# EXPERIMENTAL ARTICLES

UDC 577.21 https://doi.org/10.15407/biotech13.06.030

# EXPRESSION OF NUCLEOCAPSID VIRAL PROTEINS IN THE BACTERIAL SYSTEM OF *Escherichia coli*: THE INFLUENCE OF THE CODON COMPOSITION AND THE UNIFORMITY OF ITS DISTRIBUTION WITHIN GENE

*E. G. Fomina*
*E. E. Grigorieva*
*V. V. Zverko*
*A. S. Vladyko*

State Institution "Republican Scientific and Practical Center for Epidemiology and Microbiology", Republic of Belarus, Minsk

*E-mail: feg1@tut.by*

Received 10.06.2020
Revised 12.11.2020
Accepted 30.12.2020

A heterologous host has got a unique expression ability of each gene. Differences between the synonymous sequences play an important role in regulation of protein expression in organisms from *Escherichia coli* to human, and many details of this process remain unclear. The work was aimed to study the composition of codons, its distribution over the sequence and the effect of rare codons on the expression of viral nucleocapsid proteins and their fragments in the heterologous system of *E. coli*. The plasmid vector pJC 40 and the BL 21 (DE 3) *E. coli* strain were used for protein expression. The codon composition analysis was performed using the online resource (*www.biologicscorp.com*). Ten recombinant polypeptides were obtained, two of them encoding the complete nucleotide sequence of nucleocapsid proteins (West Nile and hepatitis C viruses) and the fragments including antigenic determinants (Lassa, Marburg, Ebola, Crimean-Congo hemorrhagic fever (CCHF), Puumala, Dobrava-Belgrade, Hantaan, and lymphocytic choriomeningitis viruses). Hybrid plasmid DNAs provided efficient production of these proteins in the prokaryotic system. The recombinant protein yield varied from 5 to 40 mg per one liter of bacterial culture. No correlation was found between the level of protein expression and the frequency of rare codon occurrence in the cloned sequence: the maximum frequency of rare codon occurrence was observed for the West Nile virus (14.6%), the minimum one was for the CCHF virus (6.6%), whereas the expression level for these proteins was 30 and 5 mg/l of culture, respectively. The codon adaptation index (CAI) values, calculated on the basis of the codon composition in *E. coli*, were in the range from 0.50 to 0.58, which corresponded to the average expressed proteins. The analysis of the CAI distribution profiles indicated the absence of rare codons clusters that could create difficulties in translation. Difference between the frequencies of the amino acids distribution and their content in E. coli was statistically significant for the nucleocapsid proteins of the Marburg, Ebola, West Nile, and hepatitis C viruses.

***Key words***: recombinant nucleocapsid proteins, expression, rare codons, codon adaptation index.

The development of recombinant DNA technology laid the foundation for the expression of proteins in a wide range of various cellular systems (from bacterial to eukaryotic and cell-free ones). Thus, the process of recombinant proteins production has become faster and easier in comparison with their natural counterparts. Until now, *Escherichia coli* remains the main host for protein production. The main advantages of this system are potentially very high expression levels, rapid growth of culture, low

cost of media and simple cultivation conditions are the main advantages of this system. The scale of production of therapeutic, diagnostic, and industrially important proteins and/or enzymes ultimately depends on protein expression the in a heterologous system. The expression ability of each gene is unique in a heterologous host, and not all proteins are successfully synthesized in *E. coli* cells. Many factors, such as a vector and a host, the promoter strength, the inducers concentration, the composition of the medium, and a number of others, affect the efficiency of protein production [1]. It has been known for a long time that gene expression in a heterologous host is disrupted due to differences in the use of codons by an organism [2]. Degenerate coding of 20 amino acids by 61 nucleotide triplets makes it possible to synthesize the same protein sequence using a huge number of synonymous mRNAs. Differences between synonymous sequences play an important role in the regulation of protein expression in the organisms from *E. coli* [3–5] to human, and many details of this process remain unclear [6]. Most literature on codon use are focused on a study of rare codons such as AUA codon for Ile and AGA, AGG and CGG for Arg, insufficient amount of the corresponding specialized tRNA [7]. However, this assumption was refuted by the studies on ribosome profiling, which showed that the net rate of translation-elongation, as a rule, was constant and did not depend on the use of codons [8]. The level of protein expression could be influenced not only by the presence of so-called rare codons, but also by their location and distribution within the gene. It has been reported that the presence of the AAA codon at position of +2 gene increases expression [9], while the NGG codon has the opposite effect at position +2 [10]. In this context, the codon composition of a heterologous gene can be optimized for expression in a particular host.

Studying the effect of mRNA sequence on protein expression is complicated by the fact that the changes in synonymous sequences (optimization) simultaneously affect many parameters, including identity, codon homogeneity and mRNA folding, as well as other features of local and whole sequences, ranging from the effects of codon pairs to theirs general content of A/U/C/G [11].

The aim of this research was to study the composition of codons, its distribution over the gene, and the effect of rare codons on the expression of viral nucleocapsid proteins and their fragments in the heterologous system of *Escherichia coli*.

## Material and Methods

Lassa virus (LASV) (Josiach strain), Marburg virus (MARV) (Voege strain), Ebola virus (EBOV) (Zair strain) were obtained from Dr.G. van der Groen (Institute of Tropical Medicine, Belgium). Lymphocytic choriomeningitis virus (LCMV) (Armstrong strain) was obtained from the Center for Disease Control and Prevention (Atlanta, USA). Crimean-Congo hemorrhagic fever (CCHF) virus (strains 22263, Astrakhan and Uzbekistan); Hantaan virus (HTNV) (strain 4950); Dobrava-Belgrade virus (strain Aa 118) were obtained from the Ivanovsky Institute of Virology of RAMS (Moscow, Russian Federation). Puumala virus (PUUV) (strains CG-1820 and K-27) was obtained from the Center of the Ministry of Health of the Russian Federation for Combating Hemorrhagic Fever with Renal Syndrome (headed by Professor E.A. Tkachenko, Doctor of Medical Sciences) (Moscow, Russian Federation). Virus-containing fluid from a patient with laboratory-confirmed diagnosis of hepatitis C virus was used to obtain a nucleotide sequence encoding the hepatitis C virus nucleocapsid protein.

*E. coli* strain DH5 α (*supE44 lacU169 (f80 lacZΔM15) hsdR17 recA1 endA1 gyrA96 thi-1 relA1*) was used for genetic engineering. *E. coli* strain BL21 (DE3) (*F– ompT gal dcm lon hsdSB* (rB —mB –) λ (DE3 [*lacI lacUV5-T7p07 ind1 sam7 nin5*]) [*malB*+]K-12 (λS)) was used for the expression of recombinant proteins. Plasmid pJC 40 [12] was used as an expression vector providing transcription of cloned genes under the control of the T7 polymerase promoter. The plasmid contains an additional fragment encoding 10 histidine residues (His), which is localized at the N-terminal part of the polypeptide during translation, allowing protein purification by metal chelate chromatography.

Topological analysis of nucleocapsid proteins amino acid sequences to search for antigensignificant sites was carried out using the computer program wxGeneBee.

Oligonucleotide sequences (primers) for the amplification of DNA fragments encoding regions of nucleocapsid proteins were synthesized by Primetech (Republic of Belarus).

Viral RNA obtained from virus-containing fluid was used as an initial template for reverse transcription. RNA was extracted using NucleoSpin RNA reagent kit (MACHERY-Nagel, Germany).

The reverse transcription reaction was performed using the Reverta-L reagents kit (Amplisense, Russian Federation) according

to the manufacturer's instructions. The complementary DNA (cDNA) was used for PCR. Reaction mixture contained 25 pmol of primers, 5 µl of 10x Taq buffer (Primetech, Republic of Belarus), 4 µl of 25mM MgCl2, 1 µl of dNTP (10 mM each), 10 µl of RT product (cDNA), 2.5 units of Taq DNA polymerase (Primetech, Republic of Belarus), deionized water to a final volume of 50 µl.

Analysis of DNA fragments synthesized in PCR was carried out by electrophoresis in 1.5–2% agarose gel. DNA was visualized in UV light by staining the gel with ethidium bromide.

DNA was hydrolyzed with restriction enzymes (1 U each) in accordance with the instructions (Thermo Scientific, USA) at 37 °C for 2 hours in a final volume of 20 µl. After incubation the enzymes were inactivated for 15 min at 65 °C.

DNA fragments were ligated in a volume of 20 µl at 22 °C for an hour. T4 DNA Ligase (Thermo Scientific, USA) was used as a ligating enzyme according to the manufacturer's instructions.

Isolation of recombinant plasmid DNA was performed using the QIAprep Spin Miniprep Kit (Qiagen).

Bacterial *E. coli* cells, BL 21 (DE3) strain, transformed with plasmid DNA, were cultured in a liquid LB medium containing ampicillin as a selective agent at a concentration of 50 µg/ml (Sigma, United States) with constant shaking at 37 °C until the cell culture reached the optical density OD600 = 0.3. Isopropyl-β-D-thiogalactopyranoside (IPTG, Sigma, USA) was added to the medium and cells were further incubated for 3.5 hours under the same conditions. Inducer concentration, cultivation temperature, medium composition were optimized to provide the maximum yield of each recombinant polypeptide.

Analysis of recombinant polypeptides in polyacrylamide gel (PAGE) was performed according to the method proposed by Laemmli.

Analysis of the codon composition and calculation of the codon adaptation index (CAI) were performed using an online resource (www.biologicscorp.com). Each cloned sequence was divided into segments of 10 codons for assessing the distribution of CAI index values within a gene. The index value was determined for each of these segments.

Statistical data processing was carried out using the STATISTICA 6.0 package.

## Results and Discussion

During development of immunodiagnostic kits for highly pathogenic viruses, the use of biotechnological methods is now of great importance, allowing the production of antigens and antibodies to them by artificial means. This eliminates the need to work with a highly contagious pathogen and is more cost-effective and technologically advanced. Recombinant technologies make it possible to express not only full-length proteins, but also their individual antigenic fragments, as well as mosaic molecules consisting of separate sections of one or several antigenic determinants.

The nucleocapsid protein (N-protein) packs the genome, acts as an RNA chaperone, provides intracellular protein transport, participates in DNA degradation and translation processes of the host cell. And it is commonly used as an antigen in the most serological studies, since it has a pronounced antigenic and immunogenic activity along with a high conservatism.

This article presents data on the expression of full-length nucleocapsid proteins and their fragments of ten highly dangerous viruses from various families of Arenaviridae (Lassa virus and lymphocytic choriomeningitis virus), Bunyaviridae (Hantaan, Puumala, Dobrava-Belgrade and Crimean-Congo hemorrhagic fever viruses), Filoviridae (Marburg and Ebola viruses), Flaviviridae (hepatitis C and West Nile viruses) in the prokaryotic system (*Escherichia coli* BL 21 (DE 3)) and expression vector pJC 40.

The potential antigenic determinants exposed on a molecule surface were found. The topological maps of the amino acid sequences of viral nucleocapsid proteins were compiled using the wxGeneBee computer program. The program is based on the scales of hydrophobicity and hydrophilicity proposed by Hopp and Woods (1981) and Kyte and Doolitle (1982). The evaluation criteria include such parameters as solubility, charge, distance from the NC backbone, the presence of helices, and the presence of sulfhydryl residues.

Analysis of antigenic determinants in nucleocapsid proteins showed their sequential arrangement throughout the entire sequence for the proteins of Lassa, LCM and hepatitis C viruses; C-terminal localization in Marburg, Ebola, West Nile viruses; placement at the N-terminus for Puumala, Hantaan, Dobrava-Belgrade viruses; central location in the CCHF virus. These studies served as a basis for selection and subsequent cloning into the expression vector of the following sequences: 432 amino acids (aa) of the Lassa virus (6/8 of amino acid sequence from position 137 aa to 569 aa); 320 amino acids of LCMV

(from 41 to 361 aa). For etiological agents of hemorrhagic fever with renal syndrome (HFRS), the most extended antigenic regions within the first 117 amino acids were found in the N-terminal part of the molecule (amino acids 1 to 117). For the CCHF virus the choice was made in favor of a sequence of 105 amino acid bases (from 202 to 306 aa), including one antigensignificant site with the highest antigenicity index located in the center of the protein. Five hydrophilic regions for the Marburg virus and six for the Ebola virus were identified as potential antibody binding sites in the C-terminal part of the protein. The most extended antigen-significant amino acid areas are localized in the region 442–695 aa of Marburg virus nucleocapsid, represented by four B-sites, and in region 434–739 aa of Ebola virus nucleocapsid, including all six antigenic determinants. The profile of antigenic sites in the nucleocapsid proteins of hepatitis C and West Nile viruses showed their uniform distribution in the sequence for the hepatitis C virus and the C-terminal location for the West Nile virus. Small size of the West Nile virus nucleocapsid protein and identification of only one antigenic determinant were the reasons to clone the full-length core proteins of these viruses. Specific fragments of 190 amino acids and 103 amino acids encoding full-length nucleocapsid proteins of hepatitis C virus and West Nile virus, respectively, were cloned into the expression vector.

The results of the experiments showed that the level of recombinant polypeptides expression in optimal for each of them conditions varied to a large extent (by 8 times): from 40 mg per liter of culture for representatives of arenoviruses (Lassa virus, LCMV) to 5 mg per liter for the CCHF virus (Fig. 1). The highest yield of recombinant nucleocapsid proteins (30–40 mg/l of cell culture) was observed for the cloned sequences of three viruses: Lassa, LCM, and West Nile (Fig. 1, *A, B, C*, respectively). The average expression level was detected for the fragments of the Hantaan, Marburg, Ebola, Puumala, Dobrava-Belgrade, and hepatitis C viruses (from 25 to 15 mg/l of cell suspension) (Fig. 1, *C, E, F, H*). CCHF virus nucleocapsid was poorely expressed (5 mg/l) (Fig. 1, *D*).

One of the explanation for the different levels of protein biosynthesis may be the different number of codons that are rare for *E. coli* in the cloned viral sequence.

Amino acid codons rarely found in *E. coli* cells include AGG/AGA/CGA triplets encoding arginine (frequency of occurrence 1.1; 2.0;

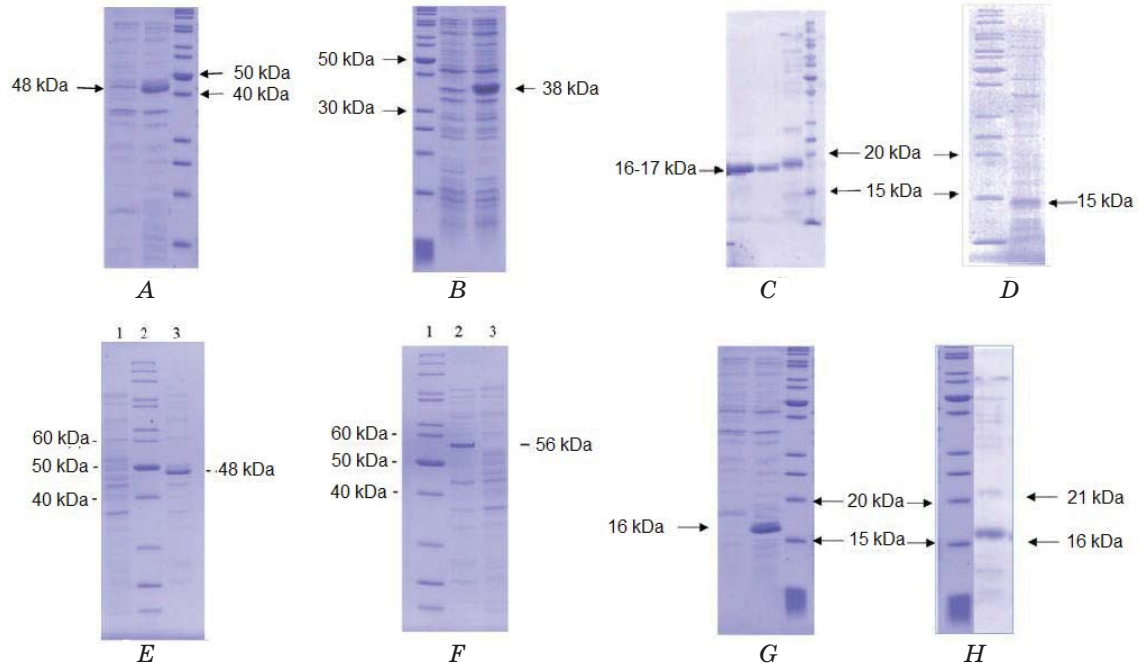


*Fig. 1.* **Electrophoretic analysis of recombinant viral nucleocapsid proteins expression**

Black arrows indicate the target viral proteins: LCM (*A*, lane 2); Lassa (*B*, lane 3); Hantaan, Dobrava-Belgrade, Puumala (*C*, lanes 1, 2, 3, respectively); CCHF (*D*, lane 2); Marburg (*E*, lane 3); Ebola (*F*, lane 2); West Nile (*G*, lane 2); hepatitis C (*H*, lane 2). The samples of bacterial lysates of the cell clones expressing recombinant viral nucleocapsid proteins after IPTG induction were used for the analysis

3.5 per 1000); CTA for leucine (3.8 per 1000); ATA for isoleucine (4.2 per 1000); CCC for proline (5.0 per 1000); GGA for glutamine (8.7 per 1000). If the distribution of codons was uniform, the frequency of occurrence for all codons would be about 16.4. The analysis of rare codon content in the cloned sequences is shown in the Table.

Analysis of the rare codons frequency (Fig. 2, *A*) showed that in the cloned sequences of most viral proteins (Lassa, LCM, Marburg, Ebola, Puumala, Dobrava-Belgrade, Hantaan) it ranged from 7.5 to 9.5%. The lowest of this value was for the CCHF virus polypeptide (6.6%), the highest one was for the triplets of the hepatitis C and West Nile viruses (12.1 and 14.6%, respectively). Based on this analysis, the maximum expression level is expected to be observed for the protein of the CCHF virus, and the minimum one should be for hepatitis C and West Nile viruses.

However, our data obtained by using pJC 40 expression plasmid, bacterial strain BL21 (DE3), and optimized for each polypeptide expression conditions indicated that the maximum protein yield was observed for the amino acid sequences of the Lassa and LCM viruses (approximately 35–40 mg per liter of a culture), and minimum one was for the CCHF virus (5 mg/l) (Fig. 2, *B*). Thus, there is no clear correlation between the yield of recombinant proteins and the total frequency of rare codons in their sequences. It should be noted that a feature of the CCHF virus cloned sequence is the fact that the AGG codon, which is the rarest for *E. coli*, is located at position +2, the closest to the initiating codon. According to the literature, not only the presence of rare codons, but also their location can be critical for protein expression. The proximity to the initiating codon can explain the low synthesis efficiency (5 mg/l) of this protein in the heterologous system.

It is interesting to note that, despite the large number of "not typical" codons for the nucleocapsid protein of the West Nile virus (14.6%), its biosynthesis proceeded quite efficiently (30 mg/l). This can be partly explained by the fact that the main contribution to the percentage of rare codons in this case is not made by the most infrequent ones. The

**Analysis of the rare codon number in cloned sequences**

| Virus (number of cloned amino acids) | Arg | | | Leu | Ile | Pro | Gln | General frequency of occurrence |
|---|---|---|---|---|---|---|---|---|
| | AGG (1.1%) | AGA (2.0%) | CGA (3.5%) | CTA (3.8%) | ATA (4.2%) | CCC (5.0%) | GGA (8.7%) | |
| Lassa virus (432 aa) | 5/1.16* | 12/2.78 | – | 3/0.69 | 6/1.39 | 4/0.93 | 11/2.55 | 41/9.5 |
| LCM virus (320 aa) | 8/2.5 | 9/2.81 | – | 4/1.25 | 2/0.63 | 2/0.63 | 2/0.63 | 27/8.4 |
| Marburg virus (254 aa) | 3/1.18 | 8/3.15 | – | – | 3/1.18 | 2/0.79 | 3/1.18 | 19/7.5 |
| Ebola virus (306 aa) | 3/0.98 | 8/2.61 | 1/0.33 | 3/0.98 | 1/0.33 | 6/1.96 | 3/0.98 | 25/8.2 |
| Puumala virus (117 aa) | – | 7/5.98 | – | – | 1/0.85 | – | 1/0.85 | 9/7.7 |
| Dobrava-Belgrade virus (116 aa) | 3/2.59 | 2/1.72 | – | 2/1.72 | 1/0.86 | – | 2/1.72 | 10/8.6 |
| Hantaan virus (117 aa) | 5/4.27 | 2/1.71 | – | – | 1/0.85 | 1/0.85 | 2/1.71 | 11/9.4 |
| CCHF virus (106 v.) | 1/0.94 | – | 1/0.94 | – | 3/2.83 | – | 2/1.89 | 7/6.6 |
| West Nile virus (103 aa) | 1/0.97 | 3/2.91 | – | 3/2.91 | 2/1.94 | 2/1.94 | 4/3.88 | 15/14.6 |
| Hepatitis C virus C (190 aa) | 7/3.68 | 1/0.53 | 1/0.53 | 1/0.53 | – | 10/5.26 | 3/1.58 | 23/12.1 |

* Absolute number of codons/frequency of occurrence, expressed as a percentage.

cloned sequence of this protein contains only one AGG triplet, located in the middle.

Codon adaptation index (CAI), which reflects the degree of non-uniformity of the codon composition of a gene, is an additional parameter that is widely used in various biological studies to determine the efficiency of translation, predict the level of cell protein synthesis and the expression of foreign genes in heterologous systems [13].

*E. coli* refers to the organisms in which the preferential use of certain codons is observed depending on the gene expression level, while the translation efficiency correlates with the uneven use of codons [14]. Highly expressed genes of *E. coli* (genes of ribosomal proteins, transcription and translation factors, outer membrane proteins) are characterized by high CAI values (over 0.75).

The CAI values for the cloned viral sequences, calculated on the basis of the codon composition in *E. coli*, are in the range from 0.50 to 0.58, which corresponds to the average expressed proteins. Therefore, the codon composition of viral nucleocapsid proteins is relatively optimal for *E. coli*. For example, green fluorescent protein (GFP), which is also foreign to *E. coli* and according to the literature [15] is synthesized in bacterial cells with a high yield, has a coefficient of 0.58. However, despite the proximity of the CAI values to each other, the results of the experiments indicate that viral polypeptides with a similar content of rare codons in the nucleotide sequence are expressed with different efficiencies.

Different levels of the recombinant proteins expression can also be associated with the peculiarities of the rare codons location. Additional information for assessing the spatial distribution of rare codons can be obtained by studying the segmental distribution of CAI values within a gene. The obtained data are represented in graphical form in Fig. 3. As a comparison the coding sequences of highly expressed *E. coli* proteins (*rspB 30S* — ribosomal protein S2, *rec A* — recombinase A), as well as heterologous green fluorescent protein GFP (*gfp*) were analyzed.

The CAI distribution profiles for cloned sequences are different. For example, more flattened curve corresponds to the genes of the Dobrava-Belgrade and CCHF viruses (Fig. 3, *A*) while an alternation of pronounced peaks are evidenced for the nucleotide sequences of the Ebola, LCM and Marburg viruses (Fig. 3, *B*). In general, distribution of CAI values for genes of viral nucleocapsid proteins is similar to that for proteins with an average expression level. The curves for *E. coli* proteins are differed by higher values of the index, which are typical for highly expressed proteins. Analysis of the CAI distribution

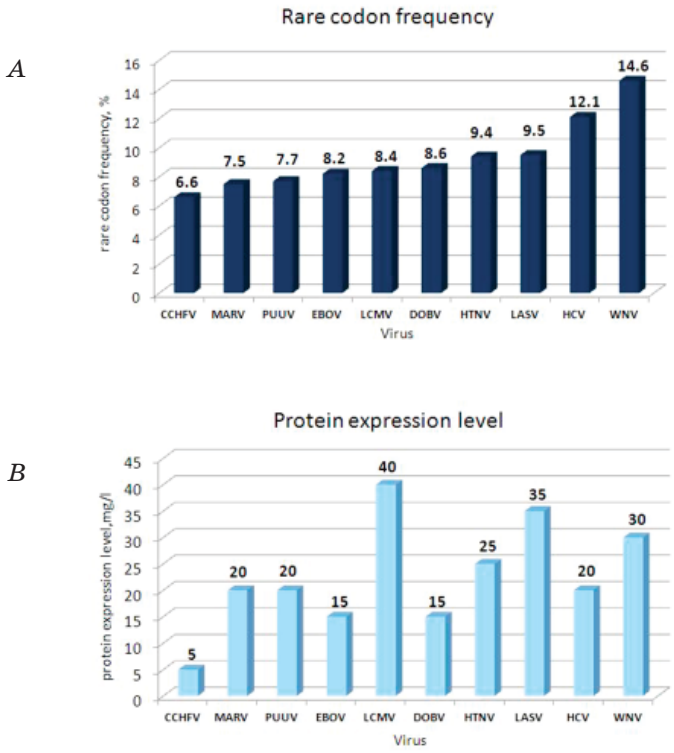


*Fig. 2.* **Analysis of rare codon occurrence in cloned sequences and recombinant proteins expression level**

profiles in the cloned sequences did not reveal areas of sharp decreases with extremely low (less than 0.2) index values. This may indicate the absence of rare codon clusters capable of embarrassing the translation.

As far as is known, the use of synonymous codons in the coding genetic sequences of various organisms is not accidental. Analysis of the codon adaptation index values, as well as the distribution of its values over the cloned sequence for the causative agents of LCM and HFRS, which are the most relevant for the Republic of Belarus zoonotic viral diseases, made it possible to identify some patterns in the use of codons. Higher CAI values for all studied viral sequences indicate a potentially more efficient translation of viral proteins in humans compared to *E. coli* (Fig. 4).

Almost similar distribution of CAI along the analyzed sequences of LCM, Puumala, and Dobrava-Belgrade viruses in the case of their expression in humans and rodents (Fig. 5) indicates that the coding strategy of nucleocapsid proteins of these viruses is adapted for optimal replication of pathogens in natural hosts.

The data obtained confirm the existing assumption that one of the manifestations of viruses adaptation to the host during their evolution is the use of certain synonymous codons in the viral genome, which corresponds to codon preferences that are specific for the host [16].

The frequency of the amino acids occurrence is another factor that can affect the level of heterologous protein expression in *E. coli*.

If the amino acid composition of the recombinant protein is skewed compared to the typical *E. coli* proteins, heterologous expression may result in translation disruptions. Premature termination, reading frame shift, or amino acids misincorporation can lead to a decrease in the amount or quality of the expressed protein [17–20].

Analysis of the cloned sequences amino acid composition in comparison with that typical to *E. coli* proteins [18] showed the uneven content of amino acids in the recombinant proteins. The polypeptides of the Lassa, LCM, and CCHF viruses were the closest to bacterial proteins in the percentage of amino acids in their composition (Fig. 6, *A*). For nucleocapsid

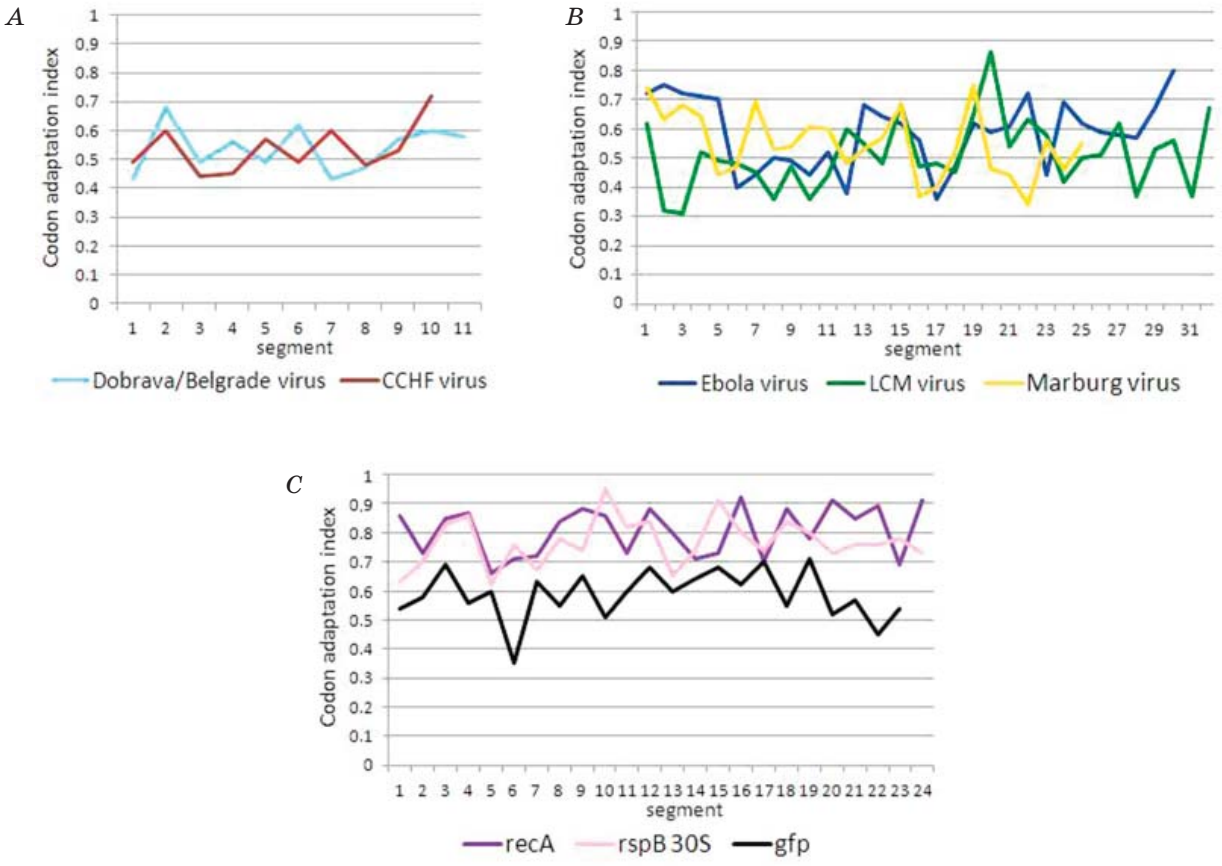


*Fig. 3.* **Segmental distribution of the CAI values within genes encoding the nucleocapsid viral proteins and highly expressed *E. coli* proteins**

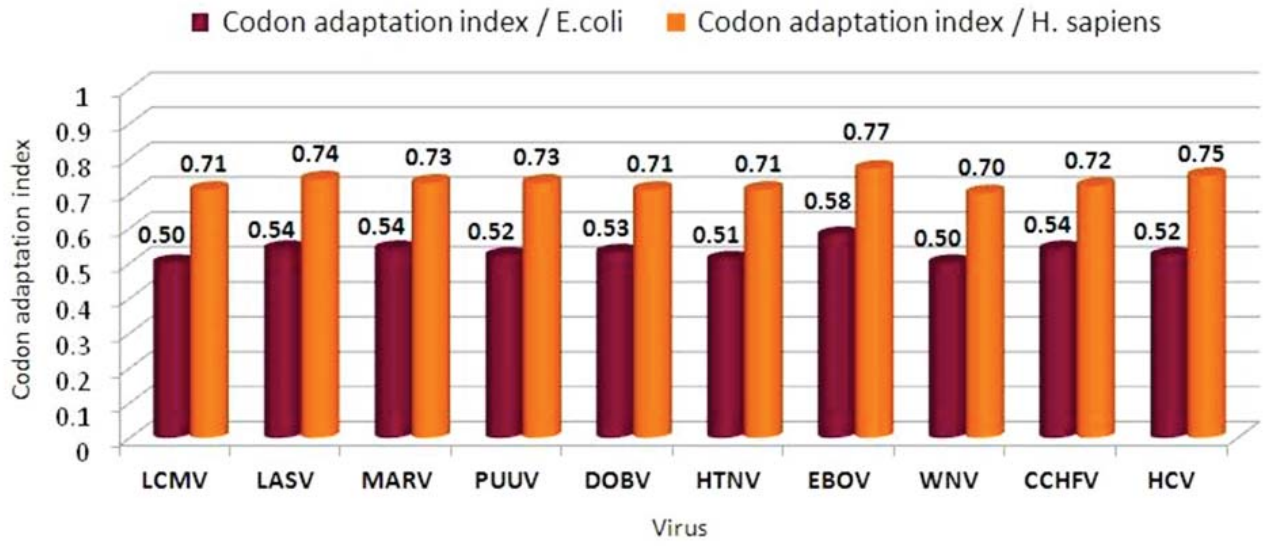


*Fig. 4.* **CAI values for viral sequences, calculated based on the codon composition of *E. coli* and humans**

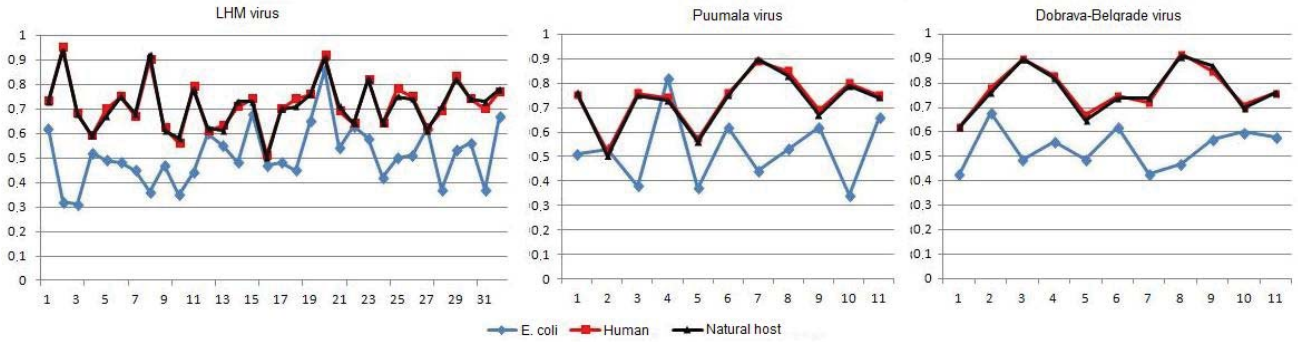


*Fig. 5.* **Distribution of CAI along the analyzed sequences of LCM, Puumala, and Dobrava-Belgrade viruses when expressed in humans and rodents**

proteins of Marburg, Ebola, West Nile and hepatitis C viruses, a statistically significant difference was observed between the frequencies of distribution of amino acids in the cloned sequences and their content in *E. coli* ($\chi^2 = 36.79$, $P = 0.008$; $\chi^2 = 37.60$, $P = 0.007$; $\chi^2 = 32.29$, $P = 0.029$; $\chi^2 = 33.96$, $P = 0.019$, respectively) (Fig. 6, *C*).

Analysis of the data obtained demonstrated the multidirectional nature of existing deviations. The frequency of various amino acids occurrence in viral proteins differed from that in the bacterial proteome both upward and downward. For example, the amino acids Cys, Trp, and Phe, which were rare for *E. coli* proteins, were not contained in the sequences of the Puumala, Dobrava-Belgrade, and Hantaan nucleocapsid proteins. For a number of cloned sequences, there was a significant excess in the frequency of occurrence of both the most rare amino acids for *E. coli* proteins (Cys in the Lassa virus sequence, Trp in the hepatitis C virus sequence), and relatively uniformly presented in the bacterial proteome (Asp in the sequences of Marburg, Ebola, Puumala, Dobrava-Belgrade viruses; Pro in the sequences of Marburg, Ebola, hepatitis C viruses, Lys in the sequences of viruses Dobrava-Belgrade, West Nile, CCHF viruses).

Our studies have shown that the nucleotide sequences of ten nucleocapsid proteins of belonging to various families dangerous and especially dangerous human viruses are expressed in a prokaryotic heterologous system quite efficiently: from 5 to 40 mg per liter of bacterial culture. Bioinformatic analysis

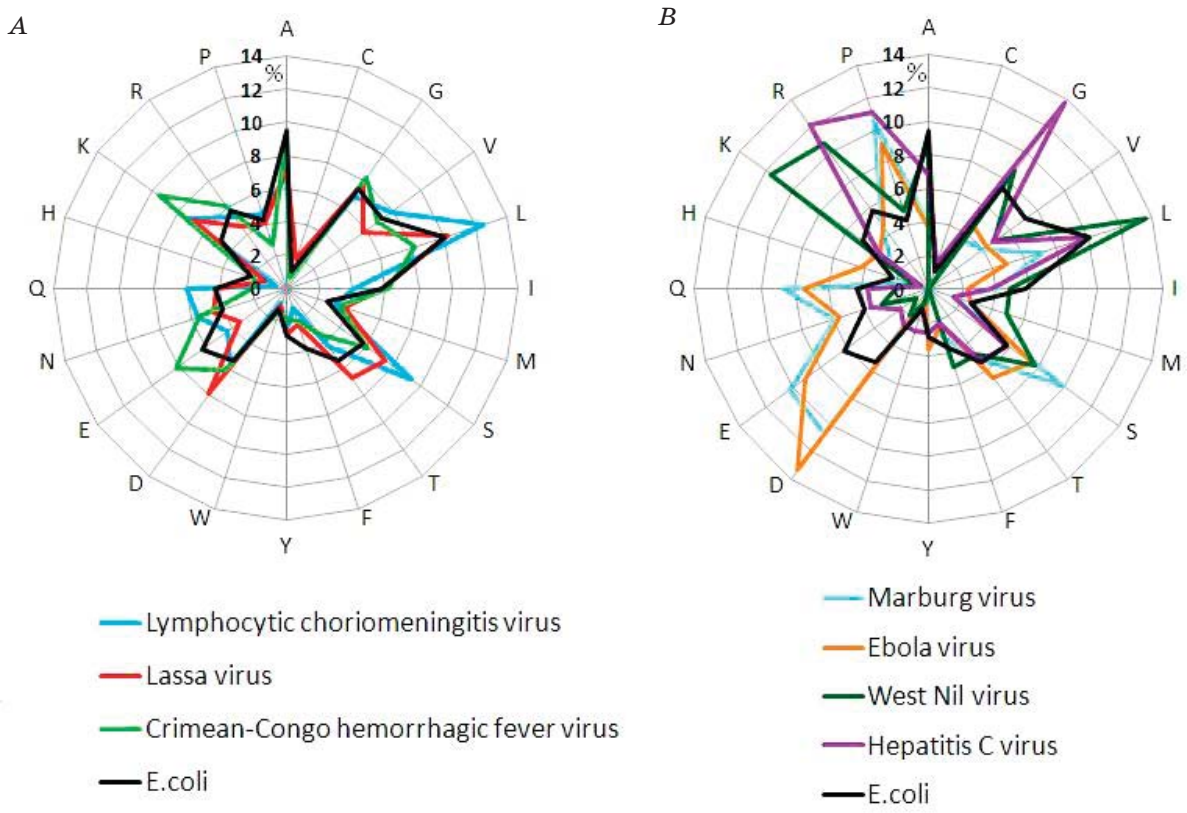


*Fig. 6*. **Histogram of the amino acid composition of the cloned viral sequences and the amino acid composition typical for *E. coli* proteins**

of the codon adaptation index classifies recombinant polypeptides as proteins with an average level of xpression in the *E. coli* system. At the same time, the efficiency of viral proteins translation in the human as a typical host is potentially higher than in *E. coli*, as evidenced by the higher CAI values for all viral sequences.

Nevertheless, significant differences in the frequency of rare codon occurrence in sequences and the absence of its correlation with the level of protein expression were revealed, which confirms the complexity of forecasting the of the protein biosynthesis process in a heterologous system and its unpredictable nature.

The authors express their sincere gratitude to the head of the laboratory of biotechnology and immunodiagnostics of especially dangerous viral infections of the Republican Scientific and Practical Center for Epidemiology and Microbiology P.A. Semizhon and leading scientific researcher E.P. Scheslenok for their help in the recombinant polypeptide production and preparing the manuscript.

The research was carried out with the financial support of the State Committee on Science and Technology of the Republic of Belarus within the framework of the project (D52) “Development of technology and mastering the production of a confirmatory diagnostic test system for detecting specific antibodies to the hepatitis C virus by immunoblotting” (State registration No. 20142189, dated 19.09. 2014), subprogram 8 “Import-substituting diagnostic tools and biological products — 2020”, the State program “Science-intensive technologies and technique” for 2016–2020.

## REFERENCES


1. *BoëlG., LetsoR., NeelyH., PriceW.N., WongK.H., Su M., Luff J., Valecha M., Everett J. K., Acton T. B., Xiao R., Montelione G. T., Aalberts D. P., Hunt J. F.* Codon influence on protein expression in *E. coli* correlates with mRNA levels. *Nature*. 2016, V. 529, P. 358–363. https://doi.org/10.1038/nature16509
2. *Robinson M., Lilley R., Little S., Emtage J. S., Yarranton G., Stephens P., Millican A., Eaton M., Humphrey G.* Codon usage can affect efficiency of translation of genes in *Escherichia coli*. *Nucleic Acids Res.* 1984, V. 12, P. 6663–6671.
3. *Goodman D. B., Church G. M., Kosuri S.* Causes and effects of N-terminal codon bias in bacterial genes. *Science*. 2013, 342 (6157), 475–479. https://doi.org/10.1126/science.1241934
4. *Castillo-Mendez M. A., Jacinto-Loeza E., Olivares-Trejo J. J., Guarneros-Pena G., Hernandez-Sanchez J.* Adenine-containing codons enhance protein synthesis by promoting mRNA binding to ribosomal 30S subunits provided that specific tRNAs are not exhausted. *Biochimie*. 2012, V. 94, P. 662–672. https://doi.org/10.1016/j.biochi.2011.09.019
5. *Bentele K., Saffert P., Rauscher R., Ignatova Z., Bluthgen N.* Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* 2013, V. 9, P. 675. https://doi.org/10.1038/msb.2013.32
6. *Cannarozzi G., Schraudolph N., Mahamadou_ Faty, Peter von Rohr, Friberg M., Roth A., Gonnet P., Gonnet G., Barral Y.* A role for codon order in translation dynamics. *Cell*. 2010, V. 141, P. 355–367. https://doi.org/10.1016/j.cell.2010.02.036
7. *Vivanco-Dominguez S., Bueno-Martínez J., León-Avila G., Nobuhiro Iwakura, Kaji A., Kaji H., Guarneros G.* Protein synthesis factors (RF1, RF2, RF3, RRF, and tmRNA) and peptidyl-tRNA hydrolase rescue stalled ribosomes at sense codons. *J. Mol. Biol.* 2012, V. 417, P. 425–439. https://doi.org/10.1016/j.jmb.2012.02.008
8. *Li G. W., Burkhardt D., Gross C., Weissman J. S.* Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*. 2014, V. 157, P. 624–635. https://doi.org/10.1016/j.cell.2014.02.033
9. *Sato T., Terabe M., Watanabe H., Gojobori T., Hori-Takemoto C., Miura K.* Codon and base biases after initiation codon of the open reading frames in the Escherichia coli genome and their influence on translation efficiency. *J. Biochem*. 2001, V. 129, P. 851–860. https://doi.org/10.1093/oxfordjournals.jbchem.a002929
10. *Gonzalez de Valdivia E. I., Isaksson L. A.* A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli*. *Nucleic Acids Res*. 2004, V. 32, P. 5198–5205. https://doi.org/10.1093/nar/gkh857
11. *Ude S., Lassak S., Starosta A., Kraxenberger T., Wilson D., Jung K.* Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science*. 2013, V. 339, P. 82–85. https://doi.org/10.1126/science.1228985
12. *Clos J., Brandau S.* pJC20 and pJC40 – two high-copy-number vectors for T7 RNA polymerase-dependent expression of recombinant genes in Escherichia coli. *Protein Expr. Purif*. 1994, V. 5, P. 133–137. https://doi.org/10.1006/prep.1994.1020
13. *Lee S., Weon S., Kang C.* Relative codon adaptation index, a sensitive measure of codon usage bias. *Evol. Bioinform. Online*. 2010, V. 6, P. 47–55. https://doi.org/10.4137/ebo.s4608
14. *Vladimirov N. V., Likhoshvai V. A., Matushkin Yu. G.* Correlation of codon biases and potential secondary structures with mRNA translation efficiency in unicellular organisms. *Mol. Biol*. 2007, 41 (5), 843–850. (In Russian).
15. *Kirienko N. V., Lepikhov K. A., Zheleznaya L. A., Matvienko N. I.* Significance of codon usage and irregularities of rare codon distribution in genes for expression of BspLU11III methyltransferases. *Biochem*. 2004, 69 (5), 647–657. (In Russian).
16. *Tyulko J. S., Yakimenko V. V.* Strategy of synonymous codon usage in encoding sequences of the Thick-borne encephalitis virus. *Voprosy virusologii*. 2015, 60 (6), 37–41. (In Russian).
17. *Kaur J., Kumar A.* Strategies for optimization of heterologous protein expression in *E. coli*: roadblocks and reinforcements. *Int. J. Biol. Macromol*. 2018, V. 106, P. 803–822. https://doi.org/10.1016/j.ijbiomac.2017.08.080
18. *Kane J.* Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol*. 1995, V. 6, P. 494–500. https://doi.org/10.1016/0958-1669(95)80082-4
19. *Jia B., Jeon C. O.* High-throughput recombinant protein expression in *Escherichia coli*: current status and future perspectives. *Open Biol*. 2016. V. 6, P. 160–196. https://doi.org/10.1098/rsob.160196
20. *Gopal G., Kumar A.* Strategies for the production of recombinant protein in *Escherichia coli*. *Protein J*. 2013, 32 (6), 419–425. https://doi.org/10.1007/s10930-013-9502-5

## ЕКСПРЕСІЯ НУКЛЕОКАПСИДНИХ ВІРУСНИХ ПРОТЕЇНІВ У БАКТЕРІАЛЬНІЙ СИСТЕМІ *Escherichia coli*: ВПЛИВ КОДОНОВОГО СКЛАДУ ТА РІВНОМІРНОСТІ ЙОГО РОЗПОДІЛУ ВСЕРЕДИНІ ГЕНА

*О. Г. Фоміна О. Є., Григор'єва, В. В. Зверко, А. С. Владико*

Державна установа «Республіканський науково-практичний центр епідеміології та мікробіології», Республіка Білорусь, Мінськ

*E-mail: feg1@tut.by*

Експресійна здатність кожного гена є унікальною у гетерологічного хазяїна. Відмінності між синонімічними послідовностями відіграють важливу роль у регуляції експресії протеїну в організмах від *Escherichia coli* до людини, і багато деталей цього процесу ще не з'ясовано. Метою дослідження було вивчити склад кодонів, його розподіл у послідовності та вплив рідкісних кодонів на експресію вірусних нуклеокапсидних протеїнів і їхніх фрагментів у гетерологічній системі *E. coli*. Для експресії протеїнів використовували плазмідний вектор pJC 40 і штам BL 21 (DE 3) *E. coli*. Аналіз складу кодонів виконано з використанням on-line ресурсу (*www.biologicscorp.com*). Отримано десять рекомбінантних поліпептидів, з них два, що кодують повну нуклеотидну послідовність нуклеокапсидних протеїнів (віруси Західного Нілу і гепатиту С) та їхні фрагменти, що містять антигенні детермінанти (вірус Ласса, Марбург, Ебола, Кримської-Конго геморагічної лихоманки (ККГЛ), Пуумала, Хантаан, Добрава-Белград і лімфоцитарного хоріоменінгіту (ЛХМ)). Гібридні плазмідні ДНК забезпечують ефективне продукування цих протеїнів у прокаріотичній системі з виходом рекомбінатного протеїну, що варіює у 8 разів: від 5 до 40 мг на 1 літр бактеріальної культури. Не виявлена кореляції рівня експресії протеїну з частотою народження рідкісних кодонів у клонованій послідовності: максимальна частота народження рідкісних кодонів у клонованій послідовності спостерігалася для вірусу Західного Нілу (14,6%), мінімальна – для вірусу ККГЛ (6,6%), тимчасом як рівень експресії для цих протеїнів становив 30 і 5 мг/л культури відповідно. Значення індексу адаптації кодонів (CAI), розраховані на основі кодонового складу у *E. coli*, для клонованих вірусних послідовностей знаходяться в діапазоні від 0,50 до 0,58, що відповідає середньоекспресованим протеїнам. Проведений аналіз профілів розподілу CAI у клонованих послідовностях свідчить про відсутність кластерів рідкісних кодонів, здатних створювати труднощі за трансляції. Статистично значущу відмінність між частотами розподілу амінокислот у клонованих послідовностях та їхнім змістом в *E. coli* спостерігали для нуклеокапсидних протеїнів вірусів Марбург, Ебола, Західного Нілу і гепатиту С.

***Ключові слова:*** рекомбінантні нуклеокапсидні протеїни, експресія, рідкісні кодони, індекс адаптації кодонів.

## ЭКСПРЕССИЯ НУКЛЕОКАПСИДНЫХ ВИРУСНЫХ ПРОТЕИНОВ В БАКТЕРИАЛЬНОЙ СИСТЕМЕ *Escherichia coli*: ВЛИЯНИЕ КОДОНОВОГО СОСТАВА И РАВНОМЕРНОСТИ ЕГО РАСПРЕДЕЛЕНИЯ ВНУТРИ ГЕНА

*Е. Г. Фомина, Е. Е. Григорьева, В. В. Зверко, А. С. Владыко*

Государственное учреждение «Республиканский научно-практический центр эпидемиологии и микробиологии», Республика Беларусь, Минск

*E-mail: feg1@tut.by*

Экспрессионная способность каждого гена уникальна у гетерологичного хозяина. Различия между синонимичными последовательностями играют важную роль в регуляции экспрессии протеина в организмах от *Escherichia coli* до человека, и многие детали этого процесса остаются неясными. Цель исследования: изучить состав кодонов, его распределение по последовательности и влияние редких кодонов на экспрессию вирусных нуклеокапсидных протеинов и их фрагментов в гетерологичной системе *E. coli*. Для экспрессии протеинов использовали плазмидный вектор pJC 40 и штамм BL 21 (DE 3) *E. coli*. Анализ состава кодонов выполнен с использованием on-line ресурса (www.biologicscorp.com). Получены десять рекомбинантных полипептидов, из них два, кодирующих полную нуклеотидную последовательность нуклеокапсидных протеинов (вирусы Западного Нила и гепатита С) и их фрагменты, включающие антигенные детерминанты (вирус Ласса, Марбург, Эбола, Крымской-Конго геморрагической лихорадки (ККГЛ), Пуумала, Хантаан, Добрава-Белград и лимфоцитарного хориоменингита (ЛХМ)). Гибридные плазмидные ДНК обеспечивают эффективное продуцирование этих протеинов в прокариотической системе с выходом рекомбинатного протеина, варьирующим в 8 раз: от 5 до 40 мг на 1 литр бактериальной культуры. Не выявлена корреляция уровня экспрессии протеинов с частотой встречаемости редких кодонов в клонированной последовательности: максимальная частота встречаемости редких кодонов на клонированную последовательность наблюдалась для вируса Западного Нила (14,6%), минимальная — для вируса ККГЛ (6,6%), в то время как уровень экспрессии для этих белков составлял 30 и 5 мг/л культуры соответственно. Значения индекса адаптации кодонов (CAI), рассчитанные на основе кодонового состава у *E. coli*, для клонированных вирусных последовательностей находятся в диапазоне от 0,50 до 0,58, что соответствует среднеэкспрессируемым протеинам. Проведенный анализ профилей распределения CAI в клонированных последовательностях указывает на отсутствие кластеров редких кодонов, способных создавать затруднения при трансляции. Статистически значимое отличие между частотами распределения аминокислот в клонированных последовательностях и их содержанием в *E. coli* наблюдалось для нуклеокапсидных протеинов вирусов Марбург, Эбола, Западного Нила и гепатита С.

***Ключевые слова:*** рекомбинантные нуклеокапсидные протеины, экспрессия, редкие кодоны, индекс адаптации кодонов.