TBEV Analyzer: a Platform for Evolutionary Analysis of Tick-borne Encephalitis Virus

Majid Forghani
Institute of Natural Sciences and Mathematics
Ural Federal University
Yekaterinburg, Russia
Forghani@urfu.ru

Sergey Kovalev
Institute of Natural Sciences and Mathematics
Ural Federal University
Yekaterinburg, Russia
Sergey.Kovalev@urfu.ru

Pavel Vasev
Krasovsky Institute of Mathematics and Mechanics
Ural Branch of Russian Academy of Science
Yekaterinburg, Russia
Vasev@imm.uran.ru

Mikhail Bolkov
Institute of Immunology and Physiology
Ural Branch of Russian Academy of Science
Yekaterinburg, Russia
M.Bolkov@iip.uran.ru

***Abstract*—Tick-borne encephalitis virus (TBEV) is a causative agent an extremely dangerous infectious disease, often characterized by damage to the central nervous system, which sometimes leads to serious health problems, such as disability and even death. Currently, there is no global organized platform that could help researchers to study the virus, understand its evolution and localization, conduct a variety of modern mathematical and phylogenetic analyzes, automatically draw conclusions and visualize them for interpretation and decision making. Such a system could lead, as a result, to more effective actions for predicting the virus evolution. In this paper, we propose an online platform called TBEV analyzer which serves as an analytical tool in two directions - fundamental research on the virus evolution and the solution of practical problems related to public health. The platform has been developed specifically for virologists, molecular epidemiologists, evolutionists and other researchers who work with tick-borne encephalitis virus. The platform has two major benefits: first, it allows the user to concentrate on the result of analysis without concerning about computational aspects, and second, with reproducibility of system, it provides unique results associated with specific target query.**

***Index Terms*—tick-borne encephalitis, TBEV, clusteron structure, phylogenetic tree, online platform, alignment, evolution**

I. Introduction

Tick-borne encephalitis virus (TBEV) is a causative agent, a severe vector-borne neuroinfection, which causes encephalitis, paralysis, and death. The annual number of infected cases by TBEV is up to 14 000 in which fatality rate ranging from 0.03% to 20-35% depending on location [1], [2]. The TBEV genome is a positive single-stranded RNA molecule, approximately 11,000 bases in length, which encodes three structural (capsid, C; membrane, M, that is expressed as precursor prM; envelope, E) and seven non-structural proteins [3].

The main carrier of infection are ticks of the genus *Ixodes persulcatus* and *Ixodes ricinus*. Usually, the reservoirs of the virus are warm-blooded animals and birds [4]. The spread of TBEV is territorially considered to the north of Eurasia including the countries of Europe, Kazakhstan, Mongolia, Russian Federation, Japan and the north of China [5].

Generally, there are 3 well-known subtypes of the virus - European (TBEV-Eu), Siberian (TBEV-Sib), and Far-Eastern (TBEV-FE), which belong to the genus Flavivirus of the family Flaviviridae [6]. The name of the TBEV subtype is related to its geographical distribution. It is observed that climate changes, as well as the progress of human civilization, have impacts on tick's drift to other regions which may cause emerging new types of the virus [7].

An increase in the diversity of strains and their unusual localization in non-endemic regions requires constant monitoring and the pathogenicity assessment of new strains. Many characteristics of TBEV are poorly understood, but it is already clear that differences between virus subtypes play a role in its pathogenicity [8]. This emphasizes the importance of pattern identification which plays a crucial role in the prediction of the virus evolution direction and its potential threat.

Monitoring the population and studying the dynamics of the virus evolution is a key factor in predicting its distribution and evaluating the direction of evolution. In addition to developing laboratory methods for the detection and genotyping of TBEV [9], it is necessary to design and develop bioinformatics platforms for processing data obtained using laboratory analysis, as well as monitoring of the virus epidemiology. An example of such a system is Nextflu [10], which is designed and specified for the influenza virus. Among the various biological information and analytical systems, there are several systems for ticks. Ruzhnikov et al [11] presented a modern system that is created to assess the current state and predict the future distribution of invertebrate vectors of human vector-borne diseases. Another example is suggested by Molodorov et al [12], which is based on a domain ontology and specified for tick-borne disease. Among the well-known resources, it should be mentioned the entomologic platform VectorMap (vectormap.si.edu), which provides information about statistics

on the distribution of various types of ticks around the world.

Despite the aforementioned previous works and their good promising, there is no specific platform for phylogenetic analysis, tracking evolution and monitoring the epidemiology of TBEV. The platform proposed in this article, which is mainly based on the analysis of genetic sequence, is unique with respect to the methodology and considered purpose. The uniqueness of the developed platform stems from applying "clusteron approach" that functions as the analysis core. The rest of this section is devoted to the basics of the proposed platform.

Recently, "clusteron approach" has been developed [13], [14], which enables a researcher to determine the query strain subtype, phylogenetic lineage, and clusteron. The approach gives a direct meaningful description and illustration which includes the evolutionary history of TBEV. A clusteron is a group of TBEV strains with identical amino acid sequences of the E glycoprotein fragment, as a rule phylogeographically close, and having a certain type of territorial distribution [13]. Generally, clusterons (a term derived from cluster and clone) are clonal groups or the smallest systematic units of TBEV. Isolation of clusterons within the subtypes and phylogenetic lineages of the virus provides a clusteron graph structure which is illustrated in Fig.1. Obviously, the structure gives an overall picture of virus evolution and the relationship between strains which is necessary for the major of virus evolution research. As an example of clusteron structure application, through the structure analysis, it has been determined that the emergence of the European subtype of TBEV from the Siberian subtype occurred as a result of the host-jump phenomenon, i.e. changes by the virus of the arthropod host - tick *I.persulcatus* to *I.ricinus*.

In addition to the fundamental aspect, the clusteron approach is of great practical importance, since it serves as a tool for studying and monitoring the genetic structure of virus populations not only at the global scale but also at the regional and local levels. Each clusteron has its own unique profile, a genetic signature, which had been formed in certain historical and environmental conditions. For a more elaborate description of the clusteron approach and its concepts, the reader could refer to [13], [14].

Any change in the clusteron profile (the appearance of a new or disappearance of an already known clusteron, a change in clusteron size) may indicate the existence of certain processes that affect the genetic structure of the viral population. Understanding these processes is a priority for short-term forecasting of evolution behavior, as well as predicting the emergence of new epidemiologically important variants of the virus.

The essence of monitoring any viral population is a system for collecting, recording, storing and analyzing a small number of key indicators, taken periodically which characterize the population. Previous studies have shown the informative content of such a system [15]. The clusteron approach can be successfully used for real-time monitoring of the genetic structure of viral populations.

Despite all its advantages, the clusteron approach is not yet widely used due to its difficulties related to the order and configuration of its algorithms. Furthermore, to constantly update the cluster structure, the emerging strains of virus submitted in GenBank must be manually subjected to multistage analysis of the clusteron approach.

By reporting research results, scientists document the steps they followed in which they obtain the results so that others can verify and produce similar results for comparison. Such results are called reproducible [16]. The clusteron approach uses fairly well-known methods of bioinformatics, such as alignment, phylogenetic tree, and phylogenetic network construction. Since there are different algorithms and programs for the same analysis, for example, phylogenetic tree construction, the reproducibility of the clusteron approach is still questionable. Therefore, the goal of this project is focused on creating a standard, unified and interactive platform for analyzing and monitoring the virus population which would allow online updating of the TBEV cluster structure, combined with metadata such as strain location and isolation time, it would become an indispensable tool for studying the virus evolution. The platform can be accessed through tbev.magemagix.com.

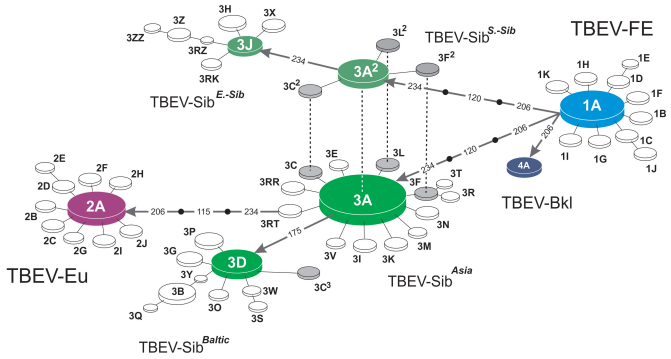

Fig. 1. The representation of clusteron structure of tick-borne encephalitis virus [14].

II. Methodology

Specifically, the proposed platform has been developed for virologists, molecular epidemiologists, evolutionists, and researchers of other specialties, who works on TBEV. It facilitates the application of clusteron approach and allows the researcher to concentrate only on the results of analysis obtained through the platform.

The platform requires applying a sequence of phylogenetic and mathematical methods, each of which needs a specific setting and configuration. There are some concerns related to the difficulties and complexity of the clusteron approach. As an example, it is known the topology of the phylogenetic tree structure is sensitive to the genetic distance method as well as the algorithm of tree construction from the genetic distance matrix. Since the obtained result from clusteron approach strongly depends on the type of its algorithms, in order to preserve the reproducibility of the platform, the main goal of the project is automatically set the necessary setting to obtain desired results.

Analytical platforms usually consist of three sections: data preparation and preprocessing, applying analytical methods, and finally the representation of the results, which can include visualization. The general scheme of the clusteron approach is illustrated in Fig.2.

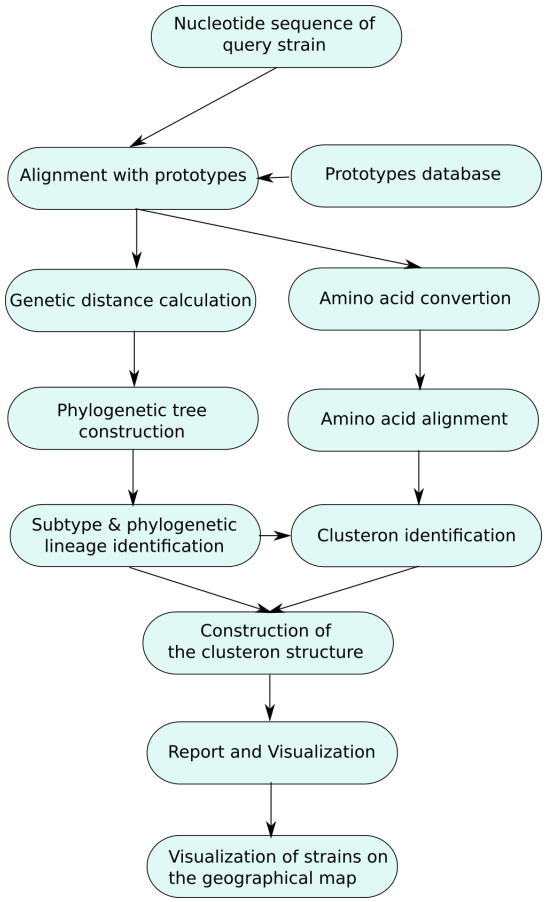

Fig. 2. The overall schema of the developed platform.

To implement this project, FLASK framework [17], which is used for creating web applications in the Python programming language, has been integrated with a developed visualization module called PhyloCat which is written in Javascript programming language. Details of the proposed platform are explained in the next sections.

A. Data preparation

As previously mentioned, the concept of a clusteron basically is determined by a fragment of amino acid sequence glycoprotein E, which consists of 151 amino acid residues (position 104 - 254 aa) and is encoded by 454 nucleotides (position 309 - 762 nt) part of the viral genome (gene E).

This fragment has been chosen as platform input for a number of reasons:

- Firstly, it contains both conservative and variable regions.
- Secondly, it includes unique amino acid substitutions at positions 175, 206 and 234, on the basis of which TBEV subtypes and phylogenetic lineages are distinguished.
- Thirdly, in GenBank, about 90% of the viral genome sequences include this fragment, which allows working with the maximum amount of data.
- Fourthly, the fragment length is a compromise between sufficient information content and the possibility of its amplification, which allows studying the maximum number of virus samples obtained per season from natural foci of TBEV, almost in real-time mode.
- Finally, the results of phylogenetic analysis, carried out using the selected fragment of the virus genome, are quite comparable in their informativeness with those of with complete genome sequences.

The mentioned fragment of the nucleotide sequence was determined and taken from the work [14] for all currently existing clusterons, which are called prototypes in the rest of this paper.

In order to simplify using the platform, a python algorithm was developed that automatically determines the location of the target fragment in the amino acid sequence of query strain through alignment with prototypes. Since the length of this fragment is conserved, if after alignment it becomes clear that contains deletion or insertion of nucleotide, the system refuses to continue further actions on the query.

To perform alignment and create a file in the FASTA format which contains fragments of prototypes, as well as the extracted fragment of the query strain, the well-known Biopython [18] and MUSCLE [19] packages, are applied. Biopython is a freely available python tool that aims to provide libraries for computational molecular biology and bioinformatics. MUltiple Sequence Comparison by Log- Expectation (MUSCLE) is a computer program of multiple alignment which can be applied to both nucleotide and protein sequences. MUSCLE was chosen for this project due to the high accuracy and speed of computation.

B. Phylogenetic Analysis

The majority of evolution researches deal with phylogenetic analysis. In our case, the phylogenetic analysis consists of two stages, which are carried out by the widely-used and well-known PHYLIP package [20], [21] explained as followings:

1) Genetic Distance Calculation: Since the fragments in the obtained FASTA format file from the previous section are closed together in terms of genetic distance, Kimura "2-parameter" model [22] is applied to calculate the genetic distance between strains of the file. The output is represented in a matrix-like shape in which the genetic distance between any pair strains of FASTA file is provided.

2) Tree construction algorithm: There are various methods for constructing a phylogenetic tree from a distance matrix. Due to the high computation speed and simplicity of the task, the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [23] was chosen for this part of the platform. UPGMA is a hierarchical clustering method which assigns the local topological relationship based on decreasing similarity that later is used to build the dendrogram. Despite conventional tree construction methods, UPGMA algorithm assumes

a constant substitution rate, over time and phylogenetic lineages (known as the molecular clock hypothesis) [24]. One prominent advantage of this method is its high computation speed which allows working with large trees. By applying PHYLIP package, a rooted UPGMA tree is built and saved into a Newick format file which will further be visualized.

C. Results and Visualization

As can be seen in Fig.1, the main challenge here is to declare the belonging of query strain to the clusteron structure in three different levels: subtype, phylogenetic lineage, and clusteron. Currently, there are four known TVEB's subtype [14]: European (TBEV-Eu), Siberian (TBEV-Sib), Baikalian (TBEV-Bkl), and Far-Eastern (TBEV-FE). Moreover, the Siberian subtype consists of four different phylogenetic lineages: Baltic (TBEV-Sib-Balt), Asian (TBEV-Sib-Asia), South-Siberian (TBEV-Sib-S.-Sib), and East-Siberian (TBEV-Sib-E.-Sib). The belonging of a given strain to a known clusteron is declared by constructing a nucleotide-based phylogenetic tree and verified through applying amino acid sequence alignment of their E gene fragment (see Fig.3).

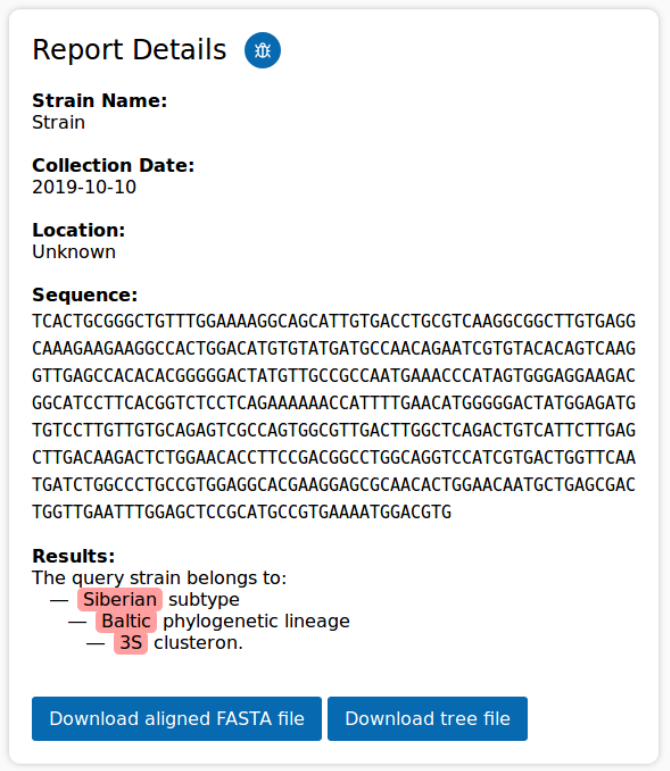

Fig. 3. The elementary report of result for a query strain.

In order to visualize the obtained phylogenetic tree on the web page, PhyloCat package was employed. PhyloCat is a Javascript package, developed by authors of this paper, that parses and converts the Newick file into a table data type which can simply be used to infer and determine the ownership and properties of the query strain in the tree. Additionally, PhyloCat is also able to visualize the phylogenetic tree, highlight the query strain path to the root, customize the subtype clade appearance, and annotate them. A typical visualization of the phylogenetic tree by PhyloCat is illustrated in Fig.4.

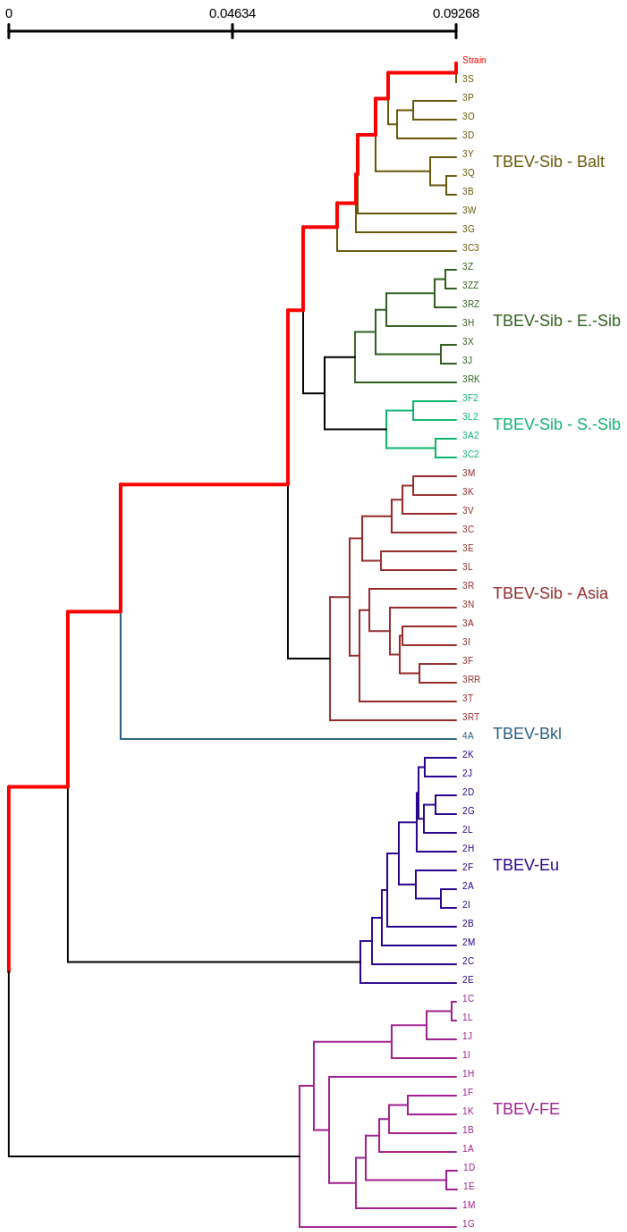

Fig. 4. The visualization of the phylogenetic tree by PhyloCat.

To facilitate the representation of alignment results, a Javascript module has been developed which visualizes the result in a table form and customizes the nucleotide base color (see Fig.5). The table also includes pair alignment score which serves as a similarity measure. Each table row represents a strain and rows are arranged in decreasing alignment score.

The visualization of the phylogenetic tree is configured so that the topmost branch belongs to the query strain. In order to conveniently identify the subtype or phylogenetic lineage closest to the given strain, subtrees associated with the sub-

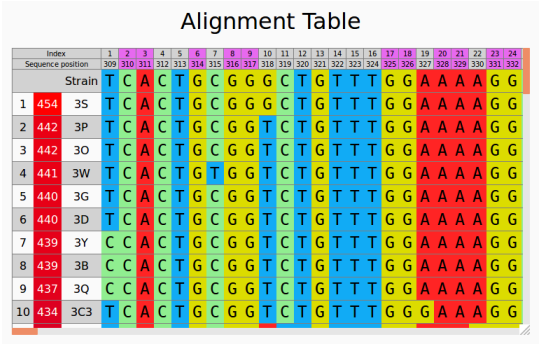

Fig. 5. The representation of alignment results.

types and lineages are uniquely colored. All clusterons name, phylogenetic lineages, and subtypes are separated in space and highlighted with the same color as their corresponding subtree (see Fig.4).

The conclusions obtained from the phylogenetic tree are necessary for the elementary assessment in which determines the belonging of the test strain to a multiscale clusteron structure. As mentioned, belonging of the query to a clusteron must satisfy two conditions:

- the query and clusteron must be sister taxa,
- they have similar amino acid fragment signature after alignment

In a special case, where the query strain does not belong to any known clusteron, the system informs the user and saves the information of the query in the server, since the emergence of new similar strains may lead to the formation of new clusteron.

However, the presented platform is the first stage of our project, which is planned to be expanded in the future due to the additional features described in the next section.

III. Future works

Generally speaking, the main task of this project is not only to determine the clusteron membership but to build and supplement the clusteron structure in order to find out the evolutionary history of the virus and also to understand which factors have an impact on the evolutionary dynamics of the virus. Therefore, a phylogenetic network can be applied as in [14], which describes the evolutionary history of TBEV in detail taking into account the clusteron structure. Consequently, the next step of the project can be the development of a method for building, customizing and visualizing a specialized phylogenetic network for TBEV by combining the advantages of both Fluxus (Fluxus-engineering.com) and PHYLOViZ [25] platforms which can well describe and characterize the clusteron structures.

The prevalence of TBEV strains is usually localized by geographical area. The detection of drift strains in new territories is an important task of virology and epidemiology. Therefore, one of the platform capabilities should be visualizing the collection location of query strain as well as known clusterons on the geographical map such as Google or Yandex map (see Fig.6).

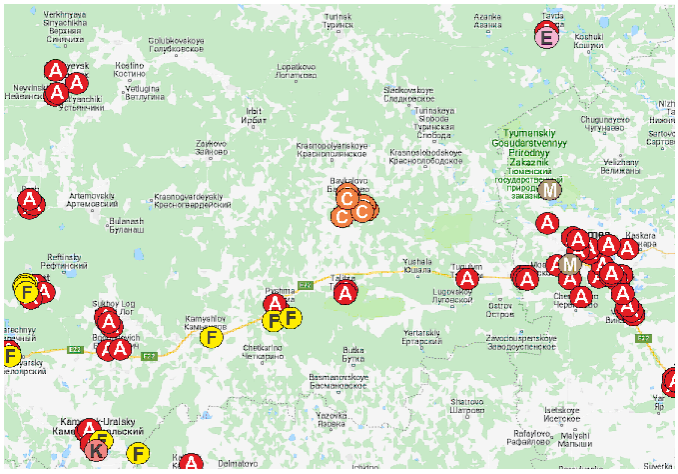

Fig. 6. The visualization of TBEV strains associated with clusterons on Google map [14].

Researchers, who work on TBEV, usually upload genomic data to the GenBank database [26]. Manually monitoring of the GenBank is labor-intensive and time-consuming. Thus, for the convenience of using the platform, an interface will be added which allows the user to extract the nucleotide sequence of GenBank strain using its ID in order to analyze it through the platform. The interface will also provide the ability to automatically update the cluster structure in real-time by monitoring TBEV sequences recently added to GenBank. It is remarkable that by submitting more TBEV samples, the overall picture of TBEV evolution becomes more and more clear.

Typically, antigens, which are located on the surface of the virus, contain important amino acid positions that play a crucial role in binding. Studying the amino acid substitutions in the structure of protein E and their relation to the clusteron structure can be facilitated by visualizing them on the surface of protein 3D model, e.g. PDB [27] file, which also helps the researcher better understand the biological role of mutations. Thus, it is planned to consider such visualization which highlights the amino acid signatures for each clusteron. A possible way to improve the quality of cognitive visualization is by employing a virtual reality environment. Accordingly, the 3D visualization of protein will be further equipped with a virtual reality environment by Viewlang platform (viewlang.ru).

Besides the aforementioned capabilities, the user's requested analysis also will be considered for future development.

IV. Conclusions

The implementation of the automatic analytical platforms for the genetic database has been gaining more and more interest in the last decade. With the accumulation of genetic sequence database and the development of technology, it becomes necessary to create specialized platforms that will allow researchers to automatically apply a required analysis and draw conclusions. Thus, the researcher can significantly

increase the efficiency of his work and concentrate his attention only on the obtained results.

In applying bioinformatics methods for processing genetic data, it is necessary to consider the reproducibility of the results, especially when the work relates to the study of evolution. Therefore, it is very important to consider this factor before developing analytical platforms.

This article introduces the project of creating a global unified interactive platform for monitoring and visualizing tick-borne encephalitis virus populations based on the recently developed clusteron approach. Clusteron approach addresses several factors such as subtype, phylogenetic lineage, and clusteron through which the strain can be characterized. In general, it can be said that the proposed platform will provide unique capabilities and a variety of analyzes for studying the evolution and behavior of TBEV which can especially be employed by virologists, molecular epidemiologists, and evolutionists. The main goal is performing the analysis and obtaining the results with minimal human effort. To further enhance the platform's practical significance, it is planned to add other capabilities such as visualization of results on a geographical map, such as Google and Yandex map, applying the phylogenetic network and considering an interface with GenBank.

It is expected that with the increase in the use of the presented platform, evolutionary information of TBEV will be continuously supplemented which updates the clusteron structure, as well as yields the extension and enhancement of analyzes for a better understanding of TBEV evolution. This also will lead to a comprehensive representation of the virus distribution on the geographical map.

REFERENCES

[1] V. Loktev, "Tick-borne encephalitis virus. genetic features and its variability in the modern world," *The Siberian Scientific Medical Journal*, no. 4, 2007.

[2] E. Chausov, V. Ternova, E. Protopopova, S. Konovalova, S. Kononova, N. Tupota, N. Moskvitina, V. Romanenko, N. Ivanova, N. Bolshakova, *et al.*, "Molecular genetic analysis of whole genome of the siberian subtype tick-borne encephalitis virus by the example of the modern isolate kolarovo-2008," *Problems of Particularly Dangerous Infections*, no. 4, 2011.

[3] T. J. Chambers, C. S. Hahn, R. Galler, and C. M. Rice, "Flavivirus genome organization, expression, and replication," *Annual review of microbiology*, vol. 44, no. 1, pp. 649–688, 1990.

[4] T. G. Jaenson, M. Hjertqvist, T. Bergström, and Å. Lundkvist, "Why is tick-borne encephalitis increasing? a review of the key factors causing the increasing incidence of human tbe in sweden," *Parasites & vectors*, vol. 5, no. 1, p. 184, 2012.

[5] D. Amicizia, A. Domnich, D. Panatto, P. L. Lai, M. L. Cristina, U. Avio, and R. Gasparini, "Epidemiology of tick-borne encephalitis (tbe) in europe and its prevention by available vaccines," *Human vaccines & immunotherapeutics*, vol. 9, no. 5, pp. 1163–1171, 2013.

[6] D. Ruzek, T. A. Županc, J. Borde, A. Chrdle, L. Eyer, G. Karganova, I. Kholodilov, N. Knap, L. Kozlovskaya, A. Matveev, *et al.*, "Tick-borne encephalitis in europe and russia: Review of pathogenesis, clinical features, therapy, and vaccines," *Antiviral research*, 2019.

[7] N. Tokarevich, A. Tronin, B. Gnativ, B. Revich, O. Blinova, and B. Evengard, "Impact of air temperature variation on the ixodid ticks habitat and tick-borne encephalitis incidence in the russian arctic: the case of the komi republic," *International journal of circumpolar health*, vol. 76, no. 1, p. 1298882, 2017.

[8] L. Pulkkinen, S. Butcher, and M. Anastasina, "Tick-borne encephalitis virus: a structural view," *Viruses*, vol. 10, no. 7, p. 350, 2018.

[9] L. Karan, S. Braslavskaya, and A. Myazin, "The development of methods for tick-borne encephalitis virus detection and genotyping based on amplification technologies," *Voprosy virusologii*, vol. 6, pp. 17–21, 2007, (in Russian).

[10] R. A. Neher and T. Bedford, "nextflu: Real-time tracking of seasonal influenza virus evolution in humans," *Bioinformatics*, vol. 31, no. 21, pp. 3546–3548, 2015.

[11] G. Ruzhnikov, G. Danchinova, R. Fedorov, M. Khasnatinov, V. Paramonov, and A. Lyapunov, "Modern technologies for informational and analytical evaluation of activity and forecast of spatial distribution of ixodid ticks by the example of irkutsk city," *Bulletin of the Siberian Branch of the Russian Academy of Medical Sciences*, vol. 32, no. 6, pp. 55–59, 2012, (in Russian).

[12] Y. I. Molorodov and D. A. Khodorchenko, "Development of the information system for tick-hazard on the basis of the objective ontology," *Vestnik NSU. Series: Information Technologie*, vol. 16, no. 4, p. 107–114, 2018, (in Russian) DOI 10.25205/1818-7900-2018-16-4-107-114.

[13] S. Y. Kovalev and T. A. Mukhacheva, "Clusteron structure of tick-borne encephalitis virus populations," *Infection, Genetics and Evolution*, vol. 14, pp. 22–28, 2013.

[14] S. Kovalev and T. Mukhacheva, "Reconsidering the classification of tick-borne encephalitis virus within the siberian subtype gives new insights into its evolutionary history," *Infection, Genetics and Evolution*, vol. 55, pp. 159–165, 2017.

[15] S. Y. Kovalev and T. A. Mukhacheva, "Clusterons as a tool for monitoring populations of tick-borne encephalitis virus," *Journal of medical virology*, vol. 86, no. 2, pp. 283–289, 2014.

[16] V. Stodden, "The scientific method in practice: Reproducibility in the computational sciences," 2010.

[17] A. Ronacher, "Flask (a python microframework)," *URL: http://flask.pocoo.org (visited on 10/09/2019)*, 2019.

[18] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython: freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[19] R. C. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[20] J. Felsenstein, "Phylip-phylogeny inference package (version 3.2) cladistics 5: 164–166," 1989.

[21] "Phylip (phylogeny inference package) version 3.6. distributed by the author. department of genome sciences," 2005.

[22] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *Journal of molecular evolution*, vol. 16, no. 2, pp. 111–120, 1980.

[23] R. R. Sokal, "A statistical method for evaluating systematic relationships," *Univ. Kansas, Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.

[24] M. Weiß and M. Göker, "Molecular phylogenetic reconstruction," in *The Yeasts*, pp. 159–174, Elsevier, 2011.

[25] A. P. Francisco, C. Vaz, P. T. Monteiro, J. Melo-Cristino, M. Ramirez, and J. A. Carriço, "Phyloviz: phylogenetic inference and data visualization for sequence based typing methods," *BMC bioinformatics*, vol. 13, no. 1, p. 87, 2012.

[26] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "Genbank," *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.

[27] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The protein data bank nucleic acids research, 28: 235-242," *URL: www. rcsb. org Citation*, 2000.