Mathematical Biology and Bioinformatics
2018. V. 13. № S. P. t39-t58. doi: 10.17537/2018.13.t39

Translation into English of the original article published in Russian:
Zimin A.A., Mikoulinskaia G.V., Nigmatullina L.F., Nazipova N.N.
Mathematical Biology and Bioinformatics. 2012. V. 7. № 2. P. 611-631. doi: 10.17537/2012.7.611

========================== **BIOINFORMATICS** ===========================

UDC: 578.81, 51-76

Comparative Analysis of Amino Acid Sequences in Particular Domains of Hoc Proteins in *Teequatrovirinae* Subfamily Bacteriophages

Zimin A.A.[1], Mikoulinskaia G.V.[2], Nigmatullina L.F.[1], Nazipova N.N.[3]

[1]*G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms, Pushchino, Moscow Region, Russia*
[2]*Branch of M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Pushchino, Moscow Region, Russia*
[3]*Institute of Mathematical Problems of Biology RAS – the Branch of Keldysh Institute of Applied Mathematics, Pushchino, Moscow Region, Russia*

Abstract. The work presents the results of comparative research into immunoglobulin-like domains in the genomes of T4-related bacteriophages. Hoc proteins are proposed to be used for classification of the *Teequatrovirinae* phage subfamily. Particular domains in 31 Hoc proteins of the subfamily were subjected to phylogenetic analysis. The number of domains in Hoc proteins of different bacteriophages in the subfamily was shown to vary from one to five. Based on this, bacteriophages can be divided into six subgroups. The phylogenetic tree of the domains in *hoc* gene product proteins of T4-related bacteriophages forms three major branches. These are the branches of C-terminal, N-terminal and intermediate domains. The obligatory occurrence of the C-terminal domain in all Hoc proteins is indicative of its functional and structural significance for the formation of the protein and its attachment to phage's capsid. Hypothetical schemes for the evolutionary origin of repeated amino acid sequences in Hoc proteins were formulated.

Key words: *bacteriophage T4, Hoc protein, immunoglobulin-like proteins, domains, phylogenetic tree, protein evolution.*

INTRODUCTION

Bacteriophages are the most abundant population of biological beings on the planet [3]. Research shows that the diversity in this population of bacterial viruses is great [22]; this is provided for by the active role of the horizontal transfer of genetic material observed in bacteriophages. The taxonomic classification of bacteriophages is not a simple task. For a long time, the only classification system was that of the International Committee on the Taxonomy of Viruses (ICTV) [15], which was based on a number of physiological and morphological characters of organisms and on the nature of their genetic material (double- or single-stranded DNA, double- or single-stranded RNA). In the early 21 century, works appeared that made use of genomic sequences; new data were published that characterized the bacteriophage diversity [21, 22]. Unlike other species of organisms, for bacteriophages the use of ribosomal RNA genes – these taxonomic markers universal for most biological beings – is impossible because they have no ribosomal RNA. Besides, it has been shown that there is no universal protein marker that could be used for classification of bacteriophages.

The genomic classification of phages (the development of the phage proteomic tree) was performed based on the idea of the similarity of the proteome in related bacteriophages [22]; it agrees well with the ICTV taxonomic system. The authors of the genomic classification assumed that different proteins should be used for each branch of the phage taxonomic tree as markers.

There is an approach to the phage classification based on the theory of the modular evolution of phages [25]. According to this concept, phage genomes consist of genetic elements which they exchange in the process of evolution.

The goal of this work was to refine the classification of *Teequatrovirinae* subfamily bacteriophages (related to bacteriophage T4) using a synthetic approach based on the main ideas of the mentioned methods. The subfamily of T4-like *Teequatrovirinae* phages is divided by the similarity of amino acid sequences into two groups of genera (T4-like phages and KVP40-like viruses). We investigated the first of these groups, which contains four subgroups (T4-type, 44RR2.8-type, RB43-type, RB49-type) [20]. Each subgroup was studied separately.

Proteins containing immunoglobulin-like domains are attributed to the immunoglobulin superfamily [4, 28, 27, 8]. Immunoglobulin-like proteins have been found in all types of organisms of the animal kingdom [28, 27, 11], in bacteria [4] and viruses [5]. They perform the most diverse functions from adhesion up to manifestation of elasticity. There is special software developed for the search of segments characteristic of immunoglobulin-like domains in amino acid sequences [8].

Bacteriophage T4 – a large virus that infects *Escherichia coli* – belongs to the subfamily *Teequatrovirinae*, family *Myoviridae*, order *Caudovirales* [20]. It consists of an extended icosahedral head and a contractile tail that ends up with a basal plate, to which six long fibrillae are attached (see Fig. 2). The head (capsid) that comprises the genomic double-stranded DNA of 171,000 base pairs (bp) is constructed from three proteins, products of genes 20, 23 and 24.

Six proteins 23 form major capsomers, of which 20 icosahedral faces are formed; eleven vertices of the capsid are formed by pentamers of protein 24, and a special portal vertex to which the tail is attached is formed by twelve copies of protein 20 [10]. There are also two decoration proteins (Hoc (head outer capsid protein) and Soc (small outer capsid protein)), which are attached to the assembled capsid [16–18]. Hoc and Soc proteins are not required for assembling the capsid. Deletion of one or both genes encoding them does not lead to decrease production of the phage, its viability or infectivity under standard laboratory conditions.

Hoc protein is attached to the centre of the capsomer formed by hexamers of proteins 23; the stability of the capsid is not significantly affected by this [23]. Using cryoelectron microscopy, the molecule of Hoc protein was shown to have a dumb-bell shape. Probably, two balls of a dumbbell represent two functional modules. One of them, the conserved module, interacts with the capsid surface [24], and the variable module remote from the capsid interacts with the surface of the bacterial cell [9]. The amino acid subsequence ESRNG responsible for the attachment to the capsid is localized in 25 C-terminal amino acids, which contain a conserved predicted loop enclosed by two β-structures that orient the loop to the interaction with the main capsid protein [24]. The structural organization of the conserved module forms the point of support on the surface of the virus, and the variable module can be adjusted for interaction with various surfaces, including that of the bacterial cell.

The sequence analysis of T4 Hoc protein showed it to consist of four tandem immunoglobulin-like (Ig-like) domains [5]. The first three domains have a typical Ig-like folding, which usually consists of seven β-strands connected into two antiparallel β-sheets packed into a β-sandwich [12, 6].

All known Ig-like domains are combined into four groups: Ig domains (I-Set), fibronectin3 domains (FN3), bacterial Ig-like domains (Big2), polycystic kidney disease (PKD) domains ([14].

The first two domains of Hoc protein are similar by their amino acid sequence with PKD domains, and the third domain is similar to I-Set domains [24]. The fourth domain has a immunoglobulin-like packing similar to telokin domains [4]. The role of Hoc protein in the development of the bacteriophage and its organization has not been established yet.

Hoc proteins of various T4-related bacteriophages are heterogeneous along the length and can contain various numbers of domains similar by their amino acid sequence and structure [9]. The nature of the heterogeneity of Hoc proteins as well as the pathways of their formation in their current shape can be studied by comparative analysis of the amino acid sequences of particular domains. Such a study will enable a classification of phages inside the subfamily, and, besides, will shed light on the evolution pathways of Hoc and other proteins containing the repeated immunoglobulin-like domains. Search for proteins that have sequences similar to particular Hoc domains can help in establishing the biological function of this protein.

METHODS

To perform the task, we compared amino acid sequences of particular domains in Hoc proteins of phages T4 and RB49 with the protein sequence databases on the NCBI server [2] by the PSI-BLAST algorithm [1] with a confidence level of the results E < 0.0001. Herewith, the iterated profile search method for the conservative motives in particular domains was performed until each of the next successive iteration revealed new local similarities in the GenBank database. Four or five steps of profile search were sufficient to convergence of algorithm in each of cases. Therefore, five iterations of PSI-BLAST were sufficient for meaningful results.

Based on the obtained results, we performed a phylogenetic analysis of particular domains in 31 proteins of *Teequatrovirinae* subfamily bacteriophages. For this, the sequences were aligned by ClustalX [13]; the phylogenetic tree was constructed using the Mega4 software package [26].

RESULTS AND DISCUSSION

1. Classification of *hoc* gene product proteins

The amino acid sequence of Hoc protein in bacteriophage T4 has a length of 376 amino acids (a.a.) and consists of four like segments, each of which has a pronounced similarity with immunoglobulins. The first three segments are rather strongly similar between themselves. The repeated fragments in them are 94 a.a. in length. For the first three similar segments from the related protein of phage RB49, as the result of X-ray diffraction analysis, it was shown that each of them formed a particular domain [9]. That work has also shown that Hoc protein is capable of attaching to the *E. coli* cell surface. For this reason, we call repeated Hoc protein fragments domains. Homologous Hoc proteins were found in most phages of the subfamily *Teequatrovirinae*. A total of 31 amino acid sequences were analyzed. A classification of *hoc* gene product proteins is given in Table 1.

By the number of repeated domains we subdivide homologous Hoc proteins into six main groups: single-domain, two-domain, three-domain, four-domain; four-domain with C-terminal elongation of the third domain; five-domain proteins.

The first of the groups is distinguished by the greatest diversity of proteins by length (61–167 a.a.). The minimal-length protein of phage RB43 (61 a.a.) does not contain the sequence ESRNG involved in the attachment to the capsid. The *hoc* gene product protein of phage RB16 contains one usual C-terminal domain, which has a C-terminal tail in the form of a 71 amino-acids long sequence. This segment of the sequence has no similarity with the repeated sequences of Hoc proteins. As it is known [17], Hoc protein is the major antigen of bacteriophage T4. Probably, the additional sequence of Hoc protein in phage RB16 is responsible for the new antigen properties of this phage.

The group of two-domain variants varies by length from 177 up to 180 a.a.; three-domain variants contain from 264 up to 282 a.a.; four-domain variants, from 367 up to 377 a.a. Five-domain variants insignificantly vary by length and contain from 469 up to 474 a.a. Four-domain variants with C-terminal tail of the third domain are homogenous by length, they are 404 a.a. each.

Table 1. Classification of *hoc* gene product proteins

№	Phage name	Protein length	Number of domains	Protein identifier	Maximal identity to Hoc protein of T4 (%)	Maximal identity to Hoc protein of RB49 (%)
Single-Domain Proteins						
1	Enterobacteria phage RB43	61	1	AAX78759.1	33	55
2	Klebsiella phage KP15_01	91	1	YP_003580072.1	36	49
3	Acinetobacter phage Acj61	106	1	YP_004009803.1	68	38
4	Enterobacteria phage RB16	167	1 (2?)	ADJ55528.1	40	48
Two-Domain Proteins						
5	Aeromonas phage 31	180	2	AAX63659.1	29	32
6	Aeromonas phage 44RR2.8t	180	2	AAQ81490.1	29	32
7	Aeromonas phage 25	177	2	ABF72722.1	24	31
8	Klebsiella phage KP15_02	177	2	ADE35027.1	45	41
9	Aeromonas phage phi AS4	178	2	YP_003969113.1	25	32
Three-Domain Proteins						
10	Aeromonas phage 65_01	264	3	YP_004300980.1	37	39
11	Shigella phage SP18	282	3	YP_003934817.1	59	
12	Enterobacteria phage Bp7	282	3	AEN93941.1	59	34
13	Enterobacteria phage JS98	282	3	YP_001595307.1	59	35
14	Enterobacteria phage IME08_01	288	3	YP_003734322.1	43	37
15	Enterobacteria phage JS10_01	286	3	YP_002922526.1	29	33
Four-Domain Proteins						
16	Enterobacteria phage T4	376	4	AAD42581.1	100	35
17	Enterobacteria phage vB_EcoM-VR7	367	4	YP_004063879.1	45	35
18	Enterobacteria phage IME08_02	377	4	YP_003734323.1	54	33
19	Enterobacteria phage RB32	376	4	ABI95002.1	89	34
20	Enterobacteria phage RB14	376	4	YP_002854513.1	86	35
21	Enterobacteria phage RB51	376	4	YP_002854135.1	91	36
22	Shigella phage Shf12	376	4	YP_004415072.1	89	34
Four-domain proteins with C-terminal elongation of the third domain						
23	Enterobacteria phage RB49	404	4	AAQ15404.1	35	100
24	Enterobacteria phage Phi1	404	4	YP_001469514.1	36	97
25	Enterobacteria phage JSE	404	4	YP_002922245.1	36	93
Five-Domain Proteins						
26	Enterobacteria phage JS10_02	469	5	YP_002922527.1	50	34
27	Enterobacteria phage RB30	472	5	AAM52483.1	77	34
28	Enterobacteria phage AR1	474	5	BAI83192.1	77	37
29	Enterobacteria phage RB69	471	5	AAP76093.1	53	37
30	Enterobacteria phage wV7	474	5	AEM00840.1	77	37
31	Enterobacteria phage ime09	472	5	AEK12435.1	76	34

From the viewpoint of modern classification, the following remarks can be made. Phages assigned to the T4 group of the subfamily *Teequatrovirinae* have three-domain, four-domain and five-domain variants of these proteins. Phages of the RB43 group have one-domain variants of this protein. The only exception is bacteriophage KP15. Its genome encodes both the one-domain and two-domain variants. Phages of the RB49 group carry only four-domain variants with the C-terminal tail of the third domain. Phages of the 44RR2.8 group carry two-domain variants of Hoc protein. Phages of bacteria of the genera *Aeromonas* and *Acinetobacter* have one-domain, two-domain and three-domain variants. *E. coli*-infecting bacteriophages carry all variants of Hoc proteins except two-domain variants. Phages of *Shigella* bacteria have a three-domain and four-domain variants.

A number of bacteriophage genomes encode two Hoc proteins. These are phages JS10 and IME08, whose genomes encode the three- and four-domain variants of Hoc protein, as well as phage KP15, whose genome encodes the one-domain and two-domain variants of this protein.

Hoc proteins of bacteriophages T4 and RB49 are the most investigated of the set. Hoc protein of bacteriophage T4 has been intensively studied biochemically and structurally [16-18, 24]. For Hoc protein of bacteriophage RB49, the spatial structures of the first three domains have been determined [9]. These two proteins are identical only by 20%. They are at the different ends of the homogeneity spectrum of Hoc proteins' primary structures. For this reason, it was of interest to search the protein database for the similarity with amino acid sequences of particular domains in namely these proteins. A comparative study of amino acid sequences in the domains of Hoc proteins known to date can clarify the evolutionary pathways for the origin of the diversity of these proteins.

2. Comparative study of amino acid sequences in particular domains of T4 Hoc protein

We applied the following approach for a detailed investigation of the diversity of Hoc protein and the evolution of its sequences. An earlier work [4] has shown this protein to consist of four like sequences 94 a.a. each. These segments are its domains. We performed a position-specific comparison separately of each domain with the GenBank protein database. The results yielded by the PSI-BLAST program are given in ADDITIONAL MATERIALS.

2.1. Analysis of the first domain

The first domain of Hoc protein in phage T4 has the highest degree of identity with domains of Hoc proteins in related bacteriophages RB32, RB14, RB30, RB51, JS10, RB69, JS98. The best identity is characteristic of phage RB30 (95 % of identical positions), as well as of phages RB32, RB51 (88 %). Substitutions in the alignment preserve the chemical nature (polarity, hydrophobicity) of substituted amino acids. The second iteration of the position-specific search in the database yielded the similarity in domains of Hoc protein of bacteriophage T4 and bacteriophages Ae25, 44RR2, Ae31, JSE, RB49 and Phi1. We should note the occurrence of conserved subsequences FTA (22–24 a.a.), TY**h**W**xx**D (36–42 a.a.), where **h** is hydrophobic amino acid and **x** any amino acid. At the third stage of comparison, mammalian cytokin receptor proteins appear (E = 10–11 to 10–12): cytokine receptor-like factor 1 precursor (*Mus musculus*), cytokine receptor-like factor 1 (*Rattus norvegicus*), as well as cytokine receptor-like factor 1 (*Homo sapiens* and *Macaca mulatta*). The biological effect of cytokines is known to be implemented by specific interaction with the cell receptor. Many bacteriophages are found in animal blood; with the blood flow they get into tissues and organs. This fact can indicate that blood is a normal medium for a virus to exist in. Presumably, bacteriophages in the animal organism could be selected for immunoglobulin-like domains that provided for the phage to reside in the blood until it reached the host bacterium. In this connection, we admit such an evolutionary pathway for the development of Hoc protein as its molecular mimicry for components of the immune system. It should be noted that a group of Polish researchers points to the possible eukaryotic origin of major antigen protein of bacteriophage T4 [7].

The fourth stage of comparison added 15 more cytokine receptors and 2 proteins with immunoglobulin-like domains to a list of similar protein sequences. At the fifth stage, the following interesting objects were retrieved: the major tail subunit Vibrio phage VP5 and the tail component encoded by prophage CP-933P in *E. coli*. The prophage-encoded tail components and the major tail subunit of phage VP5 are similar with various segments of the first domains in Hoc protein of bacteriophage T4. The first domain exhibits some similarity with capsid protein of bacteriophage VP2 of the family *Podoviridae*, as well as a similarity with the tail subunit of phage VP5 in the C-terminal region of this protein.

2.2. Analysis of the second domain

The first stage of comparison expectedly revealed Hoc proteins of seven related bacteriophages. The high degree of identity (~80% for RB32, RB30, RB51, RB14, and 57% for RB69) is indicative of the evolutionary proximity of these bacteriophages and their major antigens. In all proteins, the first six amino acids coincide. Our attention was attracted by the sequence 169–176 a.a. of phage T4 (TDYDALS). In Hoc proteins of phages RB51 and RB32, this segment is substituted by another conserved site (ENYNEKE). This fact indicates the possible location of Hoc proteins of these two related bacteriophages on the parallel branches of the evolutionary development of the protein. The second domain of Hoc protein of phage T4 reveals a high similarity (E ~ 10–17) with two segments 94 amino acids long in the Hoc protein sequence of bacteriophage RB30, which are the second and third domains of this protein (Fig. 1).

```
T4       1   TLAVTPASPAAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVGGETNSTFSYTPTTSG  60
             TLAVTPASPAAGVIGT V+FTAALASQP GASATYQWYVDDS V   T++TF+YTP TSG
RB30    98   TLAVTPASPAAGVIGTAVEFTAALASQPSGASATYQWYVDDSPVSEATSATFNYTPDTSG  157

T4      61    VKRIKCVAQVTATDYDALSVTSNEVSLTVNKKTM  94
              VK+IKC AQVTAT+YDALSVTSNEVSLTVNKKT
RB30   158    VKKIKCTAQVTATNYDALSVTSNEVSLTVNKKTQ  191
```

a)

```
T4      1    TLAVTPASPAAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVGGETNSTFSYTPTTSG  60
             TLAVTPASP+AGVIGTPVQFTAALASQPDGASATYQWYVDDSQV GETNSTF+YTPTT+G
RB30   194   TLAVTPASPSAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVSGETNSTFNYTPTTNG  253

T4      61    VKRIKCVAQVTATDYDALSVTSNEVSLTVNKKTM  94
              VKRIKCVAQVTA DY+A  VTSNEVSLTVNKKTM
RB30   254    VKRIKCVAQVTADDYNAKEVTSNEVSLTVNKKTM  287
```

b)

Fig. 1. a) A sequence alignment fragment of Hoc proteins in phage T4 (second domain) and phage RB30 (second domain); b) a sequence alignment fragment of Hoc proteins in phage T4 (second domain) and phage RB30 (third domain), amino acid sequences of RB30 protein that coincide with T4 Hoc are shown by blue letters. Grey highlights are motifs characteristic of major antigen protein.

This fact gives grounds to believe that evolutionally protein of phage RB30 became five-domain protein as the result of the duplication of the *hoc* gene segment. The protein acquired an additional immunoglobulin-like domain, by more than 90 % similar to the second domain in Hoc protein of phageT4. Based on this, it could be assumed that the attachment of protein to the capsid surface of bacteriophage T4 is done not owing to the first three N-terminal domains as it was assumed earlier [4], but owing to the C-terminal domain. The N-terminal domains of Hoc – three in T4 and four in RB30 – in this case can be exposed on the surface of the capsid and provide a high antigenicity of Hoc protein.

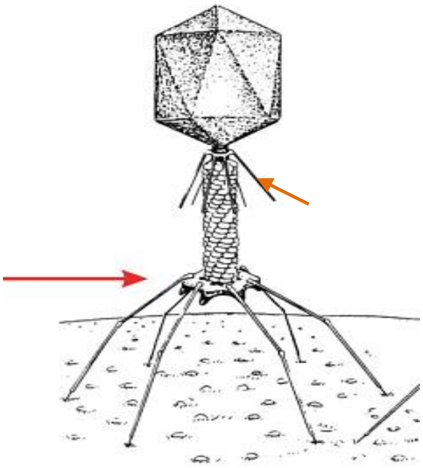

Fig. 2. Bacteriophage T4. The red arrow points to tail fibrillae; the orange arrow, to the phage collar with whiskers formed by Wac proteins.

At the second stage of comparison, we found (E < 0.005) Hoc proteins of bacteriophages of *Aeromonas* (25, 44RR2.8t, 31), *E. coli* (JS98 and JS10); in the additional list (at E > 0.005) proteins of *E. coli* bacteriophages (Phi1, RB49 and JSE). Besides, the second domain of T4 Hoc exhibits a similarity with a tail component of various bacteriophages (e.g., of phage BP-4795 of *E. coli* O157:H7, cryptic prophage CP-933X of *E. coli* O157:H7 EDL933), as well as of other prophages of various *E. coli* strains. We should also note a similarity of T4 Hoc protein and phage Vibrio VP2 capsid protein. In the course of studying the features of the second domain, at the third stage of comparison we found its similarity with phage-RB43 Wac protein, which is a constituent of the phage collar. We assume that these protein sequences can be exposed on the surface of collar whiskers (Fig. 2). It should be noted that Hoc protein of bacteriophage RB43 consists of only one domain homologous to the fourth domain of T4 Hoc. Possibly, the occurrence of sequences, similar to the second domain of Hoc, in Wac protein of RB43 compensates for their absence in Hoc protein of this phage. At the third, fourth and fifth stages of comparison, we found many major tail proteins in prophages of various *E. coli* strains, various variants of the leukocyte immune-type receptor and immunoglobulin I-set domain-containing proteins, as well as IgG Fc receptors.

2.3. Analysis of the third domain

At the first stage of comparison, we found proteins of bacteriophages RB32, RB14, RB51, RB30, JS10, RB69, JS98, JSE. Hoc proteins of these phages exhibit various degrees of similarity; still, all of them preserve the following motif in their sequence: YSWKKD**x**S (Fig. 3).

```
T4   189  KTMNP-QVTLTPPSINVQQDASATFTANVTGAPEEAQITYSWKKDSSPVEGSTNVYTVDT  247
          KTM+   VTLTP SINV +   A+F A V GAP  A   YSWKKD SPV G+T+   +DT
JS98 93   KTMSGVSVTLTPESINVPEGTPASFKATVAGAPAGATFAYSWKKDGSPVVGTTDTLAIDT  152
```

Fig. 3. A motif (highlighted in grey) occurring in the third domain of T4 Hoc; the figure presents an example of a sequence alignment fragment of the third domain in T4 Hoc and JS98 Hoc.

Besides, Hoc proteins of bacteriophages RB32, RB14, RB51, RB30 totally coincide with respect to the first 29 amino acids. The second, third and fourth domains of Hoc protein of phage JS10 exhibit a similarity with the third domain of T4 Hoc. This homology of the domains, as well as the similarity of the second and third domains of RB30 Hoc protein can be indicative of their origin owing to intragenic duplication. In contrast with JS10, the three-domain Hoc protein of JS98 exhibits a similarity with the third domain of T4 Hoc only in the region of its second domain (93–187 a.a.). The fourth domain of RB30 Hoc is identical to the third domain of the investigated protein of T4 by 86 %, which points to the evident evolutionary affinity of proteins of these bacteriophages and the origin of RB30 Hoc protein by duplication of a *hoc* gene segment of phage T4 (Fig. 4).

```
T4     188   KKTMNPQVTLTPPSINVQQDASATFTANVTGAPEEAQITYSWKKDSSPVEGSTNVYTVDT   247
             KKTMNPQVTLTPPSINVQQDASATFTANVT APEEAQI YSWKKDSSPVEGSTNVYTVDT
RB32   188   KKTMNPQVTLTPPSINVQQDASATFTANVTDAPEEAQIEYSWKKDSSPVEGSTNVYTVDT   247
                                           a)
T4     188   KKTMNPQVTLTPPSINVQQDASATFTANVTGAPEEAQITYSWKKDSSPVEGSTNVYTVDT   247
             KKTMNPQVTLTPPSINVQQDASATFTANVT APEEAQITYSWK+DSSPVEGSTNVYTVDT
RB30   284   KKTMNPQVTLTPPSINVQQDASATFTANVTDAPEEAQITYSWKRDSSPVEGSTNVYTVDT   343
                                           b)
```

Fig. 4. **a**) A sequence alignment fragment of Hoc protein in phage T4. The similarity pattern (shown not in full) is the same for RB32, RB14, RB51; **b**) for Hoc protein of phage RB30 the similarity pattern (shown not in full) is alike but belongs to the amino acid sequence 284–378, a segment of the fourth domain.

At the second stage of comparison, in the additional list we found the major tail protein of phage VP5. The similarity of the latter with the Hoc sequence can be considered to be significant, if we take into account that the alignment of sequences is given without gaps, and many differing amino acids are similar by their chemical characteristics to those in the sequence of T4 Hoc. Besides, the FTA consensus (22–24 a.a.), occurring in many major antigenic proteins of T4-type bacteriophages, is preserved (Fig. 5). Possibly, tail proteins containing such domains possess, as Hoc, an increased antigenicity.

```
T4    193   PQVTLTPPSINVQQDASATFTANVTGAPEEAQITYSWKKDSSPVEGSTNVYTVDTSSVG   251
            P V+++P S +  + A  TFTA+V        +T  W  + + V+     YT   ++VG
VP5   393   PTVSISPVSASPLEPAPVTFTASVVDDGGAPPVTLKWYLNGNLVQNGGTTYTSPPTAVGQ   452
```

Fig. 5. A sequence alignment fragment of Hoc protein in phage T4 and of tail protein in phage VP5; grey highlights are motifs; yellow highlights, like amino acid substitutions.

Besides, Hoc capsid protein of phage Phil was found; its first domain has a similar segment (55 a.a.) carrying a conserved sequence TY**h**W**x**KD**xx**P (226–235 a.a.), where **h** is hydrophilic amino acid. A similar segment of Hoc capsid protein of phage RB49 has a 37 % identity to its homolog from phage T4 (Fig. 6). The length of the similar region is 53 a.a.

```
T4      211   TFTANVTG-APEEAQITYSWKKDSSPVEGSTNVYTV-DTSSVGSQTIEVTATVTAA   264
              T TA V G  P  + +TY+W KD  P E +T   TV D +S  + + +VT   T
Rb49    21    TLTATVAGDEPLPSNLTYTWTKDDQPHENNTATLTVADATSENAGSYKVTVQDTDT    76
```

Fig. 6. A sequence alignment fragment of Hoc proteins in phages T4 and RB49.

At the third stage of comparison, we found the sequence of phage T5 tail protein, in which we should note the presence of the consensus sequence TY**h**W**x**KD**x**SP**h** (226–236 a.a.) slightly differing from that earlier indicated for T4-type phages. At this stage of comparison, tail proteins of the following bacteriophages are found: YYZ-2008, BP-4795, prophage CP-933N; the above mentioned motif is also present in them.

At the fourth stage of comparison, the program revealed numerous proteins belonging to the superfamily of immunoglobulins or protein fragments containing immunoglobulin-like domains. This similarity of the investigated protein and these domains can be explained by the similarity of protein folding, because it is such an immunoglobulin-like folding that is energetically advantageous and/or evolutionally ancient. Molecular mimicry of Hoc protein for immune system components to avoid elimination in the animal organism can also be assumed. However, an opposite version of the possible inhibition and overcoming of the system of animal immune response owing to the molecular similarity with cytokines, selectins, fragments of immunoglobulin receptors, is also probable. Thus, at this stage of comparison, a similarity of the third domain of T4 Hoc with IgG Fc receptor in the region of the N-terminal (344 a.a.) is found.

Also in the fourth stage, we revealed a similarity with the sequence of tail protein of *E. coli* O157:H7 str. Sakai prophage (E = 0.045); similar segments of proteins are presented without gaps, there is the Y**h**WKKD motif (Fig. 7).

```
T4    205 QQDASATFTANVTGAPEEAQITYSWKKDSSPVEGSTNVYTVDTSSVGSQTIEVTATVTAA 264
             + A  T +V+         Y+WKKD  PV+G T + + + T VT +
Sakai 158 TVNTGALLTMSVSANGGTPPYKYAWKKDGQPVDGQTTDTFSKPGAQSADAGKYTCVVTDS 217
```

Fig. 7. A sequence alignment fragment of Gp hoc in phage T4 and of tail protein in prophage *E. coli* O157:H7 str. Sakai, grey highlights are Y**h**WKKD motifs.

The concluding fifth stage of comparison revealed no new protein sequences, but in the list of similar proteins the number of immunoglobulin-like domains, as well as proteins with an insignificant similarity of sequences (titin, paladin, obscurin, fibronectin, etc.), increased.

2.4. Analysis of the fourth domain

At the first stage of comparison, we found a significant similarity of the fourth domain of T4 Hoc protein with Hoc proteins of 14 related T4-type bacteriophages: RB69, RB14, RB30, RB32, RB51, JS10, JS98, Phi1, JSE, RB49, Ae31, 44RR2.8, Ae 25, RB43. The highest identity (96 % and more) was found in Hoc proteins of bacteriophages RB69, RB14, RB30, RB32, RB51. Earlier, we pointed out the possible fact of intragenic duplication in the *hoc* gene of phage RB30, owing to which this protein acquired an additional domain. The program showed an evident similarity of the fourth domain of Hoc protein in phage T4 and the fifth domain of this protein in phage RB30 (98 % identity). Our interest was also drawn to the major antigen protein of phage JS10. Its amino acid sequence exhibits a similarity with the fourth domain of T4 Hoc also in the region of the fifth domain (87 % identity; herewith, the substituting amino acids in the main preserve their chemical characteristics, e.g., E→D, K→R etc.). An analogous pattern of similarity (85 % identity) is presented for Hoc of phage JS98; however, in it the third domain is similar to the sequence of the fourth domain. Indeed, in contrast with five-domain Hoc proteins of phages RB30 and JS10, JS98 Hoc is three-domain protein. Hoc proteins of other bacteriophages (Phi1, JSE, RB49, Ae31, 44RR2.8, Ae 25, RB43) exhibit a much lower degree of similarity with the investigated protein.

All above mentioned major antigen proteins feature a conserved amino acid segment, ESRNG (355–359 a.a.) (Table 2), which provides for the attachment to the capsid [24].

It should also be noted that at the first stage of comparison, besides the evident similarity of bacteriophages' capsid proteins, we revealed some similarity of the fourth domain with immune system proteins – E-selectins (*Equus caballus*), P-selectin precursor (*Salmo salar*). The aligned amino acid sequences are presented with a large number of gaps. However, it is noteworthy that the fourth (C-terminal) domain of Hoc is similar to the N-terminal sequence of selectins. It is well known that selectins are adhesive molecules, N-terminus of which (lectin-like domain) provides for the adhesion of leukocytes to endothelial cells. Thus, the similarity of the C-terminal domain of Hoc with them indirectly indicates the involvement of namely this domain in the attachment to phage's capsid surface.

At the second stage of comparison, among sequences similar to the fourth domain we had protease inhibitor proteins of T4-type bacteriophages (RB43, Ae25, Ae 44RR2.8, JS10, JS98, RB32, RB14, RB30, RB51, RB69, JSE, Phi1, RB49). The extent of similarity of these proteins with the investigated sequence is not high; still, we should note the occurrence of a consensus sequence that had been revealed at the first stage of comparison. In the case of protease inhibitors, it looks like this: **hh**ESRN (353–358 a.a.). Probably, this sequence, as in Hoc proteins, can provide for the attachment to the capsid surface.

The third, fourth and fifth stages of comparison revealed no additional protein sequences similar to the fourth domain of T4 Hoc protein.

Table 2. Multiple alignment of amino acid sequences of Hoc protein C-terminal fragments by CLUSTALX software. The conserved motifs are highlighted in red. The first column gives phage names; the second column, the coordinates of conserved elements with respect to the beginning of fragments

Phage name	Position number	Multiple-alignment fragment
RB49	347	WRDREVYSTSK-YAKDLETIAAAEEKYSDCTCMESRNGFMYQSKELHKLDRETLERVLR
Phi1	347	WRDREVYSTSK-YAKDLETIAAAEEKYSDCTCMESRNGFMYHSKELHKLDRETLERVLR
44RR2.8t	135	WRDDPVNS--P-WPKVTYAIDKAVTDYGDCLMQESRNGYIYKASQFVKS---------
Ae31	135	WRDDPVNS--P-WPKVTYAIDKAVTDYGDCLMQESRNGYIYKASQFVKS--------
RB32	326	WKTDDPDS--K-YYLHRYTLQKMMKDYPEVDVQESRNGYIIHKTALETGIIYTYP---
T4	326	WKTDDPDS--K-YYLHRYTLQKMMKDYPEVDVQESRNGYIIHKTALETGIIYTYP---
JS98	230	WKTDDPDS--P-YYLHRYTLQKMITDYPEVDVQESRNGRIIHRTALEAGIIYDYVY--
RB69	420	WKTEDPDS--K-YYLHRYTLQKMMKDYPEVDVQESRNGYIIHKTALETGIIYTYP--
RB30	421	WKTDDPDS--K-YYLHRYTLQKMMKDYPEVDVQESRNGYIIHKTALETGIYYTYP
Ae25	133	WRQTPAES--P-WPIITFSIDRAVEEYGECLMQESRNGYIYKASAYTK---------
RB43	40	WRDYAEEK----YAKEFVTLKQAYID

3. Comparative study of amino acid sequences of particular domains in Hoc protein of bacteriophage RB49

Hoc protein of bacteriophage RB49 strongly differs from that of phage T4: it is identical by 35 % in the zone of the best alignment and by 20 % along the entire length. For this reason, it was of interest to compare the amino acid sequences of its domains with the protein sequence databases similar to the comparison of T4 Hoc protein domains. The first three domains of Hoc proteins of bacteriophage T4 were crystallized, and their structure was resolved by X-ray diffraction analysis [10]. In the course of the analysis, the first and second domains were found to contain, respectively, 90 and 91 amino acids, and the third domain was much longer, 123 a.a. All the three domains were shown to form a linear structure. It proved impossible to crystallize Hoc protein of bacteriophage RB49 as a whole, because there was a flexible joint between the third and fourth domains.

3.1. Analysis of the first domain

At the first stage of comparison, we found Hoc proteins of Rb49-related bacteriophages Phi1 and JSE with a high percentage of identity, 98 % and 93 %, respectively. Hoc protein of bacteriophage RB69 possessed a lower identity (36 %). Tail proteins in *E. coli* bacteriophages EcoS-CEV2 and T5C possessed a greater percentage of maximal identity (44 % and 41 %, respectively). Also, we found the assumed tail proteins of *E. coli* prophages ED1a, E110019, DEC9D; *Salmonella enterica* and *Shigella boydii* Sb227; similar proteins of prophages in other *E. coli* strains; the assumed capsid protein of enterobacterial bacteriophage pb10 H8; Wac proteins of bacteriophages RB43 and RB16; proteins of *Opitutus terrae*, *Desulfitobacterium hafniense*, *E. coli* STEC_B2F1, *E. coli* OK1180 and of the bacterium Ellin514 that contain immunoglobulin I-set domain-containing protein, CD22 *Cricetulus griseus* B-cell receptor, titin protein of *Alligator sinensis* and *Crocodylus siamensis*.

At the second stage of comparison, we additionally found tail proteins of enterobacterial phages EPS7 and Felix 01, of phage Yersinia PY54, Vibrio phages pVp-1, VP5 and phage BP-4795, Wac protein of *Klebsiella* phage KP15, *Klebsiella pneumonia* bacterial surface protein containing Ig-like domains.

The third, fourth and fifth stages of comparison additionally yielded a large number of Fc-fragment receptors of antibodies from various eukaryotic organisms. These findings, as in the case of Hoc protein of phage T4, can be indicative of molecular mimicry of bacteriophages for immune system components. The first domain of Hoc protein of bacteriophage RB49 exhibits a similarity with many phage tail proteins. Probably, as Hoc, these proteins carry immunoglobulin-like domains. A similarity with such proteins was also found in the analysis of the first three domains of T4 Hoc protein.

3.2. Analysis of the second and third domains

When comparing the second and third domains with the database of the known protein sequences in all five iterations of the PSI-BLAST algorithm in the credibility region ($E < 0.0001$) of the results, we found only a similarity with the sequences of corresponding domains in Hoc proteins of closely related phages Phi1 and JSE.

The absence of similarity in the second and third domains of RB49 Hoc protein with immunoglobulins and immunoglobulin-like proteins was an unexpected result. It is indicative of a significant difference in the antigenic and other properties of Hoc protein of phage RB49 (as well as Phi1 and JSE) from the corresponding protein of phage T4 and a number of other related bacteriophages.

3.3. Analysis of the fourth domain

At the first stage of comparison, we found a similarity with Hoc proteins of enterobacterial bacteriophages Phi1, JSE, RB16, CC31, RB43, wV7, AR1, T4, ime09, RB51, RB30, RB32, vB_EcoM-VR7, T4T, RB14, JS10, IME08, JS98, Bp7, as well as of bacteriophages of the genera *Acinetobacter* (Acj61 and Acj9), *Aeromonas* (65, phiAS4, 44RR2.8t, 31), *Shigella* (SP18 and Shfl2) and *Klebsiella* (KP15).

At the second stage of comparison, Hoc proteins of phage Aeromonas 25 and the second variant of Hoc protein of enterobacteriophage JS10 were found.

At the third stage of comparison, additionally in the credibility region of comparison we found Inh protein, protease inhibitor of bacteriophage RB43.

At the fourth stage of comparison, we found 32 prohead protease inhibitors of enterobacterial bacteriophages RB16, CC31, Bp7, JS10, wV7, JS98, IME08, RB32, ime09, T4T, RB14, AR1, RB51, T4, RB30, vB_EcoM-VR7, RB69, Phi1, JSE, RB49; *Acinetobacter* phages Acj9, 133, Acj61, Ac42; *Aeromonas* phages 44RR2.8t, phiAS4, 65, 25, phiAS5; *Shigella* bacteriophages Shfl2 and SP18; *Klebsiella* phage KP15.

At the fifth stage of comparison, we found two more Inh proteins of *Aeromonas* bacteriophages Aeh1 and PX29 and immunoglobulin heavy chains of various animals. Inh protein – a prohead protease inhibitor in *Teequatrovirinae* bacteriophages – has a C-terminal domain similar to the C-terminal domain of Hoc protein. What is more, close to the C-terminal this protein contains the subsequence ESRN similar to the sequence ESRNG, which in Hoc protein has been shown to be involved in the interaction with the phage surface [24]. The C-terminal domain of Hoc proteins is responsible for the attachment to the bacteriophage head. Probably, the prohead protease inhibitor, Inh protein, owing to its C-terminal domain can temporarily attach to the head, thus inhibiting the protease involved in its maturation.

An interesting result is the similarity of the last domain of RB49 Hoc protein with immunoglobulin heavy chains. Earlier, the fourth domain of Hoc was considered to have no similarity with immunoglobulins [24]; however, a similarity was found to exist, which is consistent with the data by the group of C. Chotia [4].

Comparison of particular domains of bacteriophage RB49 with the protein sequence database differs in a substantial way from such an analysis of particular domains of phage T4. All domains of bacteriophage T4 exhibited a similarity with Hoc proteins of many related bacteriophages. In bacteriophage RB49 for the first three domains the PSI-BLAST program revealed a similarity only with Hoc proteins of closely related bacteriophages Phi1 and JSE assigned to the same RB49 group of phages. The only exception is the similarity of the first domain with the first domain of Hoc protein of bacteriophage RB69 belonging to the T4 group. The similarity of Hoc protein of bacteriophage RB49 with Hoc proteins of other bacteriophages is limited to the fourth domain responsible for the attachment to the capsid surface. Therefore, an antigen differing from other bacteriophages is exposed on the surface of capsids of bacteriophages RB49, Phi1 and JSE.

4. The phylogenetic tree of particular domains in Hoc proteins of the *Teequatrovirinae* subfamily bacteriophages

Within the framework of the analysis of the role of particular domains in providing for the structure and function of Hoc proteins, we constructed a tree of the similarity of particular domains in *hoc* gene product proteins of T4-related bacteriophages. The tree is shown in Figure 8.

Hoc proteins of various bacteriophages are designated after bacteriophages themselves; the abbreviations are as follows: Ae31, *Aeromonas* phage 31; 44RR, *Aeromonas* phage 44RR2.8t; phi AS4, *Aeromonas* phage phi AS4; JS98, enterobacterial phage JS98; VR7, enterobacterial phage vB_EcoM-VR7; RB30, enterobacterial phage RB30; RB69, enterobacterial phage RB69; RB49, enterobacterial phage RB49; AR1, enterobacterial phage AR1; RB16, enterobacterial phage RB16; RB43, enterobacterial phage RB43; Acj9, *Acinetobacter* phage Acj9; Acj61, *Acinetobacter* phage Acj61; T4, enterobacterial phage T4; Ae25, *Aeromonas* phage 25; Ae65, *Aeromonas* phage 65; Phi1, enterobacterial phage Phi1; JSE, enterobacterial phage JSE; RB32, enterobacterial phage RB32; RB14, enterobacterial phage RB14; RB51, enterobacterial phage RB51; wV7, enterobacterial phage wV7; ime09, enterobacterial phage ime09. Some phages have two Hoc proteins; they are designated by numbers after the abbreviated phage name: IME01, enterobacterial phage IME08 (the three-domain variant of IME08 protein); IME02, enterobacterial phage IME08 (the four-domain variant of IME08 protein); JS1001, enterobacterial phage JS10 (the three-domain variant of JS10 protein); JS1002, enterobacterial phage JS10 (the four-domain variant of JS10 protein); KP1501, *Klebsiella* phage KP15 (the one-domain variant of KP15 protein); KP1502, *Klebsiella* phage KP15 (the two-domain variant of KP15 protein). The number after the abbreviated phage name designates the domain number counting from the N-terminus of Hoc protein.

The tree has three branches: the branch of C-terminal domains, the branch of N-terminal domains and the branch of intermediate domains. This corresponds to the results of modelling the structure of T4 Hoc protein, which have shown that the first and last domains are brought together, organize the protein structure and are involved in the attachment of protein to the phage surface [24].

The role of the intermediate domains is different – they are exposed on the phage surface and, most likely, are responsible for the interaction with antibodies and *E. coli* cell surface [9]. The results of sequence analysis for particular domains indirectly support these conclusions.

It can be noted that one-domain proteins of phages RB43, RB16 and KP1501 are assembled into one subbranch inside the branch of C-terminal domains and are at close distances one from another. They form the common branch with C-terminal domains of two-domain proteins of *Aeromonas* phages 44RR2.8t, 31, 25 and phi AS4. This can indicate that they have evolved in parallel owing to deletions occurring in genes of longer precursor proteins. At the same time, one-domain protein of phage Acj61 exhibits a greater similarity with C-terminal domains of proteins similar to T4 Hoc. One-domain Hoc proteins are separately existing C-terminal domains. This implies the prime significance of namely C-terminal domains for functioning. Also, it can be assumed that C-terminal domains are evolutionally more ancient than other domains of Hoc proteins.

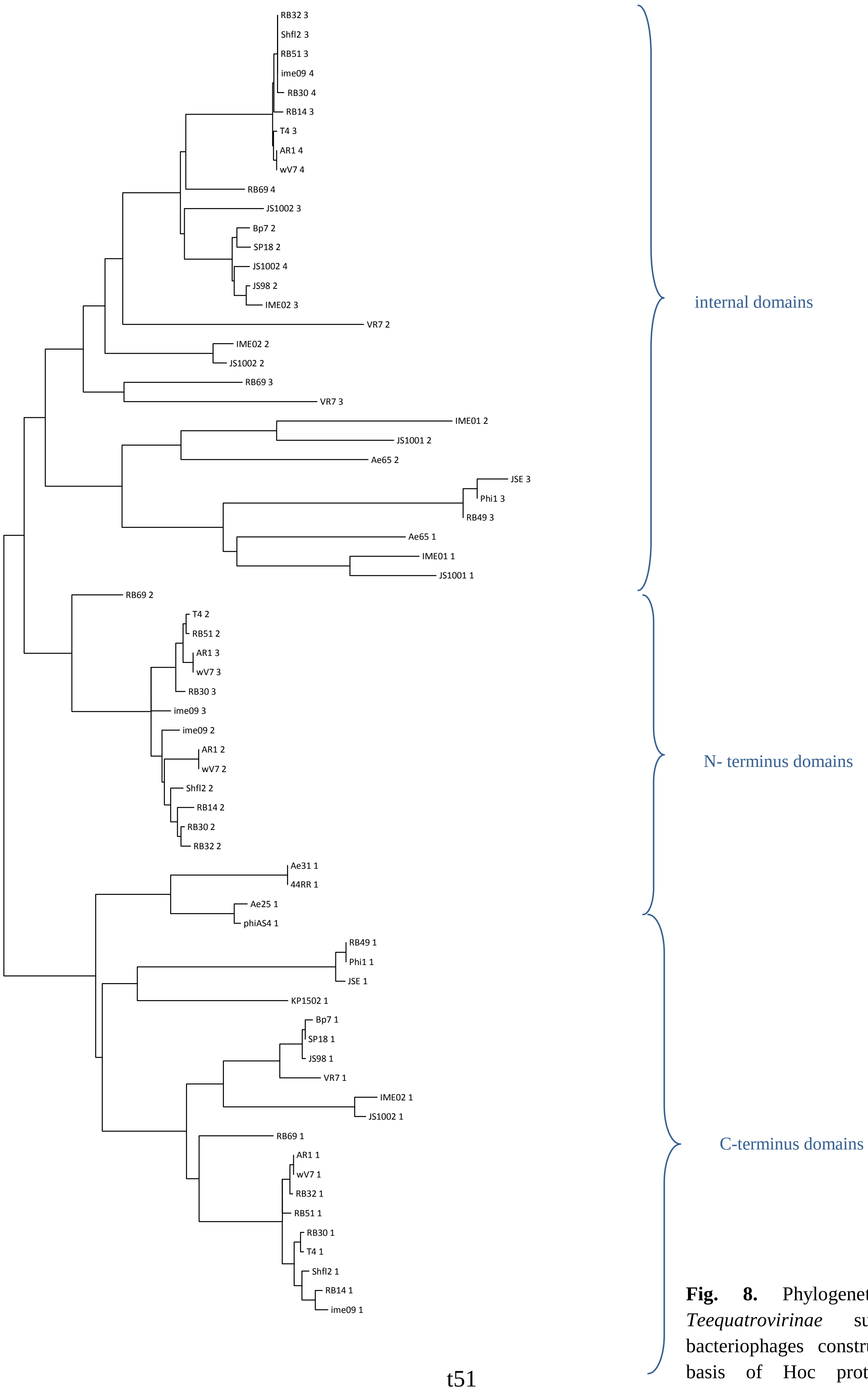

Fig. 8. Phylogenetic tree of *Teequatrovirinae* subfamily of bacteriophages constructed on the basis of Hoc protein domains similarity.

Possibly, they were precursors of other domains and evolved in the direction of refining the property of attachment to phage’s capsid.

On the branch of C-terminal domains, there is an anomalous unique subbranch of second domains of Hoc proteins of bacteriophages RB49, Phi1, JSE. Probably, the similarity of these domains with C-terminal domains is related to the origin of the second domain in proteins of bacteriophages RB49, Phi1, JSE from the C-terminal domain.

On the branch of the intermediate domains, there is also a unique subbranch of the third domains of Hoc proteins of bacteriophages RB49, Phi1, JSE and the N-terminal domains of three-domain proteins of phages Ae65, IME08, JS10. This can be indicative of the evolutionary relation of the third domain in Hoc protein of bacteriophages RB49, Phi1, JSE and the N-terminal domain.

The second and third domains of five-domain proteins of bacteriophages RB30, AR1, wV7, ime09 are on one subbranch of the common tree, which can be indicative of their origin due to intragenic duplication. On the same subbranch, there are the second domains of four-domain Hoc proteins of bacteriophage T4 and related bacteriophages RB14, RB32, RB51. Probably, these domains could be duplicated in the course of evolution, which led to the emergence of the second and third domains of five-domain proteins. The second and third domains of phage RB30 differ only by 17 out of 94 a.a. (Fig. 9), the second and third domains of phages AR1 and ime09 differ by 25 out of 94 a.a. (Figures 1 and 2 in Supplementary Materials), and the corresponding domains in Hoc protein of bacteriophage wV7 differ by 24 amino acids (Fig. 3 in Supplementary Materials). The amino acid substitutions differ in different phages, which indicates an independent evolution of duplicated domains.

```
RB30_3    191   QTTTLAVTPASPSAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVSGETNSTFNYTPT
RB30_2     95   QTTTLAVTPASPAAGVIGTAVEFTAALASQPSGASATYQWYVDDSPVSEATSATFNYTPD
T4_2       95   QTTTLAVTPASPAAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVGGETNSTFSYTPT
                ************:******.*:*********.************ *.   *.:**.***

RB30_3    251   TNGVKRIKCVAQVTADDYNAKEVTSNEVSLTVNKKT
RB30_2    155   TSGVKKIKCTAQVTATNYDALSVTSNEVSLTVNKKT
T4_2      155   TSGVKRIKCVAQVTATDYDALSVTSNEVSLTVNKKT
```

```
RB30_3    191   QTTTLAVTPASPSAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVSGETNSTFNYTPT
RB30_2     95   QTTTLAVTPASPAAGVIGTAVEFTAALASQPSGASATYQWYVDDSPVSEATSATFNYTPD
T4_2       95   QTTTLAVTPASPAAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVGGETNSTFSYTPT
                ************:******.*:*********.************ *.   *.:**.***

RB30_3    251   TNGVKRIKCVAQVTADDYNAKEVTSNEVSLTVNKKT
RB30_2    155   TSGVKKIKCTAQVTATNYDALSVTSNEVSLTVNKKT
T4_2      155   TSGVKRIKCVAQVTATDYDALSVTSNEVSLTVNKKT
                *.***:***.***** :*:* .**************
```

Fig. 9. Multiple sequence alignment of the second and third domains of Hoc protein in phage RB30 and of the second domain of Hoc protein in phage T4.

The second and third domains in five-domain Hoc protein of bacteriophage RB69 are identical by 40 % (38 out of 96 amino acids coincide) (Fig. 4 in Supplementary Materials). They could also occur due to intragenic duplication and, judging by the number of distinctions, this is the result of a more ancient evolutionary process.

It can be concluded that, first, five-domain variants are formed from four-domain variants by duplication of genetic material in the segment that encodes the second domain of the four-domain variant. As the result, two domains are formed that are similar with each other and with the parent domain. Second, the duplication in these phages proceeded at different times: in phage RB30, comparatively recently; in phage AR1 (Fig. 1 in Supplementary Materials), slightly earlier; and in phage RB69, considerably earlier (Fig. 10, Fig. 4 in Supplementary Materials). Third, it could be assumed that if the duplication can occur at different times in the same site, then there is a genetic mechanism that generates duplications. Indeed, at a relatively small distance from this segment of T4 genome there is the site of nuclease SegE, which is the recombination hotspot [19]. Incorrect resolution of recombination structures frequently emerging in this region of the genome can serve as a mechanism of duplications to emerge. Intragenic duplication led to the formation of five-domain variants of the Hoc protein sequence (Fig. 2 in Supplementary Materials, Fig. 3 in Supplementary Materials).

```
RB69_2      94    ENNSTVAVTPASPAAVEIGTATTFTANVSNQPSGAAIAYTWKVDGVAVDGQKQSTFEYTP
T4_2        95    -QTTTLAVTPASPAAGVIGTPVQFTAALASQPDGASATYQWYVDDSQVGGETNSTFSYTP
RB69_3     190    ANSSTLKITPESPTT-VFGVPITLTANVSGAPSGATTSFQWSMDDSNILDATSATYKFTP
                   :.:*: :** **::  :*..  :** ::. *.**: :: * :*.  : . ..:*:.:**

RB69_2     154    TSEGTKSITCSVTVTATDYVDKTVESSAVSLTVNKK-
T4_2       154    TTSGVKRIKCVAQVTATDYDALSVTSNEVSLTVNKKT
RB69_3     259    TEVGSKTLKCTVSVSATNYVTKEISAEATVVTNNATF
                  *  * * :.* . *:**:*    : :. . :* * .
```

Fig. 10. Multiple sequence alignment of the second and third domains of Hoc protein of phage RB69 and of the second domain of Hoc protein of phage T4.

The second and third domains in the four-domain protein of phage vB_EcoM-VR7 are identical by 32 % (Fig. 11); however, the similarity between them is greater than with the first domain (26 %). This degree of similarity is sufficiently large to assume that they originated as the result of a duplication of the precursor domain sequence. Thus, we can also assume the origin of the four-domain vB_EcoM-VR7 from the three-domain variant by intragenic duplication.

Thus, the intermediate domains in proteins of all phages (except phages RB49, Phi1, JSE) are on a separate branch of the domain comparison tree, as we reported earlier, on one of the three main branches. This is indicative of a greater similarity of the intermediate domains between themselves than with the terminal domains, and of the possible involvement of intragenic duplications in their evolution.

5. Evolutionary schemes of the formation of Hoc protein amino-acid sequence variants

The three-domain variants of Hoc protein occur in bacteriophages of various bacterial genera – *E. coli*, *Aeromonas*, *and Shigella*. The three-domain variant of Hoc protein possesses all major features of the assumed structure. Its C-terminal domain is responsible for the attachment to the phage surface, the N-terminal domain interacts with the C-terminal domain and forms a loop structure [24], the intermediate domain performs the binding of antibodies and attachment to the cell surface [9]. As the three-domain variant possesses all major features of the structure, we assume that it is this variant that underlies Hoc proteins' diversity. Let us try to model the origin of the diversity of Hoc protein sequences from the three-domain variant.

```
VR7_2      94    MSATATLTTSTPTVKVGEEYNASADVTGEPGGATIAYLWS--TGEITKDITRTATVAGPV
VR7_3     189    ---VEISGPSTATVDV--PFNLTASVSPAIPGATLAYKWDDNSTEATRAIT--ESTEGLK
                    .    .**.**.*   :* :*.*:    ***:** *.  : * *: **   :. *

VR7_2     152    SLTCQITVSAADYDNQVINPEAVVVTVENN-TFPEFT----
VR7_3     142    SYTCEVTASQTGFTNSVKSGNKSVTVEEAEPVIPEECPLIY
                 * **::*.* :.: *.* . :  *.. * : .:**
```

Fig. 11. Pairwise sequence alignment of the second and third domains of Hoc protein in phage vB_EcoM-VR7.

A scheme of the evolution due to intragenic duplication could include two stages. First, the duplication of a gene part encoding the intermediate domain to form a four-domain variant from the three-domain variant; second, further duplication of one of the intermediate domains to form a five-domain variant from the four-domain variant (Fig. 12).

The driving force of evolution by increasing the number of intermediate domains could have been their ability to bind the surface of the bacterial cell, which can be a preliminary phage–bacterium interaction stage preceding the infection. It is also essential for changing the sequence to have a recombination hotspot at this site of the genome; probably, even more essential than the positive selection for variants with additional middle domains.

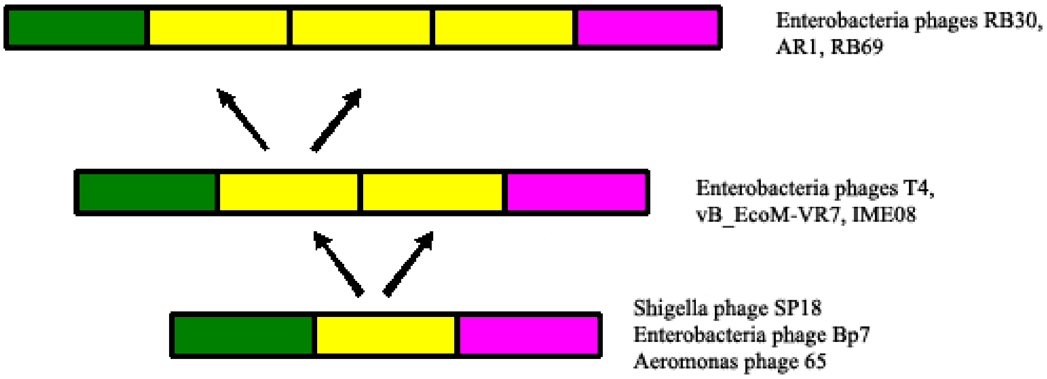
Enterobacteria phages RB30, AR1, RB69
Enterobacteria phages T4, vB_EcoM-VR7, IME08
Shigella phage SP18
Enterobacteria phage Bp7
Aeromonas phage 65

Fig. 12. Scheme of intragenic duplications in the evolution of the three-domain variant of Hoc protein into a five-domain variant. N-terminal domains are highlighted in green; intermediate domains, in yellow; C-terminal domains, in magenta. The duplication process is shown by arrows. Shown on the right-hand side are phages that carry Hoc protein variants corresponding to stages of the process.

Mutation of *hoc* gene has almost no effect on the harvest of the phage or its resistance to external impacts [17, 18]. On the other hand, the absence of this protein on the capsid surface leads to an increased aggregation of particles [29], which can affect the infectivity of phage populations. From this point of view, any of the Hoc protein variants, including the one-domain and two-domain variants, is more preferable than the complete absence of a domain.

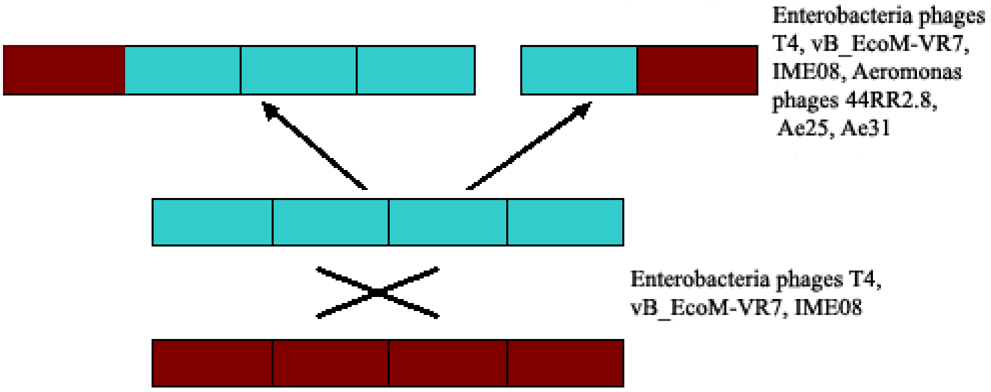
Enterobacteria phages T4, vB_EcoM-VR7, IME08, Aeromonas phages 44RR2.8, Ae25, Ae31
Enterobacteria phages T4, vB_EcoM-VR7, IME08

Fig. 13. Scheme of domain mixing in the recombination between *hoc* genes of various closely related bacteriophages. A variant of domain substitution and deletion of intermediate domains is considered.

A broad variability of one-domain proteins' lengths suggests that they are a product of deletions. Possibly, recombination between polydomain variants could also produce deletion variants consisting of one or two domains (Fig. 13); herewith, they were selected by only one property, the possibility of attachment to the phage surface. Aggregation of the phage can decrease its infectivity, so at least a one-domain variant is required on the surface of the head.

As some bacteriophages contain two variants of the *hoc* gene, they may recombine to form new variants of this gene (Fig. 13).

For four-domain Hoc proteins of bacteriophages RB49, Phi1 and JSE, an alternative mechanism for the origin of the four-domain variant can be assumed. The position of the second domains of bacteriophages RB49, Phi1, JSE on the common phylogenetic tree enables retracing the similarity of these domains with C-terminal domains of both these and other phages. In turn, the third domains of bacteriophages RB49, Phi1, JSE belong to one branch with the first (N-terminal) domains in Hoc proteins of phages Ae65, IME08, JS10. Such a similarity makes it possible to assume another model for the origin of Hoc proteins of bacteriophages RB49, Phi1, JSE – due to the duplication of the gene of the two-domain precursor (Fig. 14).

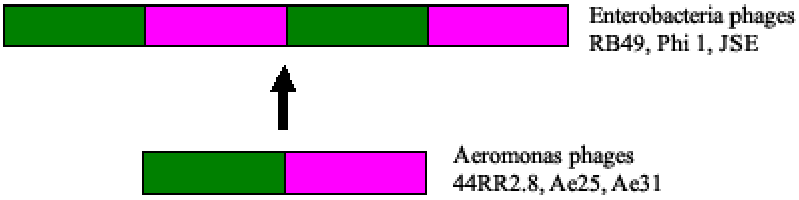

Fig.14. Scheme of the duplication in the evolution of the two-domain variant of Hoc protein into a four-domain variant. N-terminal domains are highlighted in green; C-terminal domains, in magenta.

Thus, we assumed three variants of the evolution of Hoc proteins' amino acid sequence: the variant of intragenic duplications of a gene part encoding the intermediate domain, the variant of domain mixing in recombination between the *hoc* genes of various closely related bacteriophages and the variant of the origin of four-domain protein at the duplication of two-domain protein. Most likely, the emergence of a great diversity of amino acid sequences in Hoc protein, the major phage antigen, involved all three variants of the recombination evolution of their genes.

CONCLUSION

Scientific literature and databases abound in descriptions of T4-like bacteriophages. They are singled out from the steadily growing number of various biological objects. This enables a statement that sequenced genomes of T4-like bacteriophages that are at the disposal of researchers are only a minor part of the diversity of these viruses, which really exists in nature. The development of methods of their classification is of special significance. The approach developed here is consistent with the known methods of classification by genome sequences [20] and is based on more natural criteria. As a genetic marker, we took proteins containing immunoglobulin-like domains. These markers exist in genomes of almost all representatives of the subfamily. The classification based on the number and quality of such domains is adequate to generally recognized views, according to which the evolutionary relation between phages is masked by the horizontal transfer of genetic material.

The number of immunoglobulin-like domains in Hoc protein of various bacteriophages of the subfamily *Teequatrovirinae* varies from one up to five and determines the antigenic properties of the protein. The use of the number of domains in Hoc protein as a criterion for classification of bacteriophages leads to the division of all bacteriophages of the subfamily into six groups (Table 1):

- bacteriophages having one-domain variants of protein,
- bacteriophages having two-domain variants of protein,
- bacteriophages having three-domain variants of protein,
- bacteriophages having four-domain variants of protein,
- bacteriophages having four-domain variants of protein with the C-terminal tail of the third domain,
- bacteriophages having five-domain variants of protein.

The phylogenetic tree of T4-related bacteriophages constructed by particular domains of *hoc* gene product proteins forms three major branches. These are the branch of C-terminal domains, the branch of N-terminal domains and the branch of intermediate domains.

The obligatory occurrence of the C-terminal domains in all Hoc proteins is indicative of its functional and structural significance for the formation of protein and its attachment to phage’s capsid. The similarity of this domain with adhesive proteins selectins and the occurrence in it of the conserved protein sequence ESRNG also indirectly indicates its involvement in the attachment to the capsid.

When analyzing the first three domains of T4 Hoc protein, we revealed a pronounced similarity with protein components of the mammalian immune system. This can indicate that blood is a normal medium for the existence of the phage in it. In connection with this, its molecular mimicry for immune system components is possible. The occurrence of such domains in tail proteins of other bacteriophages can imply a broad distribution of such a molecular mimicry among phages.

Comparative analysis of particular domains in Hoc protein of bacteriophage RB49 showed that, together with Hoc proteins of bacteriophages Phi1and JSE, it forms an isolated group.

This group is characterized by the absence of similarity of the second and third domains of Hoc protein with immunoglobulins and immunoglobulin-like proteins, as well as with corresponding domains of Hoc proteins of phage T4 and a number of other related bacteriophages. This indicates a significant difference in the antigenic and other properties of Hoc protein of phage RB49 (as well as Phi1 and JSE).

Comparative analysis of the domain organization of Hoc enables singling out three possible pathways of the evolution of its genes:

a) intragenic duplications of intermediate domains;

b) domain interchanging between *hoc* genes of close related bacteriophages in the process of recombination and

c) duplications of the gene of the two-domain variant of Hoc protein.

REFERENCES

1. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997. V. 25. P. 3389–3402.
2. *Basic Local Alignment Search Tool*. URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed 19 June 2018).
3. Balter M. Virology. Evolution on life’s fringes. *Science*. 2000. V. 289. P. 1866–1867.
4. Bateman A., Eddy S.R., Chothia C. Members of the immunoglobulin superfamily in bacteria. *Protein Sci*. 1996. V. 5. № 9. P.1939–1941.
5. Bateman A., Eddy S.R., Mesyanzhinov V.V. A member of the immunoglobulin superfamily in bacteriophage T4. *Virus Genes*. 1997. V. 14. P. 163–165.
6. Bork P., Holm L., Sander C. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J. Mol. Biol*. 1994. V. 242. P. 309–320.
7. Dabrowska K., Zembala M., Boratynski J., Switala-Jelen K., Wietrzyk J., Opolski A., Szczaurska K., Kujawa M., Godlewska J., Gorski A. Hoc protein regulates the biological effects of T4 phage in mammals. *Arch. Microbiol*. 2007. V. 187. № 6. P. 489–498.
8. De Bono B., Chothia C. Exegesis: a procedure to improve gene predictions and its use to find immunoglobulin superfamily proteins in the human and mouse genomes. *Nucleic Acids Res*. 2003. V. 31. № 21. P. 6096–6103.
9. Fokine A., Islam M.Z., Zhang Z., Bowman V.D., Rao V.B., Rossmann M.G.. Structure of the three N-terminal immunoglobulin domains of the highly immunogenic outer capsid protein from a T4-like bacteriophage. *J. Virol*. 2011. V. 85. № 16. P.8141–8148.
10. Fokine A., Leiman P.G., Shneider M.Mю, Ahvazi B., Boeshans K.M., Steven A.C., Black L.W., Mesyanzhinov V.V., Rossmann M.G. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc Nat. Acad. Sci. USA*. 2005. V. 102. P. 7163–7168.

11. Fong S., Hamill S.J., Proctor M., Freund S.M., Benian G.M., Chothia C., Bycroft M., Clarke J. Structure and stability of an immunoglobulin superfamily domain from twitchin, a muscle protein of the nematode *Caenorhabditis elegans*. *J. Mol. Biol*. 1996. V. 264. № 3. P. 624–639.
12. Halaby D.M., Poupon A., Mornon J. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng*. 1999. V. 12. P. 563–571.
13. Jeanmougin F., Thompson J.D., Gouy M., Higgins D.G., Gibson T.J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci*. 1998. V. 23. P. 403–405.
14. Jing H., Takagi J., Liu J.H., Lindgren S., Zhang R.G., Joachimiak A., Wang J.H., Springer T.A. Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. *Structure*. 2002. V. 10. P. 1453–1464.
15. *ICTV Virus Taxonomy: 2011 Release*. URL: http://ictvonline.org/virusTaxonomy.asp?version=2011 (accessed 19 June 2018).
16. Ishii T., Yamaguchi Y., Yanagida M. Binding of the structural protein Soc to the head shell of bacteriophage T4. *J. Mol. Biol*. 1978. V. 120. P. 533–544.
17. Ishii T., Yanagida M. Molecular organization of the shell of the T-even bacteriophage head. *J. Mol. Biol*. 1975. V. 97. P. 655–660.
18. Ishii T., Yanagida M. The two dispensable structural proteins (Soc and Hoc) of the T4 phage capsid; their purification and properties, isolation and characterization of the defective mutants, and their binding with the defective heads *in vitro*. *J. Mol. Biol*. 1977. V. 109. P. 487–514.
19. Kadyrov F.A., Shlyapnikov M.G., Kryukov V.M. A phage T4 site-specific endonuclease, SegE, is responsible for a non-reciprocal genetic exchange between T-even-related phages. *FEBS Lett*. 1997. V. 415. № 1. P. 75–80.
20. Lavigne R., Darius P., Summer E.J., Seto D., Mahadevan P., Nilsson A.S., Ackermann H.W., Kropinski A.M. Classification of *Myoviridae* bacteriophages using protein sequence similarity. *BMC Microbiol*. 2009. V. 9. Article No. 224.
21. Lawrence J.G., Hatfull G.F., Hendrix R.W. Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. *J. Bacteriol*. 2002. V. 184. P. 4891–4905.
22. Rohwer F., Edwards R. The phage proteomic tree: a genome-based taxonomy for phage. *J. Bacteriol*. 2002. V. 184. № 16. P. 4529–4535.
23. Ross P.D., Black L.W., Bisher M.E., Steven A.C. Assembly-dependent conformational changes in a viral capsid protein. Calorimetric comparison of successive conformational states of the gp23 surface lattice of bacteriophage T4. *J. Mol. Biol*. 1985. V. 183. P. 353–364.
24. Sathaliyawala T., Islam M.Z., Li Q., Fokine A., Rossmann M.G., Rao V.B. Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. *Mol. Microbiol*. 2010. V. 77. № 2. P. 444–455.
25. Susskind M.M., Botstein D. Molecular genetics of bacteriophage. *Microbiol. Rev*. 1978. V. 42. P. 385–413.
26. Tamura K., Dudley J., Nei M., Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol*. 2007. V. 24. P. 1596–1599.
27. Teichmann S.A., Chothia C. Immunoglobulin superfamily proteins in *Caenorhabditis elegans*. *J. Mol. Biol*. 2000. V. 296. № 5. P. 1367–1383.
28. Vogel C., Teichmann S.A., Chothia C. The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. *Development*. 2003. V. 130. № 25. P. 6317–6328.
29. Yamaguchi Y., Yanagida M. Head shell protein hoc alters the surface charge of bacteriophage T4. Composite slab gel electrophoresis of phage T4 and related particles. *J. Mol. Biol.* 1980. V. 141. P. 175–193.

Received 20 March 2018.
Published 25 June 2018.