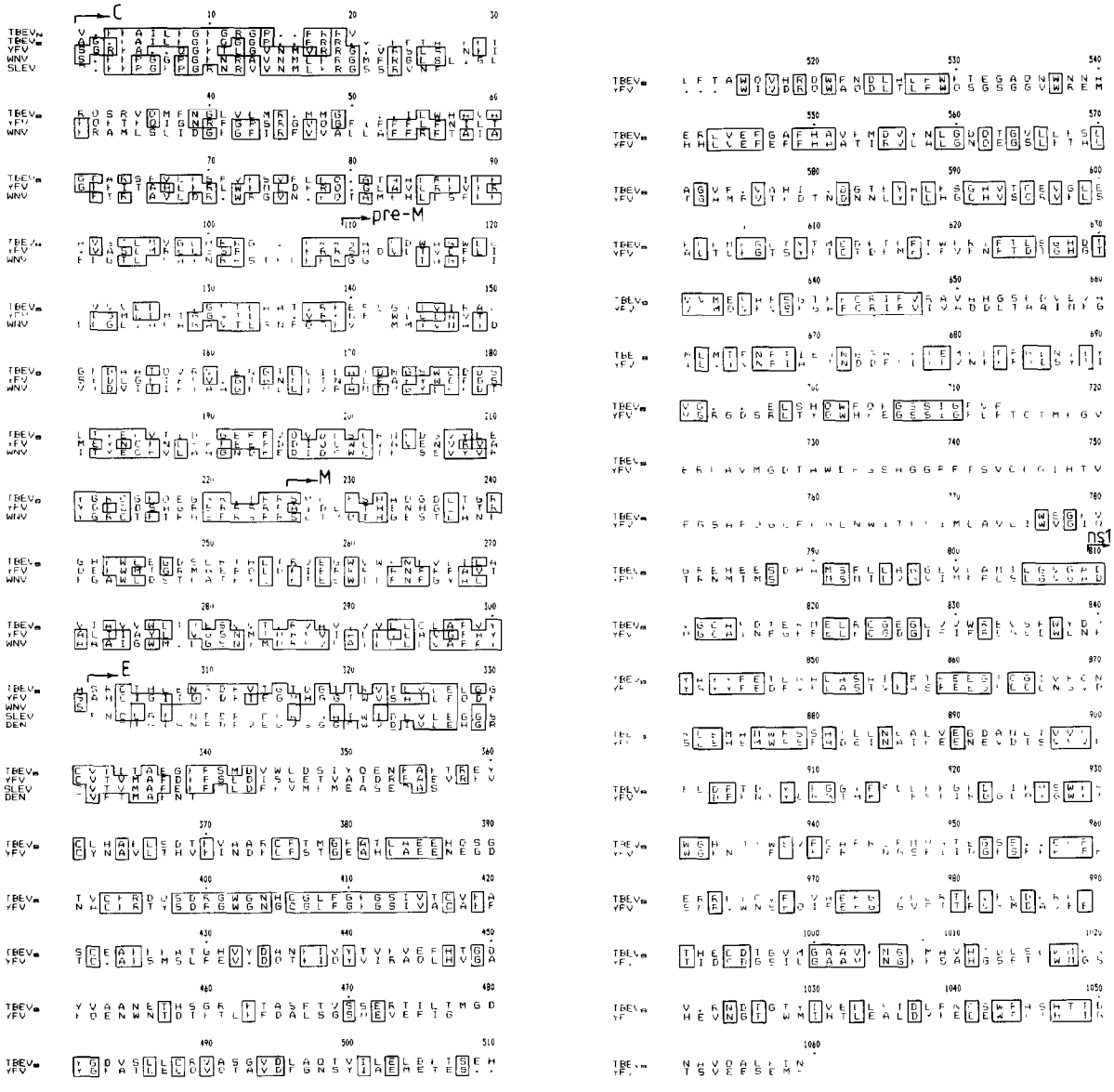

Fig.3. Amino acid sequences of the proteins C, pre-M, E and NS1 of some flaviviruses. $TBEV_N$, N-terminal sequence of protein C of TBEV (strain Neudörfl) [9]; N-terminal sequences of protein E from SLEV (Saint Louis encephalitis virus) and DENV (Denge-2 virus) are given according to [10]; amino acid sequences of proteins C and pre-M of WNV according to [5], and those of YFV according to [4]; $TBEV_S$, amino acid sequences of the proteins of TBEV (strain Sofjin) following from the nucleotide sequences of the inserts of plasmids of the clones 4, 10 and 2 (see fig.2). Arrows show the N-termini of the corresponding proteins. The sequences are presented according to the IUPAC-IUB single-letter code: A-Ala; C-Cys; D-Asp; E-Glu; F-Phe; G-Gly; H-His; I-Ile; K-Lys; L-Leu; M-Met; N-Asn; P-Pro; Q-Gln; R-Arg; S-Ser; T-Thr; V-Val; W-Trp; Y-Tyr.

our data [11], this sequence is adjacent to the processing site Arg-Arg-Ser (residues 224–226) and belongs most probably to protein M. This discrepancy may be due to two alternative reasons. Firstly, the protein studied by Boege et al. [9] could be fused M-E protein which escaped cleavage by

cellular maturation proteases. Secondly, it may be that there is actually no processing site between proteins M and E in TBEV (tick-borne), while this site exists in the polycistronic proteins of WNV and YFV (mosquito-borne).

Protein E is a glycoprotein and is the major antigen of TBEV. Two potential glycosylation sites are present in its structure, Asn-Glu-Thr (455–457) and Asn-Pro-Thr (667–669). Homology between proteins E of TBEV and YFV is 35%. The Cys residues occupy the same positions in these two proteins suggesting a similarity of their three-dimensional structures and a common origin. The proteins are terminated by clusters of noncharged amino acid residues including the sequence -Leu-Gly-Val-Gly-Ala. The hydrophobic signals near the terminus may help to hold it on the cell membranes where assembly of virions takes place. Earlier we were unable to elucidate the amino acid sequence of protein E of TBEV over the region 715–777 because we failed to find a plasmid covering the gap between the inserts of clones 4 and 2.

The insert of the plasmid of clone 2 contains the gene of the first nonstructural protein NS1. The function of this protein is unknown. The protein NS1 is a glycoprotein; its homology with NS1 of YFV is 37.9%.

The homology of amino acid sequences of the three flaviviruses TBEV, WNV and YFV may be due to their common origin. The homology is even more obvious in the nucleotide sequences. Taking into account deletions and inserts, the homology between TBEV and YFV genomes is 45%, and 40% of the differences occur at the third positions of codons and do not change the amino acid residues.

Hence, flaviviruses are conservative in the structure of both nucleic acids and proteins. The genes are located on genomes in a similar way, and the strategy of expression also seems to be similar. Remarkably, the structures of the genomes of flaviviruses have an organization which is drastically different from that characteristic of alphaviruses also belonging to the Togaviridae. Firstly, flaviviruses do not have sub-genome RNA. Secondly, the genes of structural proteins are located to the left of the nonstructural protein genes on the genome, while the situation is opposite in alphaviruses. Thirdly, the polycystronic mRNA of flaviviruses does not contain translation termination codons between the regions of structural and nonstructural genes. In the organization of genomes, flaviviruses are more similar to picornoviruses than to alphaviruses.

REFERENCES

[1] Westaway, E.G., Schlesinger, R.W., Dalrymple, J.M. and Trent, D.W. (1980) Intervirology 14, 114–117.
[2] Chumakov, M.P., Kusov, Yu.Yu., Rubin, S.G., Salnikov, Ya.A., Semashko, I.V., Georgiev, G.P., Chumakov, P.M., Grachev, M.A., Shamanin, V.A. and Pletnev, A.G. (1983) Bioorg. Khim. 9, 276–279.
[3] Pletnev, A.G. and Yamshchikov, V.F. (1985) Bioorg. Khim. 11, 1681–1684.
[4] Rice, C.M., Lenches, E.M., Eddy, S.R., Shin, S.J., Sheets, R.L. and Strauss, J.H. (1985) Science 229, 726–733.
[5] Castle, E., Nowak, T., Leidner, U., Wengler, G. and Wengler, G. (1985) Virology 145, 227–236.
[6] Chumakov, M.P., Kusov, Yu.Yu., Rubin, S.G., Semashko, I.V., Salnikov, Ya.A., Reingold, V.N., Pressman, E.K. and Tsekhanovskaya, N.A. (1984) Acta Virol. 6, 694–701.
[7] Maxam, A. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560–564.
[8] Chuvpilo, S.A. and Kravchenko, V.V. (1983) Bioorg. Khim. 9, 1634–1637.
[9] Boege, U., Heinz, F.X., Wengler, G. and Kunz, C. (1983) Virology 126, 651–657.
[10] Bell, J.R., Kinney, R.M., Trent, D.W., Lenches, E.M., Dalgarno, L. and Strauss, J.H. (1985) Virology 143, 224–229.
[11] Baram, G.I., Grachev, M.A., Nazimov, I.V., Pletnev, A.G., Pressman, E.K., Rubin, S.G., Salnikov, Ya.A., Semashko, I.V., Chumakov, M.P., Shemyakin, V.V. and Yamshchikov, V.F. (1985) Bioorg. Khim. 11, 1677–1680.
[12] McLachlan, A.D. (1971) J. Mol. Biol. 61, 409–424.
[13] Stalen, R. (1980) Nucleic Acids Res. 8, 3673–3694.
[14] Svitkin, Yu.V., Ugarova, T.Yu., Chernovskaya, T.V., Lyapustin, V.N., Lashkevich, V.A. and Agol, V.I. (1981) Virology 110, 26–34.
[15] Svitkin, Yu.V., Lyapustin, V.N., Lashkevich, V.A. and Agol, V.I. (1984) Virology 135, 536–541.
[16] Shope, R.E. (1980) in: The Togaviruses (Schlesinger, R.W. ed.) pp.47–82, Academic Press, New York.
[17] Westaway, E.G., Speight, G. and Endo, L. (1984) Virus Res. 1, 333–350.