Recognition of Herpes Viruses on the Basis of a New Metric for Protein Sequences

V Sulimova\textsuperscript{1}, O Seredin\textsuperscript{1} and V Mottl\textsuperscript{2}

1. Tula State University, Lenin Ave., 92, Tula, 300012, Russia
2. Federal Research Center “Computer Science and Control” of RAS, Moscow, Russia

vsulimova@yandex.ru

Abstract. This paper addresses the problem of intellectual human herpes viruses recognition based on the analysis of their protein sequences. To compare proteins, we use a new dissimilarity measure based on finding an optimal sequence alignment. In the previous work, we proved that the proposed way of sequence comparison generates a measure that has properties of a metric. These properties allow for more convenient and effective use of the proposed measure in further analysis in contrast to the traditional similarity measure, such as Needleman-Wunch alignment. The results of herpes viruses recognition show, that the metric properties allow to improve the classification quality. In addition, in this paper, we adduce an updated computational scheme for the proposed metric, which allows to speed up the comparison of proteins.

1. Introduction

Human herpes viruses are pathogens that establish lytic and latent infections. These viruses usually are not life-threatening, but in some cases they might cause serious infections of eyes and brain that can lead to blindness and, possibly, death. An effective drug (acyclovir and its derivatives) is available against these viruses. Therefore, early detection and identification of these viral infections is highly important for an effective treatment [1,2].

In this work, we propose the way to recognize herpes viruses basing only on their protein sequences. It is evident that in this case the sequence comparison quality plays an important role for the resulting quality of recognition.

Traditionally, comparison of proteins consists in computing similarity measures based on finding an optimal pair-wise alignment [3–6]. But they don’t provide the possibility for using advantages of popular and effective linear methods initially designed for feature spaces, like SVM for the two-class pattern recognition problem [7].

Partially, the problem of harnessing linear methods in featureless situations can be solved by constructing special similarity measures called positive definite kernel functions [8–10], which embed a set of proteins into some hypothetical linear space and play the role of inner product in it [10]. But constructing mathematically correct and at the same time biologically motivated kernel functions is, as a rule, theoretically and computationally hard problem [10–12]. Besides, there are entire continuous classes of kernels, which are equivalent from the viewpoint of the final decision rule [13].

At the same time, obviously, it is not the feature vectors of entities in some linear space that are the actual basis of machine learning and data mining algorithms but rather the respective metric, i.e., pairwise distance between entities [14]. In this connection there is a natural desire to use comparison measures possessing metric properties, especially since a metric allows for embedding any set of...
proteins into a linear space, and apply linear methods in it in accordance with the generalized linear approach to dependence estimation [15].

However, the dependence-estimation methods effectiveness crucially depends on the choice of the metric between entities, which has to satisfy the compactness hypothesis [16,17]. This means that the values of the accepted metric between the proteins performing the same function in a living organism must be small and, respectively, they must be large for proteins that perform different functions.

We know a number of ways to introduce metrics on a set of sequences [14], but in the case of amino acid sequences none of them has satisfactory interpretation from the biological point of view. In this connection, it is unlikely that the compactness hypothesis would hold true in the respective metric space of proteins. This point has multiple confirmations in practice [18,19], which gave rise to an entire series of papers aimed at improving the initial metric via various algebraic constructions of different complexity rate (Metric Learning), [17–21], including constructing metrics on the basis of kernel functions (Metric Kernel Learning) and constructing secondary features or new metrics on the basis of some similarity measures. Also, a number of papers tries to involve structural and protein-protein interaction information [22,23]. However, in this paper our goal is to provide recognition of human herpes viruses via constructing appropriate metrics for protein sequences without involving any additional information.

In this paper, we describe a simple way of constructing a mathematically correct metric on proteins set. This method follows the traditional Needleman-Wunsch algorithm, it is based on finding a global optimal alignment of sequences, and leans upon the probabilistic model PAM (Point Accepted Mutation) [24] of single amino acids. But, at the same time, it differs from the traditional approaches in the optimization criterion and in the way of comparing amino acids. In the previous paper [26], we proved that the proposed comparison measure possesses the properties of a metric. These properties allow for more convenient and effective use of the initial comparison measure.

In this paper, we adduce an updated computational scheme for the proposed metric, which allows to speed up the process of protein comparison. It is essential for real data processing.

This paper contains a more detailed experimental study of the proposed metric. Our results have shown that the metric’s properties allow for improving the recognition quality of herpes viruses as distinct from the traditional Needleman-Wunsch procedure.

2. Comparison of protein sequences

2.1. A metric on the set of protein sequences

Let \( \Omega \) be the set of all protein sequences. In this work, we consider only the primary structures of proteins, thus, we consider sequences over the alphabet of 20 known amino acids \( A = \{\alpha_1^*, \ldots, \alpha_m^*\} \), \( m = 20 \).

Let also \( \omega' = (\alpha_1', \alpha_2', \ldots, \alpha_N') \in \Omega \) and \( \omega'' = (\alpha_1'', \alpha_2'', \ldots, \alpha_N'') \in \Omega \) be two sequences of lengths \( N' \) and \( N'' \) that contain amino acids \( \alpha_i', \alpha_j'' \in A \), \( i = 1, \ldots, N' \), \( j = 1, \ldots, N'' \).

It is evident that any protein sequences comparison should be based on the comparison of amino acids forming them. The main theoretical concept underlying the proposed amino acids comparison method is the well-known probabilistic model of amino acids evolution by Margaret Dayhoff, called PAM (Pointed Accepted Mutation) [24]. Its main instrument is the notion of the Markov chain of amino acids evolution at some single point of protein’s chain, which is represented by the matrix of transitional probabilities \( \Psi = \left( \psi_{ij}(\alpha_j' | \alpha_i') \right) \) of changing amino acid \( \alpha_i' \) into amino acid \( \alpha_j' \) at the next step of evolution. The index “1” in the brackets means that the initial one-step Markov chain is considered.

In accordance with the PAM model, it is supposed that this Markov chain is an ergodic and reversible random process, i.e., a process that possesses a final probability distribution
\[ \sum_{\alpha' \in \mathcal{A}} \xi(\alpha') \psi_{|\mathcal{A}|} (\alpha' | \alpha^j) = \xi(\alpha^j) \]

and satisfies the reversibility condition
\[ \xi(\alpha^i) \psi_{|\mathcal{A}|} (\alpha^i | \alpha^j) = \xi(\alpha^j) \psi_{|\mathcal{A}|} (\alpha^i | \alpha^j) \]

Let us consider the probabilistic process of evolution with a greater evolutionary step \( s > 1 \), i.e., a sparse Markov chain with the matrix of transitional probabilities \( \Psi_{|s|} = [\Psi_{|1|} \times \cdots \times \Psi_{|1|}]_s \). Previously we proved \([10]\) that, for any \( s \), similarity measures
\[ \kappa_s (\alpha^i, \alpha^j) = \frac{\psi_{|\mathcal{A}|} (\alpha^i | \alpha^j)}{\xi(\alpha^j)} \]

form nonnegative definite matrices of amino acids pairwise similarity. So, each of them is kernel function that embeds the set of amino acids \( \mathcal{A} \) into a respective hypothetical linear space \( \tilde{\mathcal{A}} \subset \mathcal{A} \) with Euclidean metric \([25]\)
\[ r(\alpha^i, \alpha^j) = \left( \kappa(\alpha^i, \alpha^j) + \kappa(\alpha^j, \alpha^i) - 2 \kappa(\alpha^i, \alpha^j) \right)^{1/2} \]  

(1)

This is the same metric we use here to compare amino acids. Hereafter, we shall omit lower index \( \rho(\alpha^i, \alpha^j), \kappa(\alpha^i, \alpha^j) \) assuming that the evolutionary step \( s > 1 \) is predefined.

Further, we define a metric on the amino acid sequences set on the basis of global pairwise sequence alignment \( \mathbf{w}(\omega', \omega'') \), which is understood as a way of sequence transformation by inserting gaps in some positions of aligned sequences for reducing them to a common length with preserving correspondences between their elements.

If the position \( w_i, i = 1, \ldots, |w| \), of an alignment \( \mathbf{w}(\omega', \omega'') \) of two sequences \( \omega' = (\alpha_1', \ldots, \alpha_{N'}') \in \Omega \) and \( \omega'' = (\alpha_1'', \ldots, \alpha_{N''}) \in \Omega \) does not contain a gap, it explicitly defines two aligned amino acids \( (\alpha^*, \alpha^*) \).

We define the comparison measure of any two sequences as the optimal value of the specific optimization criterion
\[ r(\omega', \omega'') = \min_{\mathbf{w}} \sqrt{\sum_{i=1}^{|w|} \left[ I(w_i) \beta^2 + (1 - I(w_i)) \rho^2 (\alpha^*, \alpha^*) \right]} \]

(2)

where \( I(w_i) = 1 \) if the \( i \)-th position of the alignment \( \mathbf{w} \) contains a gap and \( I(w_i) = 0 \) otherwise, and coefficient \( \beta \) in (2) has the meaning of gap penalty. In our previous work \([26]\) we proved that for any
\[ \beta \geq 0.5 \max_{\alpha^i, \alpha^j} \rho(\alpha^i, \alpha^j), \quad \forall \alpha \in \mathcal{A} \]

(3)

the comparison measure (2) possesses all the properties of a metric.

2.2. Dynamic programming procedure for computing the metric on the set of protein sequences

The criterion (2) falls into the class of pairwise separable goal functions. A minimum of such function can be found using dynamic programming procedure, which is similar to the Needleman-Wunsch procedure \([3]\) for finding the optimal global alignment of any two amino acid sequences, which maximizes their similarity.
The idea of this algorithm consists in recurrent computing of unknown dissimilarity values $F_{i,j}$ for growing fragments of two protein sequences $(\alpha'^1_1,\alpha'^1_2,\ldots,\alpha'^1_i)$ and $(\alpha''^1_1,\alpha''^1_2,\ldots,\alpha''^1_j)$ on the basis of dissimilarity values, which are assumed to be already computed:

$$F_{i,j} = \min \left\{ F_{i-1,j-1} + \rho^2(\alpha'^1_i,\alpha''^1_j); \quad F_{i-1,j} + \beta^2; \quad F_{i,j-1} + \beta^2, \quad i = 1,\ldots,N'; \quad j = 1,\ldots,N'' \right\} \quad (4)$$

The computation starts from the initialization:

$$F_{0,0} = 0; \quad F_{i,0} = i\beta^2, \quad i = 1,\ldots,N'; \quad F_{0,j} = j\beta^2, \quad j = 1,\ldots,N''$$

and is finished, when the end of sequences is met: $F_{N',N''}$.

It is convenient to represent such a computation process as a table of pairwise correspondences (figure 1).

![Figure 1: A table of pair-wise correspondences between elements of two sequences being compared: an illustration of the computation process (left) and a possible optimal alignment (right).](image)

The computation process consists in consecutive passing through all the cells of the table (figure 1, left), from the top left cell to the right bottom one, making recurrent computations of incomplete dissimilarity values $F_{i,j}$, choosing (and possibly saving) the direction of the optimal movement to the current cell (horizontal, vertical or diagonal), which are to be used later to find the optimal alignment (figure 1, right).

3. **Speeding-up the metric computation**

3.1. **Speeding-up the metric computation for two protein sequences**

It is evident that the computational complexity of the dynamic programming procedure described in the previous section is proportional to the product $N_0N_00$. So, the computing time is essentially increasing, when the lengths of sequences increase. As a result, there is the problem of applying the described algorithm to real data, which are usually represented by big sets of long protein sequences. This is a well-known problem, and there is a number of heuristic realizations of similar dynamic programming procedures for comparing biological sequences (such as BLAST, FASTA, and etc.) [27–29], that allow to obtain a fast but approximate solution to the dynamic programming problem.

To speed-up the comparison of protein sequences, we adapt here the approach that was proposed for comparing discrete signals with the purpose of speech recognition [30,31].

In accordance with this approach, we find the optimal alignment among only those alignments, which fully fall into a pruned subtable of pairwise alignments (figure 2).
The respective modified recurrent computation scheme can be expressed in the form

\begin{equation}
F_{i,j} = \min \left\{ F_{i-1,j-1} + \rho^2 (\alpha_i', \alpha_j'); \right. \\
F_{i-1,j} + \beta^2; \\
F_{i,j-1} + \beta^2, \\
\left. i = 1, \ldots, N'; j = j_{\text{start}}(i), \ldots, j_{\text{end}}(i), \right. \end{equation}

where \( j_{\text{start}}(i) = \max \left( 1, \left\lfloor \frac{N^* i - N^*}{t} \right\rfloor \right) \), \( j_{\text{end}}(i) = \min \left( N^*, \left\lfloor \frac{N^* i + N^*}{t} \right\rfloor \right) \).

It should be noticed that, in the strict sense, the proposed approach does not guarantee finding the optimal value of the criterion (2). But, at the same time, the strip width parameter \( t \) allows to control the number of alignments under consideration and, thus, to balance the accuracy and the speed of computation. Our experiments with real protein sequences of Herpes Viruses, described in Sect. 3 of this paper, have shown that the initial algorithm and its approximations with \( t = 1, \ldots, 1.5 \) have absolutely the same results for all the pairs of 1532 considered proteins, the lengths of which range from 36 to 1378 amino acids.

Filling up the pruned table is simple enough, which results in increasing the computation speed even for essentially different sequences, in contrast to a more intellectual approach that was proposed for signals [32].

### 3.2. Speeding-up the computation of a metric values matrix for proteins set

Let \( \Omega = \{ \omega_1, \ldots, \omega_K \} \) be some set of \( K \) protein sequences to be analyzed. It should be noticed, that for many problems of protein analysis including herpes virus recognition, it is needed to compute the entire metric values matrix for the given set of proteins. However, due to the metric’s symmetry, it is enough to compute the triangular matrix of metric values.

Moreover, it is easy to see that all the elements of such a matrix could be computed independently. Following the latter remark, we involve the technology of parallel computing OpenMP for further computation speeding-up. We have used a simple enough scheme of parallel computing, according to which for each row containing more than \( h \) elements, \( p \) threads have been created for parallel computing elements of the respective row.

### 4. Experiments

#### 4.1. Data description
For the experiments, we have used six sets of herpes virus amino acid sequences that perform functions specified in Table 1. These are the sets from the VIDA database (Virus Database at University College London) [33]. Proteins of each set have been partitioned into classes and homologous protein families on the basis of laboratory research of herpes viruses [2].

| Set (num. of proteins) | Function                                      | Class (num. of proteins) | Description                           | Homologous Protein Families (HPFs) |
|-----------------------|-----------------------------------------------|--------------------------|---------------------------------------|-----------------------------------|
| 1 (233)               | Membrane/ Glycoprotein                        | 1 (109)                  | Glycoprotein H                        | 12, 42, 531                       |
|                       |                                               | 2 (76)                   | Glycoprotein L                        | 47, 50, 114, 256                  |
|                       |                                               | 3 (48)                   | Glycoprotein M                        | 20                                |
| 2 (407)               | Nucleotide/ repair metabolism                 | 1 (256)                  | Thymidine kinase                      | 2, 27                             |
|                       |                                               | 2 (83)                   | Alkaline exonuclease                   | 11, 51                            |
|                       |                                               | 3 (37)                   | Ribonucleotide reductase              | 33                                |
|                       |                                               | 4 (31)                   | dUTPase                               | 43                                |
| 3 (262)               | Virion Assembly                               | 1 (54)                   | Transport/ capsid assembly protein     | 7                                 |
|                       |                                               | 2 (92)                   | DNA packaging protein                 | 18, 22                            |
|                       |                                               | 3 (77)                   | Cleavage/ packaging protein           | 34, 39                            |
|                       |                                               | 4 (20)                   | Packaging and capsid formation        | 79                                |
|                       |                                               | 5 (19)                   | DNA packaging and capsid formation    | 108                               |
| 4 (99)                | Enzyme                                        | 1 (89)                   | Protein kinase                        | 29, 40                            |
|                       |                                               | 2 (10)                   | Phospholipase-like protein            | 328, 329                          |
| 5 (144)               | DNA replication                               | 1 (52)                   | Origin binding protein                | 5, 152                            |
|                       |                                               | 2 (22)                   | DNA polymerase processivity factor    | 104, 1003                         |
|                       |                                               | 3 (48)                   | Helicase/ primase associated protein  | 16                                |
|                       |                                               | 4 (22)                   | Component of DNA helicase/ primase complex | 72                               |
| 6 (195)               | Virion protein                                | 1 (47)                   | Virion tegument protein               | 21                                |
|                       |                                               | 2 (28)                   | Tegument protein / FGARAT             | 44                                |
|                       |                                               | 3 (21)                   | Tegument phosphoprotein               | 65                                |
|                       |                                               | 4 (91)                   | Tegument protein                      | 83, 86, 87, 93                    |
|                       |                                               | 5 (29)                   | Virion protein                        | 62, 106                           |

Besides, for some experiments we used an additional set of 143 protein sequences, which are not herpes viruses. These proteins were randomly chosen from the data set collected by Lanckriet et al [34]. The resulting data set contains 1532 protein sequences of essentially different length in the range from 36 to 1378 amino acids.

4.2. Experimental investigation of metric computing efficiency

For experimental investigation of metric computing efficiency, we compared the initial dynamic programming procedure (section 2.2) and its approximation with different values of the width parameter t (section 2.3), where t = 1 corresponds to the initial full-width procedure. Additionally, each of these procedures was tested with 1, 2, and 4 threads on the personal computer of the following configuration: Intel Core i5-4210U CPU 1.70 GHz, 2 processor cores with the Hyper Threading, 6Gb RAM, 3Mb L3 cache.

The time for computing the whole matrix of the metric values for 1532 protein sequences was measured in minutes 5 times for each set of parameters. The average results are presented in Table 2.

It is evident that the speed increases when the value t is increasing. Nevertheless, we have found that the initial algorithm and its heuristic approximations with t = 1,...,1.5 showed absolutely the same results.
for all the pairs of 1532 proteins studied in the experiment. So, the proposed modifications of the computation scheme allow for decreasing the computation time from 82.01 to 28.74 minutes without accuracy loss (for \( t = 1.5 \) and \( p = 4 \) threads). The acceleration in this case is about 2.85. This result is good enough for the architecture with only 2 real processor cores.

### Table 2. Average time (in minutes) for computing the matrix of the metric for 1532 proteins for different computation schemes and parameters of the computation.

| Number of threads (\( p \)) | \( t = 1 \) | \( t = 1.2 \) | \( t = 1.5 \) | \( t = 2 \) |
|-----------------------------|------------|------------|------------|------------|
| \( p = 1 \)                | 82.01      | 74.64      | 65.57      | 53.44      |
| \( p = 2 \)                | 42.15      | 39.27      | 35.43      | 30.61      |
| \( p = 4 \)                | 34.61      | 31.23      | 28.74      | 23.29      |

### 4.3. Investigation of metric properties usability for Herpes Viruses recognition

#### 4.3.1. Constructing kernel functions on the basis of comparison measures

The experiment’s main purpose is to demonstrate that the presence of comparison measure metric properties allows to increase the recognition quality. So, for this purpose, we have compared two very similar comparison measures, each of them is based on the global optimal pair-wise alignment of sequences - the Needleman-Wunch (NW) alignment and the proposed metric. These two comparison measures have similar structures, but while the first is a similarity measure, the second one possesses the metric properties.

In our experiments, we used NW algorithm from the MATLAB bioinformatics toolbox with both PAM250 and BLOSSUM62 substitution matrixes for comparing single amino acids and with the gap penalties \( g = 12 \) for start and \( b = 1 \) for continuation of gaps series.

The metric on the set of amino acids was constructed in accordance with (1) on the basis of PAM model with the evolutionary step \( s = 250 \). The parameter \( \beta \) of the proposed metric was set to \( \beta = 0.0234 \) for all the experiments.

The proposed metric, as it turned out, is the Euclidean metric for the considered sets of proteins. This fact gave us the possibility to apply the radial kernel function \( K_{\text{tr}} = K_{\text{tr}} (\omega', \omega'') = \exp(-\alpha r^2(\omega', \omega'')) \) \([9,23]\). But it should be noticed, that, in general case, the proposed way of metric construction doesn’t guarantee the Euclidean property of the proposed metric and, thus, such transformation can lead to the presence of negative eigenvalues of the respective kernel matrix for proteins set. But the practice shows that the Euclidean property usually holds true.

Also we have constructed linear kernel functions in the space of the respective secondary features \( K_{\text{tr}} = R_{\text{tr}}^T R_{\text{tr}} \), where \( R_{\text{tr}} = \{ r(\omega_i, \omega_j) \}, i,j = 1, \ldots, 1532 \) is the matrix of metric values.

As to the NW similarity measure, the respective radial function was not used for it at all, because it is a similarity measure, and, moreover, it often yields negative values. So, any heuristic transformation of such similarity measure into a metric requires moving its values into the positive range. This fact requires having a full set of proteins at the training stage, but usually it is impossible in practice. In this connection, for the NW similarity measure we constructed linear kernel functions in the space of the respective secondary features \( K_{\text{SF}}^{\text{NW}} = S_{\text{SF}}^{\text{T}} S_{\text{NW}} \), where \( S_{\text{NW}} \) is a NW protein similarity matrix.

#### 4.3.2. Recognition of classes and HPFs of Herpes viruses

For this experiment we used two sets of Herpes viruses amino acid sequences that perform, respectively, the following functions: "Membrane/glycoprotein" and "Nucleotide repair/metabolism". It is important to distinguish these classes and HPFs from one another, the more so because previous investigations showed them to be most problematic ones for recognition \([2]\).

For each indicated way of proteins comparison and for each of the two considered proteins sets, a number of two-class pattern recognition problems were solved (one-against-all and one-against-one recognition for classes and for HPFs). In each case, the training was made via SVM \([7]\). The quality of
obtained decision rules was estimated by the leave-one-out cross validation (LOO) procedure. Table 3 contains LOO-error percentages only for the cases with results that differ at least for two algorithms. Besides, the results obtained for NW with PAM250 and BLOSUM62 are practically the same in these experiments. In this connection, we included only one row for NW algorithm, which corresponds to the BLOSUM62 substitution matrix.

Table 3. LOO percentages for recognition of the most interesting HPFs and classes.

| Membrane/glycoprotein | One-against-all HPFs | Classes | One-against-one (HPFs) |
|-----------------------|---------------------|---------|-----------------------|
|                       | 12                  | 20      | 47                    | 50       | 114      | 1       | 2       | 3       | 531/32   | 531/47   | 531/12   |
|                       |                     |         |                       |          |          |         |         |         |          |          |          |
| $K_{NW}^{SF}$         | 15.02               | 0.43    | 4.72                  | 0.43     | 4.72     | 0.43    | 0.43    | 0.43    | 6.45     | 4.17     | 48.06    |
| $K_{mtr}^{SF}$        | 15.40               | 0.00    | 0.00                  | 0.00     | 0.43     | 0.43    | 0.00    | 3.22    | 2.08     | 50.00    |
| $K_{mtr}^{0.01}$      | 14.59               | 0.00    | 0.00                  | 0.43     | 0.43     | 0.43    | 0.00    | 3.22    | 2.08     | 50.00    |

Nucleotide repair/metabolism

| Comparing measure | One-against-all HPFs | classes | One-against-one HPFs | classes |
|-------------------|---------------------|---------|---------------------|---------|
| $K_{NW}^{SF}$     | 14.25               | 0.49    | 0.49                | 1.91    | 1.91     | 0.59    |
| $K_{mtr}^{SF}$    | 14.49               | 0.49    | 0.25                | 0.64    | 0.64     | 0.30    |
| $K_{mtr}^{0.01}$  | 14.09               | 0.25    | 0.25                | 0.64    | 0.64     | 0.30    |

4.3.3. Recognition of Herpes Viruses proteins performing a specified function

In this experiment, we solved 12 two-class protein recognition problems: tasks 1-6 for recognition each of 6 proteins sets (table 1) from not Herpes viruses, and tasks 7-12 for recognition each of 6 sets from all other proteins. The classical SVM was used for constructing decision rules. The quality of decisions was estimated by ROC-scores computed for 10-times cross-validation with random forming training and testing sets in the proportion 20:80. The results are presented in Table 4. The best result for each task is shown in bold.

Table 4. Average ROC scores for 12 tasks of recognition Herpes viruses.

| Comparing measure | tasks |
|-------------------|-------|
| $K_{NW}^{SF}$     | 1     | 0.923 | 0.823 | 0.758 | 0.932 | 0.804 | 0.856 | 0.856 | 0.973 | 0.935 | 0.915 | 0.938 | 0.927 |
| $K_{mtr}^{SF}$    | 0.909 | 0.848 | 0.747 | 0.972 | 0.790 | 0.856 | 0.874 | 0.968 | 0.978 | 0.882 | 0.950 | 0.933 |
| $K_{mtr}^{0.5}$   | 0.926 | 0.926 | 0.883 | 0.966 | 0.894 | 0.887 | 0.960 | 0.970 | 0.999 | 0.997 | 0.990 | 0.991 |
| $K_{mtr}^{0.2}$   | 0.962 | 0.942 | 0.907 | 0.985 | 0.927 | 0.929 | 0.966 | 0.973 | 1.000 | 0.997 | 0.994 | 0.996 |
| $K_{mtr}^{0.1}$   | 0.966 | 0.948 | 0.898 | 0.981 | 0.917 | 0.937 | 0.959 | 0.970 | 0.998 | 0.996 | 0.993 | 0.993 |
| $K_{mtr}^{0.01}$  | 0.965 | 0.944 | 0.870 | 0.975 | 0.826 | 0.935 | 0.943 | 0.969 | 0.995 | 0.994 | 0.987 | 0.986 |
| $K_{mtr}^{0.001}$ | 0.965 | 0.943 | 0.868 | 0.974 | 0.824 | 0.935 | 0.942 | 0.970 | 0.995 | 0.994 | 0.987 | 0.985 |

The obtained results have shown that the use of proposed metric instead of the traditional NW-similarity measure allows for essential increasing in the Herpes viruses recognition quality of. Of course, it should be noticed, that we have the structural parameter $\alpha$, so we need some instrument for automatic selection of its appropriate value. But this aspect is outside the scope of this investigation.
5. Conclusions and discussion
This paper proposes a proteins comparison measure based on the optimal global alignment. The proposed measure is very similar to the traditional Needleman-Wunsch similarity measure, but, in contrast to it, possesses metric properties. Also, we have proposed some ways for speeding-up its computation. The obtained results have shown that the use of proposed metric instead of the traditional NW similarity measure allows for essentially increasing the quality of Herpes viruses recognition.

6. References
[1] Huleihel M, Shufan E, Zeiri L and Salman A 2016 Detection of vero cells infected with herpes simplex types 1 and 2 and varicella zoster viruses using raman spectroscopy and advanced statistical methods PLoS One 11(4) e0153599 DOI: 10.1371/journal.pone.0153599
[2] Mc Geoch D J, Rixon F J and Davison A J 2006 Topics in herpesvirus genomics and evolution Virus Res. 117 90-104 DOI: 10.1016/j.virusres.2006.01.002
[3] Needleman S B and Wunsch C D 1970 A general method applicable to the search for similarities in the amino acid sequence of two proteins Journal of Molecular Biology 48(3) 443-453 DOI: 10.1016/0022-2836(70)90057-4
[4] Smith T F and Waterman M S 1981 Identification of common molecular subsequences Journal of Molecular Biology 147(1) 195-197 DOI: 10.1016/00222836(81)90087-5
[5] Zhang Z, Schwartz S, Wagner L and Miller W 2000 A greedy algorithm for aligning DNA sequences Journal of Computational Biology 7(1-2) 203-214 DOI: 10.1089/1066527050081478
[6] Durbin R, Eddy S R, Krogh A and Mitchison G 1998 Biological sequence analysis: probabilistic models of proteins and nucleic acids (Cambridge Univ. Press) p 356
[7] Vapnik V N 1998 Statistical learning theory (Wiley-Interscience) p 768
[8] Schölkopf B, Tsuda K and Vert J-P 2004 Kernel methods in computational biology (MIT Press) p 410
[9] Aizerman M A 1970 Potential functions method in machine learning theory (Moscow: Nauka) p 384
[10] Sulimova V V 2009 Kernel functions for analysis of signals and symbolic sequences of different length (PhD thesis) p 122
[11] Miklós I, Novak A, Satija R, Lyngso R and Hein J 2009 Stochastic models of sequence evolution including insertion-deletion events Statistical Methods in Medical Research 18(5) 453-485 DOI: 10.1177/0962280208099500
[12] Seeger M 2002 Covariance kernels from bayesian generative models Adv. Neural Inform. Process. Syst. 14 905-912
[13] Abramov V I, Seredin O S and Mottl V V 2013 Pattern recognition training by support object method in Euclidean metric spaces with affine operations Proceedings of Tula state university Natural sciences series 2(1) 119-136
[14] Pekalska E M 2005 Dissimilarity representations in pattern recognition. Concepts, theory and applications (PhD Thesis) p 344
[15] Seredin O S and Mottl V V 2015 Support object method for pattern recognition training in arbitrary metric spaces Proceedings of Tula State University Natural Sciences Series 4 178-196
[16] Braverman E M 1961 Experiments on training a machine for pattern recognition (PhD Thesis Moscow)
[17] Xing E P, Ng A Y, Jordan M I and Russel S 2003 Distance metric learning with application to clustering with side-information Advances in Neural Information Processing Systems 15 521-528
[18] Bellet A, Harbrad A and Sebban M 2013 A survey on metric learning for feature vectors and structured data CoRR Arxiv 1306.6709
[19] Wang J, Sun K, Sha F, Marchand-Maillet S and Kalousis K 2014 Two-stage metric learning Proceedings of the 31st International Conference on Machine Learning 2(32) 370-378
[20] Schultz M and Joachims T 2004 Learning a distance metric from relative comparisons Adv. Neural Inform. Process. Syst. 16 41-48
[21] Wang J, Do H, Woznica A and Kalousis A 2011 Metric learning with multiple kernels Adv. Neural Inform. Process. Syst. Curran Associates Inc. 24 1-9

[22] Cao M, Zhang H, Park J, Daniels N M and Crovella M E 2013 Going the distance for protein function prediction: a new distance metric for protein interaction networks PLoS ONE 8(10) e76339 DOI: 10.1371/journal.pone.0076339

[23] Rogen P and Fain B 2002 Automatic classification of protein structure by using Gauss integrals Proc. Natl. Acad. Sci. USA 100(1) 119-124 DOI: 10.1073/pnas.2636460100

[24] Dayhoff M, Schwarts R and Orcutt B 1978 A model of evolutionary change in proteins Atlas of Protein Sequences and Structures 5(3) 345-352

[25] Mottl V V 2003 Metric spaces admitting linear operations and inner product Reports of Academy of Sciences 67(1) 140-143

[26] Sulimova V, Seredin O and Mottl V 2016 Metrics on the basis of optimal alignment of biomolecular sequences JMLDA 2(3) 286-304 DOI: 10.21469/22233792.2.3.03

[27] Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 Basic local alignment search tool J. Mol. Biol. 215(3) 403-410 DOI: 10.1006/jmbi.1990.9999

[28] Lipman D J and Pearson W R 1985 Rapid and sensitive protein similarity searches Science 227(4693) 1435-1441 DOI: 10.1126/science.2983426

[29] Pearson W R 2000 Flexible sequence similarity searching with the FASTA3 program package Methods Mol. Biol. 185-219 DOI: 10.1385/1-59259-192-2:185

[30] Sakoe H and Chiba S 1978 Dynamic programming optimization for spoken word recognition IEEE Trans. Acoust., Speech, Signal Process 26(1) 43-49 DOI: 10.1109/tassp.1978.1163055

[31] Myers C, Rabiner L R and Rosenberg A E 1980 Performance tradeoffs in dynamic time warping algorithms for isolated word recognition IEEE Transactions on Acoustics, Speech and Signal Processing 28(6) 623-635 DOI: 10.1109/tassp.1980.1163491

[32] Silva D F and Batista G E A P A 2016 Speeding up all-pairwise dynamic time warping matrix calculation Proceedings of the SIAM International Conference on Data Mining 837-845 DOI: 10.1137/1.9781611974348.94

[33] Virus Database at University College London (VIDA) URL: http://www.biochem.ucl.ac.uk/bsm/virus/database/VIDA3/VIDA.html

[34] Lanckriet G, Bie T De, Cristianini N, Jordan M I and Noble W S 2004 A statistical framework for genomic data fusion Bioinformatics 20(16) 2626-2635 DOI: 10.1093/bioinformatics/bth294

[35] Shimodaira H, Noma K-I, Nakai M and Sagayama S 2002 Dynamic time-alignment kernel in support vector machine Adv. Neural Inform. Process. Syst. 14 921-928

Acknowledgments
This work is supported by the Russian Fund for Basic Research, grants 15-07-08967, 18-07-01087, 18-07-00942.