# Применение новой метрики, основанной на парном выравнивании биомолекулярных последовательностей, для распознавания вирусов герпеса

**В.В. Сулимова[1], О.С. Середин[1], В.В. Моттль[2], А.И. Макарова[1]**

[1]Тульский государственный университет, Ленина 92, Тула, Россия, 300012
[2]Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН, Вавилова 44/2, Москва, Россия, 119333

**Аннотация.** В данной статье рассматривается проблема интеллектуального анализа белковых последовательностей с целью распознавания среди них белков, относящихся к вирусам простого герпеса. В силу сложности введения полезного для распознавания признакового описания белков, мы придерживаемся беспризнакового подхода, основанного на парном сравнении белковых последовательностей. Для сравнения белков мы применяем новую метрику на множестве белков, основанную на оптимальном выравнивании составляющих их аминокислотных последовательностей, которая впервые была предложена нами в предыдущей работе. В данной работе мы показываем, что наличие свойств метрики, которые отсутствуют, например, у такого традиционного способа сравнения биомолекулярных последовательностей, как алгоритм Нидлмана-Вунша, позволяют более удобно и эффективно использовать предложенную меру сравнения для дальнейшего анализа и улучшить качество распознавания вирусов герпеса.

## 1. Введение

Человеческие вирусы герпеса являются патогенами, которые могут находиться в клетках человека очень длительное время без каких-либо клинических проявлений до тех пор, пока не возникнут подходящие для этого условия. Эти вирусы обычно не опасны для жизни, но в некоторых случаях они могут вызвать серьезные инфекции глаз и мозга, которые могут привести к слепоте и, возможно, смерти. Существует ряд препаратов против вирусов герпеса (такие, как ацикловир и его производные). Однако для эффективного лечения очень важны раннее выявление и идентификация конкретного вида этих вирусных инфекций [1,2].

В данной работе мы предлагаем беспризнаковый [3, 4] способ распознавания вирусов герпеса (не требующий введения их признакового описания), основанный лишь на сравнении их белковых последовательностей. Очевидно, что в данном случае возможность достижения хорошего качества распознавания во многом определяется качеством сравнения последовательностей.

Традиционно сравнение белков заключается в вычислении мер сходства, основанных на нахождении оптимального парного выравнивания [5-8]. Но они не дают возможности использовать преимущества популярных и эффективных линейных методов, изначально разработанных для линейных признаковых пространств, таких как метод опорных векторов (SVM), предложенный В.Н. Вапником [9].

Частично проблема применения линейных методов в беспризнаковых ситуациях может быть решена путем построения специальной мер сходства, называемых потенциальными [10-12], которые погружают множество белков в некоторые гипотетические линейные пространства и играют роль скалярного произведения в них [12]. Но построение математически корректных и, в то же время, биологически обоснованных потенциальных функций, как правило, является теоретически и вычислительно сложной задачей [12-14] Кроме того, существуют целые классы потенциальных функций, являющихся эквивалентными с точки зрения правила окончательного решения [15].

В то же время, очевидно, что не векторы признаков объектов в некотором линейном признаковом пространстве являются фактической основой алгоритмов машинного обучения и интеллектуального анализа данных, а парное расстояние между объектами, т.е. метрика [16]. В связи с этим возникает естественное желание использовать способы сравнения, обладающие свойствами метрики, тем более, что метрика позволяет погружать любое множество белков в линейное пространство и применять в нем линейные методы в соответствии с обобщенным линейным подходом к восстановлению зависимостей [17].

Однако следует отметить, что эффективность методов восстановления зависимостей (и распознавания белков, в частности) существенно зависит от выбора метрики между объектами, которая для успешного распознавания должна удовлетворять гипотезе компактности [18,19]. В данном случае это означает, что значения принятой метрики между белками, относящимися к вирусам герпеса одного и того же вида должны быть, как правило, небольшими и, соответственно, они должны быть существенно больше, если сравниваются вирусы герпеса разных видов или вирус герпеса и не вирус.

В мировой литературе известен ряд способов введения метрик на множестве последовательностей [16], но в случае аминокислотных (белковых) последовательностей ни один из них не имеет удовлетворительной интерпретации с биологической точки зрения. В связи с этим маловероятно, что гипотеза компактности будет иметь место в соответствующем метрическом пространстве белков. Это имеет множество подтверждений на практике [20, 21], что породило целую серию работ, направленных на улучшение исходной метрики с помощью различных алгебраических конструкций разной степени сложности (Metric Learning) [19-23], включая построение метрик на основе потенциальных функций (Metric Kernel Learning) и построение вторичных признаков или новых метрик на основе исходных мер сходства или несходства. Кроме того, в ряде работ делается попытка включить в способ сравнения информацию о вторичных структурах белков и(или) белок-белковых взаимодействиях [24, 25]. Однако в данной статье нашей целью является распознавание вирусов герпеса исключительно на основе информации о первичных структурах белков (составляющих их последовательностях аминокислотных остатков) без привлечения какой-либо дополнительной информации.

Мы предлагаем достаточно простой способ построения математически корректной метрики на множестве белковых последовательностей. Данный подход, вслед за традиционным для биоинформатики алгоритмом Нидлмана-Вунша [5], основывается на поиске оптимального глобального парного выравнивания последовательностей и опирается на вероятностную модель PAM (Point Accepted Mutation) [26] эволюционных изменений аминокислот в цепи. Но, в то же время, предложенный подход отличается от традиционных критерием оптимальности и способом сравнения аминокислот. В предыдущей работе [27] мы доказали, что предложенная мера сравнения белков (в отличие от традиционных способов сравнения) обладает свойствами метрики [28]. Наличие этих свойств дает возможность ее более удобного и эффективного использования для дальнейшего распознавания.

В данной работе мы предлагаем усовершенствованную вычислительную схему, позволяющую ускорить процесс сравнения белковых последовательностей, что очень важно при работе с реальными данными.

Кроме того, данная работа содержит существенно более детальное экспериментальное исследование предложенной метрики. Результаты экспериментов показывают, что наличие свойств метрики позволяет получить более высокое качество распознавания вирусов герпеса по сравнению с другими способами сравнения, не обладающими такими свойствами.

## 2. Сравнение белковых последовательностей

*2.1. Метрика на множестве белковых последовательностей*

Пусть $\Omega$ - множество всех возможных белковых последовательностей, т.е. последовательностей над множеством 20 известных аминокислот $A = \{\alpha^1,\ldots,\alpha^m\}, m = 20$.

Пусть также $\omega' = (\alpha'_1,\alpha'_2,\ldots,\alpha'_{N'}) \in \Omega$ и $\omega'' = (\alpha''_1,\alpha''_2,\ldots,\alpha''_{N''}) \in \Omega$ - две последовательности длин, соответственно, $N'$ и $N''$, состоящие из аминокислот $\alpha'_i,\alpha''_j \in A$, $i = 1,\ldots,N', j = 1,\ldots,N''$.

Очевидно, что сравнение аминокислотных последовательностей должно основываться на сравнении составляющих их аминокислот. Основной теоретической концепцией, лежащей в основе сравнения аминокислот, в данной работе является широко известная вероятностная модель эволюции аминокислот PAM (Point Accepted Mutation), предложенная М. Дэйхофф [26]. Основным ее инструментом является понятие Марковской цепи эволюции аминокислот в отдельной позиции цепи, представленная матрицей переходных вероятностей $\Psi = \left(\psi_{[1]}(\alpha^j \mid \alpha^i)\right)$ замены аминокислоты $\alpha^i$ на аминокислоту $\alpha^j$ на следующем шаге эволюции. Индекс "1" в квадратных скобках означает, что рассматривается исходная одношаговая Марковская цепь.

В соответствии с моделью PAM предполагается, что эта Марковская цепь является эргодическим и обратимым случайным процессом, т. е. процессом, который имеет финальное распределение вероятностей

$$\xi\left(\alpha^j\right) = \sum_{\alpha^i \in A} \xi\left(\alpha^i\right)\psi_{[1]}\left(\alpha^j \mid \alpha^i\right)$$

и удовлетворяет условию обратимости

$$\xi(\alpha^i)\psi_{[1]}(\alpha^j \mid \alpha^i) = \xi(\alpha^j)\psi_{[1]}(\alpha^i \mid \alpha^j).$$

Рассмотрим вероятностный процесс эволюции с большим эволюционным шагом $s > 1$, т.е. разреженной цепью Маркова с матрицей переходных вероятностей

$$\Psi_{[s]} = [\underbrace{\Psi_{[1]} \times \ldots \times \Psi_{[1]}}_{s}].$$

Ранее мы доказали [12], что для любых значений *s* меры сходства

$$\kappa_s(\alpha^i,\alpha^j) = \psi_{[s]}(\alpha^i \mid \alpha^j) / \xi(\alpha^i)$$

образуют неотрицательно определенные матрицы парного сходства аминокислот для $\alpha^i,\alpha^j \in A$, $i, j = 1,\ldots,20$. Таким образом, каждая из них является потенциальной функцией, погружающей множество аминокислот $A$ в соответствующее гипотетическое линейное пространство $\tilde{A}_s \subset A$ с евклидовой метрикой [28]

$$\rho_s(\alpha^i,\alpha^j) = \left(\kappa_s(\alpha^i,\alpha^i) + \kappa_s(\alpha^j,\alpha^j) - 2\kappa_s(\alpha^i,\alpha^j)\right)^{1/2},\ s = 1,2,\ldots. \qquad (1)$$

Именно метрику (1) мы используем для сравнения аминокислот. В дальнейшем для простоты мы будем опускать нижний индекс: $\rho(\alpha^i,\alpha^j)$, $\kappa(\alpha^i,\alpha^j)$.

Метрику на множестве аминокислотных (белковых) последовательностей мы определяем на основе глобального парного выравнивания $\mathbf{w}(\omega',\omega'')$, под которым понимается способ преобразования последовательностей путем вставок специальных элементов, называемых

"пропуск" (gap) в некоторые позиции выравниваемых последовательностей для приведения их к одинаковой длине.

Если позиция $\mathbf{w}_i, i = 1,\ldots,|\mathbf{w}|$ выравнивания $\mathbf{w}(\omega', \omega'')$ двух последовательностей $\omega' = (\alpha'_1,\ldots,\alpha'_{N'}) \in \Omega$ и $\omega'' = (\alpha''_1,\ldots,\alpha''_{N''}) \in \Omega$ не содержит пропуска, то она определяет однозначное парное соответствие двух аминокислот $(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}})$.

Меру сравнения белковых последовательностей мы определяем как оптимальное значение специального критерия оптимальности:

$$\mathrm{r}(\omega', \omega'') = \min_{\mathbf{w}} \sqrt{\sum_{i=1}^{|\mathbf{w}|} \left[ I(\mathbf{w}_i)\beta^2 + \left(1 - I(\mathbf{w}_i)\right)\rho^2(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) \right]}, \qquad (2)$$

где $I(\mathbf{w}_i) = 1$, если *i*-я позиция выравнивания $\mathbf{w}$ содержит пропуск и $I(\mathbf{w}_i) = 0$, если не содержит, а коэффициент $\beta$ имеет смысл штрафа за пропуск.

В нашей предыдущей работе [27] мы доказали, что для любых $\beta$, удовлетворяющих условию $\beta \geq 0.5 \max_{\alpha', \alpha'' \in A} \rho(\alpha', \alpha''), \quad \forall \alpha \in A$, функция (2) обладает свойствами метрики.

*2.2. Процедура динамического программирования для вычисления метрики на множестве белковых последовательностей*

Критерий (2) относится к классу парно-сепарабельных целевых функций. Минимум такой функции может быть найден при помощи процедуры динамического программирования, которая аналогична процедуре Нидлмана-Вунша [5] для нахождения оптимального глобального выравнивания, максимизирующего сходство пары последовательностей.

Идея данного алгоритма заключается в рекуррентном вычислении неизвестных значений несходства $\mathrm{F}_{i,j}$ для увеличивающихся начальных фрагментов аминокислотных последовательностей $(\alpha'_1, \alpha'_2, \ldots, \alpha'_i)$ и $(\alpha''_1, \alpha''_2, \ldots, \alpha''_j)$ на основе уже вычисленных значений несходства:

$$F_{i,j} = \min \begin{cases} F_{i-1,j-1} + \rho^2(\alpha'_i, \alpha''_j); \\ F_{i-1,j} + \beta^2; \\ F_{i,j-1} + \beta^2, \end{cases} \qquad i = 1,\ldots,N', \quad j = 1,\ldots,N''.$$

Вычисления начинаются с инициализации:

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

и заканчиваются при достижении концов последовательностей:

$$r(\omega', \omega'') = \sqrt{F_{N'N''}}.$$

Такой вычислительный процесс удобно представить в виде таблицы парных соответствий (рисунок 1).

Вычислительный процесс состоит в последовательном прохождении через все ячейки таблицы (рисунок 1, слева), с левой верхней до правой нижней, и выполнении рекуррентного вычисления неполно значения несходства $\mathrm{F}_{i,j}$ в каждой ячейке, выбирая (и, возможно, сохраняя) направление оптимального перемещения в текущую ячейку (по горизонтали, вертикали или диагонали), которые могут быть использованы в дальнейшем для нахождения оптимального выравнивания (рисунок 1, справа).

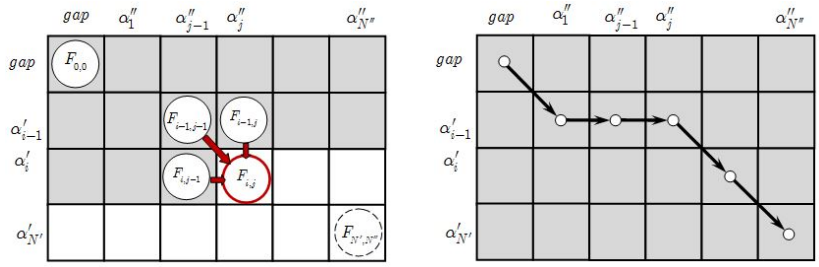


**Рисунок 1.** Таблица парных соответствий между элементами двух сравниваемых последовательностей: иллюстрация вычислительного процесса (слева) и возможное оптимальное выравнивание (справа).

## 3. Экспериментальное исследование полезности свойств метрики для распознавания вирусов герпеса

*3.1. Описание данных*

Для экспериментального исследования мы использовали шесть наборов аминокислотных последовательностей вирусов герпеса, которые выполняют функции, указанные в таблице 1. Это наборы из базы данных VIDA [35]. Белки каждого набора разделены на классы и гомологичные семейства на основе лабораторных исследований вирусов герпеса [2].

Кроме того, для некоторых экспериментов мы использовали дополнительный набор из 143 белковых последовательностей, которые не являются вирусами герпеса. Эти белки были случайным образом выбраны из набора данных, собранного Ланкриэтом и др. [36].

Полученный набор данных содержит 1532 белковых последовательности с длинами в диапазоне от 36 до 1378 аминокислотных остатков.

**Таблица 1.** Структура данных для экспериментального исследования.

| Набор данных (число белков) | Функция | Класс (число белков) | Описание | Гомологические семейства (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 Assambly | 1 (54) | Transport/ capsid assambly 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 polimerase processivity factor | 104, 1003 |
| | | 3 (48) | Helicase/ primase associated protein | 16 |
| | | 4 (22) | Component of DNA helicase/ promise 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 |

*3.2. Построение потенциальных функций на основе функций парного сравнения*

Основная цель эксперимента заключается в том, чтобы продемонстрировать, что наличие свойств метрики у функции парного сравнения позволяет повысить качество распознавания. Для достижения этой цели мы сравниваем два очень похожих способа сравнения белковых последовательностей, каждый из которых основан на поиске оптимального глобального парном выравнивании последовательностей - алгоритм Нидлмана-Вунша (NW) [5] и предлагаемую метрику. Эти два способа сравнения имеют похожее строение, но разные свойства: в то время как первая является мерой сходства, вторая - является мерой несходства и, более того, обладает свойствами метрики.

В наших экспериментах мы использовали стандартную реализацию алгоритма Нидлмана-Вунша из MATLAB bioinformatics toolbox с наиболее популярными матрицами замен аминокислот PAM250 и BLOSSUM62 для сравнения одиночных аминокислот и с принятыми по умолчанию значениями штрафа: g = 12 для начала и b = 1 для продолжения серии пропусков.

Метрика на множестве аминокислот была построена в соответствии с (1) на основе модели PAM с эволюционным шагом s = 250. Параметр $\beta$ предлагаемой метрики был принят равным 0.0234 для всех экспериментов.

В данной работе мы ориентируемся на использование такого удобного и хорошо зарекомендовавшего себя инструмента распознавания, как метод опорных векторов (SVM) [9]. Однако его использование требует введения специальных мер сходства, называемых потенциальными функциями, имеющими смысл скалярного произведения в некотором гипотетическом линейном пространстве.

Поскольку исходные способы сравнения не обладают свойствами потенциальных функций, то для применения метода опорных векторов требуется их дополнительное преобразование.

В ходе вычислительного эксперимента оказалось, что предложенная метрика является евклидовой метрикой для всех рассмотренных наборов белков. Этот факт дал нам возможность применить радиальную потенциальную функцию $K^{\alpha}_{mtr} = K^{\alpha}_{mtr}(\omega', \omega'') = \exp(-\alpha r^2(\omega', \omega''))$ [11, 25]. Однако следует отметить, что в общем случае предложенный способ построения метрики не гарантирует наличие свойств евклидовой метрики и, таким образом, такое преобразование может привести к наличию отрицательных собственных чисел у соответствующей матрицы значений потенциальной функции. Но практика показывает, что свойство евклидовости обычно выполняется.

Также мы построили линейные потенциальные функции в пространстве вторичных признаков $K^{SF}_{mtr} = R^T_{mtr} R_{mtr}$, где $R_{mtr} = \{r(\omega_i, \omega_j), i, j = 1, \ldots, 1532\}$ - матрица значений метрики.

Что касается меры сходства Нидлмана-Вунша (NW), то соответствующая радиальная потенциальная функция для нее не применима, поскольку она является мерой сходства и, более того, она часто имеет отрицательные значения. Таким образом, любое эвристическое преобразование такого сходства в метрику требует, как минимум, приведения значений к положительному диапазону. Такое преобразование требует наличия полного набора белков на стадии обучения, что обычно невозможно на практике. В связи с этим для меры сходства NW мы построили только линейные потенциальные функции в пространстве соответствующих вторичных признаков $K^{SF}_{NW} = S^T_{NW} S_{NW}$, где $S_{NW}$ - матрица значений сходства белков, вычисленная согласно алгоритму Нидлмана-Вунша.

*3.3. Распознавание классов и гомологических семейств вирусов герпеса*

В данном эксперименте мы использовали два набора аминокислотных последовательностей вирусов герпеса, которые выполняют, соответственно, следующие функции: «Membrane/Glycoprotein» и «Nucleotide/ repair metabolism». Очень важно различать эти классы и гомологические семейства (HPFs) друг от друга, тем более, что предыдущие исследования показали, что они являются наиболее проблематичными для распознавания [2].

Для каждого из указанных способов сравнения и для каждого из двух рассмотренных наборов белков было решено несколько задач двухклассового распознавания образов ("один-против-всех" и "один-против-одного", для классов и для гомологических семейств). В каждом случае обучение проводилось с использованием SVM [9]. Качество построенных решающих правил распознавания оценивалось с помощью процедуры скользящего контроля (Leave-One-Out).

Таблица 2 содержит проценты ошибок на скользящем контроле для случаев с результатами, которые отличаются, по меньшей мере, для двух алгоритмов. Кроме того, результаты, полученные для NW с PAM250 и BLOSSUM62, практически одинаковы в этих экспериментах. В этой связи в таблицу включена только одна строка для алгоритма NW, которая соответствует матрице замен аминокислот BLOSSUM62.

Как видно из таблицы 2, практически во всех случаях наименьший процент ошибок на скользящем контроле достигается при распознавании с применением предложенной метрики на множестве белковых последовательностей. Единственное исключение составляет задача отличения 531 и 12 гомологических семейств. Однако данный случай является не показательным, поскольку данные гомологические семейства содержат очень близкие белковые последовательности, в результате чего, фактически, оказывается невозможным отличить белки данных семейств основываясь только на информации о составляющих их последовательностях аминокислотных остатков.

*3.4. Распознавание вирусов герпеса, выполняющих заданные функции*

В данном эксперименте мы решаем 12 двухклассовых задач распознавания вирусов герпеса, среди которых задачи 1-6 представляют собой задачи отличения каждого из 6 наборов белков, указанных в таблице 1, выполняющих определенную функцию, от белков, не относящихся к вирусам герпеса, и задачи 7-12 являются задачами отличения каждого из этих 6 наборов белков от оставшихся, соответствующих вирусам герпеса, выполняющим другие функции.

В каждом из этих 12 случаев для обучения двухклассовому распознаванию использовался метод опорных векторов (SVM). Качество построенных решающих правил оценивалось по значению площади под ROC-кривой (AUC), усредненному по результатам десятикратной кросс-валидации со случайным разбиением множества белков на обучающее и контрольное в пропорции 20:80. Результаты представлены в таблице 3. Лучший результат по каждой задаче выделен жирным шрифтом.

**Таблица 2.** Проценты ошибок на скользящем контроле для наиболее интересных случаев распознавания гомологических семейств (HPFs) и классов вирусов герпеса.

| **Membrane/glycoprotein** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Способ сравнения | Один против всех | | | | | | | | Один против одного (HPFs) | | |
| | Гомологические семейства | | | | | Классы | | | | | |
| | 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 | 0.86 | 0.86 | 0.43 | 6.45 | 4.17 | 48.06 |
| $K_{mtr}^{SF}$ | 15.40 | 0.00 | 0.00 | 0.00 | 0.43 | 0.43 | 0.43 | 0.00 | 3.22 | 2.08 | 50.00 |
| $K_{mtr}^{0.01}$ | 14.59 | 0.00 | 0.00 | 0.00 | 0.43 | 0.43 | 0.43 | 0.00 | 3.22 | 2.08 | 50.00 |

| **Nucleotide repair/metabolism** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Способ сравнения | Один против всех | | | Один против одного | | | |
| | HPFs | классы | | HPFs | | классы | |
| | 15 | 1 | 2 | 51/2 | 51/27 | 1/2 | 2/3 |
| $K_{NW}^{SF}$ | 14.25 | 0.49 | 0.49 | 1.91 | 1.91 | 0.59 | 1.67 |
| $K_{mtr}^{SF}$ | 14.49 | 0.49 | 0.25 | 0.64 | 0.64 | 0.30 | 1.67 |
| $K_{mtr}^{0.01}$ | 14.09 | 0.25 | 0.25 | 0.64 | 0.64 | 0.30 | 0.83 |

**Таблица 3.** Усредненные значения AUC для 12 задач распознавания вирусов герпеса.

| Способ сравнения | Задачи | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| $K_{NW}^{SF}$ | 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 |

Полученные результаты показали, что использование предлагаемой метрики вместо традиционной меры сходства NW позволяет существенно увеличить качество распознавания вирусов герпеса. Конечно, следует отметить, что предложенный способ сравнения имеет структурный параметр $\alpha$, поэтому необходим инструмент для автоматического выбора его наилучшего значения. Но этот аспект выходит за рамки данного исследования.

## 4. Заключение

В данной статье предложена новая мера сравнения белков, основанная на оптимальном глобальном выравнивании составляющих их последовательностей аминокислотных остатков. Предложенная мера очень похожа на традиционную меру сходства Нидлмана-Вунша, но, в отличие от нее, обладает свойствами метрики, что дает возможность более удобного и эффективного ее использования для распознавания.

Проведенное экспериментальное исследование показало, что использование предложенной метрики вместо традиционной меры сходства Нидлмана-Вунша позволяет существенно повысить качество распознавания вирусов герпеса.

## 5. Литература

[1] Huleihel, M. Detection of Vero Cells Infected with Herpes Simplex Types 1 and 2 and Varicella Zoster Viruses Using Raman Spectroscopy and Advanced Statistical Methods / M. Huleihel, E. Shufan, L. Zeiri, A. Salman // PLoS One. – 2016. – Vol. 11(4). – P. e0153599. DOI: 10.1371/journal.pone.0153599.

[2] Mc Geoch, D.J. Topics in herpesvirus genomics and evolution / D.J. Mc Geoch, F.J. Rixon, A.J. Davison // Virus Research. – 2006. – Vol. 117. – P. 90-104. DOI: 10.1016/j.virusres.2006.01.002.

[3] Duin, R.P.W. Experiments with a featureless approach to pattern recognition / R.P.W. Duin, D. De Ridder, D.M.J. Tax // Pattern Recognition Letters. – 1997. –Vol. 18(11-13). – P. 1159-1166.

[4] Mottl, V.V. Featureless Pattern Recognition in an Imaginary Hilbert Space and Its Application to Protein Fold Classification / V.V. Mottl, S.D. Dvoenko, O.S. Seredin, C.A. Kulikowski, I.B. Muchnik // Proc. of the II-th Int. Workshop on MLDM in Pat. Rec., 2001. – P. 322-336.

[5] Needleman, S.B. A general method applicable to the search for simlarities in the amino acid sequence of two proteins / S.B. Needleman, C.D. Wunsch // Journal of Molecular Biology. – 1970. – Vol. 48(3). – P. 443-453. DOI: 10.1016/0022-2836(70)90057-4.

[6] Smith, T.F. Identification of Common Molecular Subsecuences / T.F. Smith, M.S. Waterman // Journal of Molecular Biology. – 1981. – Vol. 147(1). – P. 195-197. DOI: 10.1016/0022-2836(81)90087-5.

[7] Zhang, Z. A greedy algorithm for align- ing DNA sequences / Z. Zhang, S. Schwartz, L. Wagnerm, W. Miller // Journal of Computational Biology. – 2000. – Vol. 7(1-2). – P. 203-214. DOI: 10.1089/10665270050081478.

[8] Durbin, R. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids / R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison // Cambridge Univ. Press, 1998. – 356 p.
[9] Vapnik, V.N. Statistical Learning Theory. – Wiley-Interscience, 1998. – 768 p.
[10] Schoolkopf, B. Kernel Methods in Computational Biology / B. Schoolkopf, K. Tsuda, J-P. Vert // MIT Press, 2004. – 410 p.
[11] Айзерман, М.А. Метод потенциальных функций в теории обучения машин. – М:Наука, 1970. – 384 с.
[12] Сулимова, В.В. Потенциальные функции для анализа сигналов и символьных последовательностей разной длины // Кандидатская диссертация. – Москва, 2009. – 122 с.
[13] Miklos, I. Stochastic models of sequence evolution including insertion-deletion events / I. Miklos, A. Novak, R. Satija, R. Lyngso, J. Hein // Statistical Methods in Medical Research. – 2009. – Vol. 18(5). – P. 453-485. DOI: 10.1177/0962280208099500.
[14] Seeger, M. Covariance kernels from bayesian generative models // Adv. Neural Inf. Proc. Syst. – 2002. – Vol. 14. – P. 905-912.
[15] Абрамов, В.И. Обучение распознаванию по методу опорных векторов в Евклидовых метрических пространствах с аффинными операциями // Известия ТулГУ. Естественные науки. – 2013. – Т. 2, № 1. – С. 119-136.
[16] Pekalska, E.M. Dissimilarity representations in pattern recognition // Concepts, theory and applications. PhD Thesis, 2005. – 344 p.
[17] Середин, О.С. Метод опорных объектов для обучения распознаванию в произвольных метрических пространствах / О.С. Середин, В.В. Моттль // Известия ТулГУ. Естественные науки. – 2015. – Т. 4. – С. 178-196.
[18] Браверман, Э.М. Опыты по обучению машины распознаванию образов. Кандидатская диссертация. – Москва, 1961.
[19] Xing, E.P. Distance Metric Learning with Application to Clustering with Side Information / E.P. Xing, A.Y. Ng, M.I. Jordan, S. Russel // Advances in Neural Information Processing Systems. – 2003. – Vol. 15. – P. 521-528.
[20] Bellet, A. A survey on metric learning for feature vectors and structured data / A. Bellet, A. Harbrad, M. Sebban // CoRR, 2013.
[21] Wang, J. Two-Stage Metric Learning / J. Wang, K. Sun, F. Sha, S. Marchand-Maillet, K. Kalousis // Proceedings of the 31st International Conference on Machine Learning, Cycle 2. – 2014. - Vol. 32. – P. 370-378.
[22] Schultz, M. Learning a distance metric from relative comparisons / M. Schultz, T. Joachims // Adv. Neural Inform. Process. Syst. – 2004. – Vol. 16. – P. 41-48.
[23] Wang, J. Metric learning with multiple kernels / J. Wang, H. Do, A. Woznica, A. Kalousis // Adv. Neural Inform. Process. Syst. – 2011. – Vol. 24. – P. 1-9.
[24] Cao, M. Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks / M. Cao, H. Zhang, J. Park, N.M. Daniels, M.E. Crovella // PLoS ONE. – 2013. – Vol. 8(10). – P. e76339. DOI: 10.1371/journal.pone.0076339.
[25] Rogen, P. Automatic classi_cation of protein structure by using Gauss integrals / P. Rogen, B. Fain // Proc. Natl. Acad. Sci. USA. – 2002. – Vol. 100(1). – P. 119-124. DOI: 10.1073/pnas.2636460100.
[26] Dayhoff, M. A model of evolutionary change in proteins / M. Dayhoff, R. Schwarts, B. Orcutt // Atlas of Protein Sequences and Structures. – 1978. – Vol. 5(3). – P. 345-352.
[27] Сулимова, В.В. Метрики на основе оптимального выравнивания биомолекулярных последовательностей / В.В. Сулимова, О.С. Середин, В.В. Моттль // JMLDA. – 2016. –Vol. 2(3). – С. 286-304. DOI: 10.21469/22233792.2.3.03.
[28] Моттль, В.В. Метрические пространства, допускающие введение линейных операций и скалярного произведения // Доклады РАН. – 2003. – Т. 388, № 3. – С. 312-315.
[29] Altschul, S.F. Basic local alignment search tool / S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman // J. Mol. Biol. – 1990. – Vol. 215(3). – P. 403-410. DOI: 10.1006/jmbi.1990.9999.

[30] Lipman, D.J. Rapid and sensitive protein similarity searches / D.J. Lipman, W.R. Pearson // Science. – 1985. – Vol. 227(4693). – P. 1435-1441. DOI: 10.1126/science.2983426.
[31] Pearson, W.R. Flexible sequence similarity searching with the FASTA3 program package // Methods Mol. Biol., 2000. – P. 185-219. DOI: 10.1385/1-59259-192-2:185.
[32] Sakoe, H. Dynamic programming optimization for spoken word recognition / H. Sakoe, S. Chiba // IEEE Trans. Acoust., Speech, Signal Proces. – 1978. – Vol. 26(1). – P. 43-49. DOI: 10.1109/tassp.1978.1163055.
[33] Performance tradeos in dynamic time warping algorithms for isolated word recognition / C. Myers, L.R. Rabiner, A.E. Rosenberg // IEEE Transactions on Acoustics, Speech and Signal Processing. – 1980. – Vol. 28(6). – P. 623-635. DOI: 10.1109/tassp.1980.1163491.
[34] Silva, D.F. Speeding up all-pairwise dynamic time warping matrix calculation / D.F. Silva, G.E.A.P.A. Batista // Proceedings of the SIAM International Conference on Data Mining, 2016. – P. 837-845. DOI: 10.1137/1.9781611974348.94
[35] Virus Database at University College London (VIDA) [Electronic resource]. – Access mode: http://www.biochem. ucl.ac.uk/bsm/virus database/VIDA3/VIDA.htm.
[36] Lanckriet, G. A statistical framework for genomic data fusion / G. Lanckriet, T. De Bie, N. Cristianini, M.I. Jordan, W.S. Noble // Bioinformatics. – 2004. – Vol. 20(16). – P. 2626-2635. DOI: 10.1093/bioinformatics/bth294.
[37] Shimodaira, H. Dynamic time-alignment kernel in support vector machine / H. Shimodaira, K.-I. Noma, M. Nakai, S. Sagayama // Adv. Neural Inform. Process. Syst. – 2002. – Vol. 14. – P. 921-928.

**Благодарности**
Работа выполнена при поддержке РФФИ, гранты: 15-07-08967, 18-07-01087, 18-07-00942.

# Alignment-Based Metric for Biomolecular Sequences for Herpes Viruses Recognition

**V.V. Sulimova[1], O.S. Seredin[1], V.V. Mottl[2], A.I. Makarova[1]**

[1]Tula State University, Lenine ave. 92, Tula, Russia, 300012
[2]Computing Center of the RAS, Vavilov st. 44/2, Moscow, Russia,119333

**Abstract.** This paper deals with the problem of intellectual protein sequences analysis with the purpose of herpes virus recognition. Since it is difficult to introduce a useful feature-based description of proteins that would be useful for recognition, we adhere to a featureless approach based on pairwise comparison of protein sequences. To compare proteins, we use a new dissimilarity measure based on finding an optimal sequence alignment, that was firstly proposed by us in the previous work. In this paper we show that the presence of metric properties that are absent, for example, in such a traditional method of comparing biomolecular sequences, like the Needleman-Wunsch alignment, makes it possible to more conveniently and effectively use the proposed comparison measure for further analysis and to improve the quality of herpes virus recognition.