Метрики на основе оптимального выравнивания биомолекулярных последовательностей*

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

vsulimova@yandex.ru; oseredin@yandex.ru; vmottl@yandex.ru

[1]ФГБОУ ВО Тульский государственный университет, Россия, г. Тула, пр. Ленина, д. 92

[2]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, д. 44/2

Для биомолекулярных последовательностей наиболее адекватным является так называемый беспризнаковый подход, основанный на сравнении последовательностей (измерении их сходства или несходства), минуя явное вычисление векторов их признаков. С точки зрения передовых методов анализа данных наиболее предпочтительным является использование в качестве способа сравнения меры несходства, обладающей свойствами метрики. С другой стороны, с точки зрения молекулярной биологии важно, чтобы способ сравнения учитывал биологические особенности объектов сравнения. Кроме того, в условиях обработки больших объемов данных важно, чтобы способ сравнения был эффективен с вычислительной точки зрения и позволял в дальнейшем применять удобные и эффективные методы анализа данных, такие как метод опорных векторов (SVM — support vector machine). Известно множество способов сравнения биомолекулярных последовательностей, однако ни один из них не обладает всеми требуемыми свойствами. В данной работе предлагается достаточно простой способ построения метрик на множестве биомолекулярных последовательностей. Предлагаемый метод, как и традиционные общепринятые способы сравнения биомолекулярных последовательностей (такие, как алгоритм Нидлмана–Вунша и Смита–Ватермана), основывается на поиске их оптимального парного выравнивания и механизме мутационных замен аминокислот в ходе эволюции, но отличается от них используемым критерием оптимальности, типом оптимизации и способом сравнения элементов последовательностей. Приводится доказательство того, что предложенные меры несходства обладают свойствами метрики. Это позволяет использовать их в передовых методах анализа данных, сохраняющих вычислительные достоинства SVM, но не требующих введения признаков последовательностей и(или) скалярного произведения. Результаты экспериментов подтверждают адекватность предложенных метрик прикладным задачам на примере классификации мембранных гликопротеинов.

Ключевые слова: *метрики; сравнение последовательностей; оптимальное парное выравнивание; биомолекулярные последовательности; беспризнаковый подход*

DOI: 10.21469/22233792.2.3.03

1 Введение

Биомолекулярные последовательности, к которым относят нуклеотидные и аминокислотные последовательности, образующие полимерные молекулы белка, являются типовыми объектами анализа данных. Основной целью их анализа является определение заключающейся в них генетической информации и функций, которые они выполняют в организме. Результаты анализа биомолекулярных последовательностей крайне важны и находят применение в медицине, фармакологии, косметологии, биотехнологии, сельском

*Работа выполнена при финансовой поддержке РФФИ, проект №15-07-08967.

хозяйстве, экологии и других областях. В частности, они используются при изучении молекулярных механизмов болезней, выявлении предрасположенности человека к заболеваниям, разработке новых лекарственных средств и т. д.

Для анализа биомолекулярных последовательностей необходимо уметь сравнивать их между собой.

Традиционными способами сравнения биомолекулярных последовательностей являются меры сходства, основанные на оптимальном парном выравнивании [1–4]. Однако такие способы сравнения не позволяют при дальнейшем анализе использовать преимущества удобных и эффективных линейных методов анализа данных, разработанных для признаковых пространств, например хорошо зарекомендовавшего себя SVM [5].

В ряде случаев для обеспечения возможности применения SVM осуществляется искусственное введение так называемых вторичных (проекционных) признаков [6–11]. Однако при этом происходит искажение исходного биологически обоснованного понимания сходства последовательностей, что с точки зрения молекулярной биологии является нежелательным. Кроме того, такой подход требует знания и запоминания значений парного сходства для всех исследуемых последовательностей, что вносит существенные неудобства с вычислительной точки зрения, нейтрализуя вычислительные достоинства SVM.

Отчасти эту проблему решает использование специальной меры сходства, называемой потенциальной функцией (kernel function) [9, 12, 13]. Потенциальная функция, определенная на множестве объектов произвольной природы, погружает это множество в гипотетическое линейное пространство, в котором играет роль скалярного произведения [14]. Построению таких функций посвящено множество публикаций (см., например, [9, 13, 15–20]). Однако вычисление потенциальных функций, являющихся не только математически корректными, но и имеющими смысл с точки зрения молекулярной биологии, является непростой и, как правило, вычислительно очень трудоемкой задачей [13, 18, 21]).

В то же время, несложно убедиться, что понятие линейного пространства является избыточным. Все известные методы анализа данных опираются именно на метрику (т. е. взаимное расположение объектов в пространстве) и именно метрика, а не координаты объектов в линейном пространстве определяют в конечном счете результат решения задачи [22, 23], что хорошо видно на рис. 1. Более того, существуют целые классы потенциальных функций, порождающих одну и ту же метрику и, соответственно, являющихся эквивалентными с точки зрения получаемых решений [22, 23].

В связи с этим гораздо более естественным представляется в качестве способа сравнения использовать меру несходства, обладающую свойствами метрики, тем более что последние исследования в области беспризнакового анализа данных показывают, что в терминах метрических пространств могут быть сформулированы многие методы анализа данных, включая SVM с сохранением его основных достоинств и свойств [24–28].

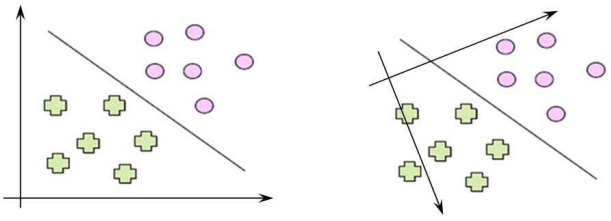

Рис. 1 Два класса объектов и решающее правило их распознавания в линейных пространствах, отличающихся выбором центра и базиса

При этом очевидно, что не любые метрики, построенные на множестве последовательностей, позволят получить приемлемое качество решения задачи анализа последовательностей, а только те, которые удовлетворяют так называемой гипотезе компактности — последовательности, относящиеся к одному классу (например, выполняющие одинаковую функцию) отображаются в более близкие точки данного пространства по сравнению с последовательностями, выполняющими разные функции.

В литературе известен ряд способов введения метрики на множестве последовательностей [29]. Однако они не имеют интерпретации с точки зрения молекулярной биологии, в связи с чем выполнение гипотезы компактности в порождаемом ими пространстве представляется маловероятным. Это находит многократные подтверждения на практике [30,31] и порождает целую серию публикаций, направленных на поиск способов улучшения исходной метрики (либо ее формирования на основе меры сходства) при помощи алгебраических конструкций различной степени сложности (Metric Learning) [30–34], включая построение метрик на основе потенциальных функций (Metric Kernel Learning) и проекционных признаков, что имеет описанные выше недостатки.

В данной работе предлагается достаточно простой способ сравнения биомолекулярных последовательностей, основанный, как и традиционные методы, на поиске оптимального парного выравнивания сравниваемых последовательностей, однако используются другой критерий оптимальности и другой способ сравнения элементов. В работе приводится доказательство, что предлагаемая функция парного сравнения обладает свойствами метрики. Экспериментальное исследование показывает, что данная метрика может успешно применяться для анализа биомолекулярных последовательностей.

2 Метрики на множестве элементов биомолекулярных последовательностей

Очевидно, что сравнение последовательностей должно базироваться на сравнении составляющих их элементов.

Типичным примером биомолекулярных последовательностей являются аминокислотные последовательности белков, т. е. последовательности над алфавитом двадцати известных аминокислот $A = \{\alpha^1, \ldots, \alpha^m\}, m = 20$.

В качестве теоретической концепции сравнения аминокислот в данной работе принята вероятностная модель эволюции Маргарет Дэйхофф, называемая PAM (Pointed Accepted Mutation) [35]. Ее основным математическим понятием является понятие марковской цепи эволюции аминокислот в отдельно взятой точке цепи, определяемой матрицей переходных вероятностей $\mathbf{\Psi} = \left(\psi_{[1]}(\alpha^j|\alpha^i)\right)$ замены аминокислоты α^i на аминокислоту α^j на следующем шаге эволюции. При этом предполагается, что эта марковская цепь представляет собой эргодический и обратимый случайный процесс, т. е. процесс, характеризующийся финальным распределением вероятностей $\xi(\alpha^j)$:

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

и удовлетворяющий условию обратимости

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

В работе [13] доказано, что для любой матрицы переходных вероятностей $\mathbf{\Psi}_{[s]} = [\underbrace{\mathbf{\Psi}_{[1]} \times \cdots \times \mathbf{\Psi}_{[1]}}_{s}]$, соответствующей разреженной марковской цепи (т. е. большему шагу эволюции), построенные на их основе меры сходства

$$\varkappa_s\left(\alpha^i,\alpha^j\right)=\frac{\psi_{[s]}\left(\alpha^i|\alpha^j\right)}{\xi\left(\alpha^i\right)}$$

обладают свойствами потенциальной функции на множестве аминокислот, образуя неотрицательно определенную матрицу значений парного сходства для любого $s>0$ и погружая множество аминокислот в гипотетическое линейное пространство $\tilde{A}\subset A$ с евклидовой метрикой [14]

$$\rho(\alpha^i,\alpha^j)=\left(\varkappa(\alpha^i,\alpha^i)+\varkappa(\alpha^j,\alpha^j)-2\varkappa(\alpha^i,\alpha^j)\right)^{1/2}, \tag{1}$$

где $\varkappa(\alpha^i,\alpha^j)$ — любая из функций $\varkappa_s(\alpha^i,\alpha^j),\quad s=1,2,\ldots$

Именно эту метрику предлагается использовать в данной работе для сравнения аминокислот.

Аналогичным образом может быть введена метрика на множестве нуклеотидов, составляющих нуклеотидные последовательности.

3 Метрики на основе оптимального выравнивания символьных последовательностей

3.1 Выравнивание символьных последовательностей и расширенные последовательности

Пусть Ω — множество всех последовательностей над некоторым конечным алфавитом $A=\{\alpha^1,\ldots,\alpha^m\}$, в частности алфавитом двадцати аминокислот или четырех нуклеотидов. И пусть $\omega'=(\alpha'_1,\alpha'_2,\ldots,\alpha'_{N'})\in\Omega$ и $\omega'=(\alpha''_1,\alpha''_2,\ldots,\alpha''_{N''})\in\Omega$ — две конкретные последовательности длины N' и N'' соответственно, состоящие из элементов $\alpha'_i,\alpha''_j\in A$, $i=1,\ldots,N'$, $j=1,\ldots,N''$.

Пусть также определена метрика на множестве элементов последовательностей, например в соответствии с (1), обладающая согласно определению следующими свойствами:

$$\left.\begin{array}{l}\rho(\alpha',\alpha''):A\times A\to R;\\ \rho(\alpha',\alpha'')\geqslant 0\ \forall\ \alpha',\alpha''\in A;\\ \rho(\alpha,\alpha)=0\ \forall\ \alpha\in A;\\ \rho(\alpha',\alpha'')+\rho(\alpha'',\alpha''')\geqslant\ \rho(\alpha',\alpha'''),\ \forall\ \alpha',\ \alpha'',\alpha'''\in A.\end{array}\right\} \tag{2}$$

Под выравниванием двух последовательностей $\omega'=(\alpha'_1,\alpha'_2,\ldots,\alpha'_{N'})\in\Omega$ длины N' и $\omega'=(\alpha''_1,\alpha''_2,\ldots,\alpha''_{N''})\in\Omega$ длины N'' понимается приведение их к одинаковой длине путем вставок так называемых «пропусков» в некоторые позиции последовательностей. Пример парного выравнивания двух последовательностей приведен на рис. 2.

α'_1	α'_2	–	α'_3	α'_4	α'_5	α'_6	α'_7
	|		|	|		|
–	α''_1	α''_2	α''_3	α''_4	–	–	α''_5

Рис. 2 Пример парного выравнивания двух последовательностей

Выравнивание естественно представить в виде таблицы парных соответствий элементов сравниваемых последовательностей, в которой пропускам соответствуют нули:

$$\mathbf{w}:\quad\left\{\begin{array}{cccccccc}\mathbf{w}_1&\mathbf{w}_2&\mathbf{w}_3&\mathbf{w}_4&\mathbf{w}_5&\mathbf{w}_6&\mathbf{w}_7&\mathbf{w}_8\\ 1&2&0&3&4&5&6&7\\ 0&1&2&3&4&0&0&5\end{array}\right..$$

Число столбцов данной таблицы определяет длину выравнивания, которая для рассматриваемого примера составляет $N_{\mathbf{w}} = 8$.

Будем называть выравнивание $\mathbf{w}$ допустимым, если оно не содержит два пропуска в одной позиции: $I_{\mathbf{w}} = \{i : \mathbf{w}_{i,1} = \mathbf{w}_{i,2} = 0\} = \varnothing$.

Множество всех допустимых выравниваний пары последовательностей длин N' и N'' будем обозначать $W_{N'N''}$.

Далее будем рассматривать только допустимые выравнивания.

Последовательность $\tilde{\omega}'$, полученную из исходной последовательности ω' путем вставки в нее пропусков, будем называть расширенной последовательностью. Символом $\tilde{\Omega}$ обозначим множество всех возможных расширенных последовательностей над расширенным алфавитом $\tilde{A} = A \cup \{-\} = \{\alpha^1, \ldots, \alpha^m, -\} = \{\tilde{\alpha}^1, \ldots, \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$.

В результате конкретного выравнивания $\mathbf{w} = \mathbf{w}(\omega', \omega'')$ пары последовательностей $\omega' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_{N'}) \in \Omega$ и $\omega' = (\alpha''_1, \alpha''_2, \ldots, \alpha''_{N''}) \in \Omega$ образуются расширенные последовательности $\tilde{\omega}' = (\tilde{\alpha}'_1, \tilde{\alpha}'_2, \ldots, \tilde{\alpha}'_{N_{\mathbf{w}}}) \in \tilde{\Omega}$ и $\omega'' = (\tilde{\alpha}''_1, \tilde{\alpha}''_2, \ldots, \tilde{\alpha}''_{N_{\mathbf{w}}}) \in \tilde{\Omega}$ одинаковой длины $N_{\mathbf{w}}$, причем

$$\tilde{\alpha}'_{\mathbf{w},i} = \begin{cases} \alpha'_{\mathbf{w}_{i,1}}, & \mathbf{w}_{i,1} \neq 0 \\ -, & \mathbf{w}_{i,1} = 0 \end{cases} ; \quad \tilde{\alpha}''_{\mathbf{w},i} = \begin{cases} \alpha''_{\mathbf{w}_{i,2}}, & \mathbf{w}_{i,2} \neq 0 \\ -, & \mathbf{w}_{i,2} = 0 \end{cases} , \; i = 1, \ldots, N_{\mathbf{w}}. \tag{3}$$

3.2 Метрика на расширенном множестве элементов последовательностей

Продолжим функцию $\rho(\alpha', \alpha'')$ на расширенное множество $\tilde{A} = A \cup \{-\} = \{\alpha^1, \ldots, \alpha^m, -\} = \{\tilde{\alpha}^1, \ldots, \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$ элементов последовательностей следующим образом:

$$\tilde{\rho}(\alpha', \alpha'') = \rho(\alpha', \alpha'') \; \forall \; \alpha', \alpha'' \in A; \quad \tilde{\rho}(-, -) = 0. \tag{4}$$

В дополнение к (4) необходимо также определить значения несходства элементов исходного множества с пропуском $\tilde{\rho}(\alpha, -), \; \alpha \in A$, которые в общем случае могут быть различны.

Теорема 1. *Для того чтобы функция $\tilde{\rho}(\alpha', \alpha'')$, определенная согласно (4), являлась метрикой на расширенном множестве элементов, достаточно, чтобы выполнялось условие:*

$$\tilde{\rho}(\alpha, -) \geqslant const = \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \; \forall \; \alpha \in A.$$

Доказательство теоремы приведено в приложении 1.

3.3 Меры несходства последовательностей, условные относительно выравнивания

Для фиксированного допустимого выравнивания $\mathbf{w}(\omega', \omega'') \in W_{N'N''}$ меры несходства пары последовательностей ω' и ω', условные относительно данного выравнивания, определим двумя способами:

$$r_1(\omega', \omega''|\mathbf{w}) = \sum_{i: \mathbf{w}_{i,1} \neq 0, \mathbf{w}_{i,2} \neq 0} \rho(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i: \; \mathbf{w}_{i,1}=0 \text{ или } \mathbf{w}_{i,2}=0} \beta;$$

$$r_2(\omega', \omega''|\mathbf{w}) = \sqrt{\sum_{i: \mathbf{w}_{i,1} \neq 0, \mathbf{w}_{i,2} \neq 0} \rho^2(\alpha'_{\mathbf{w}_{i,1}}, \alpha''_{\mathbf{w}_{i,2}}) + \sum_{i: \; \mathbf{w}_{i,1}=0 \text{ или } \mathbf{w}_{i,2}=0} \beta^2},$$

где

$$\beta = \tilde{\rho}(\alpha, -) \geqslant \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \; \forall \; \alpha \in A. \tag{5}$$

В терминах расширенных последовательностей, с учетом соответствий элементов исходных и расширенных последовательностей (3), определяемых выравниванием $\mathbf{w}$, приведенные меры несходства могут быть записаны в более компактной форме:

$$r_1(\omega', \omega''|\mathbf{w}) = \sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ ); \tag{6}$$

$$r_2(\omega', \omega''|\mathbf{w}) = \sqrt{\sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ )}. \tag{7}$$

3.4 Метрики на множестве последовательностей на основе оптимального выравнивания

Определим две меры несходства последовательностей на основе (6) и (7) соответственно:

$$\mathrm{r}_1(\omega', \omega'') = \min_{\mathbf{w}\in W_{N'N''}} r_1(\omega', \omega''|\mathbf{w}) = \min_{\mathbf{w}\in W_{N'N''}} \sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ ); \tag{8}$$

$$\mathrm{r}_2(\omega', \omega'') = \min_{\mathbf{w}\in W_{N'N''}} r_2(\omega', \omega''|\mathbf{w}) = \min_{\mathbf{w}\in W_{N'N''}} \sqrt{\sum_{i=1}^{N_\mathbf{w}} \tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w},i}, \tilde{\alpha}''_{\mathbf{w},i}\ )}. \tag{9}$$

Теорема 2. *Для любой метрики $\tilde{\rho}(\tilde{\alpha}', \tilde{\alpha}'')$, $\tilde{\alpha}', \tilde{\alpha}'' \in \tilde{A}$, на расширенном множестве элементов $\tilde{A} = A \cup \{-\} = \{\alpha^1, ..., \alpha^m, -\} = \{\tilde{\alpha}^1, ..., \tilde{\alpha}^m, \tilde{\alpha}^{m+1}\}$ меры несходства последовательностей* (8) *и* (9) *обладают свойствами метрики.*

Доказательство теоремы приведено в приложении 2.

4 Алгоритм вычисления метрики на множестве последовательностей

Критерии (8) и (9) относятся к классу парно-сепарабельных целевых функций, поскольку состоят из слагаемых, каждое из которых зависит только от двух соседних переменных. Минимум таких целевых функций может быть найден при помощи процедуры динамического программирования, аналогичной процедуре Нидлмана–Вунша для поиска оптимального глобального выравнивания последовательностей, максимизирующей их сходство [4].

Идея алгоритма заключается в рекуррентном вычислении неизвестных значений несходства $F_{i,j}$ начальных фрагментов последовательностей $(\alpha'_1, \alpha'_2, \ldots, \alpha'_i)$ и $(\alpha''_1, \alpha''_2, \ldots, \alpha''_j)$ на основе уже найденных значений. Для критерия (8) вычисление осуществляется по формуле:

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

а для критерия (9) — по формуле

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

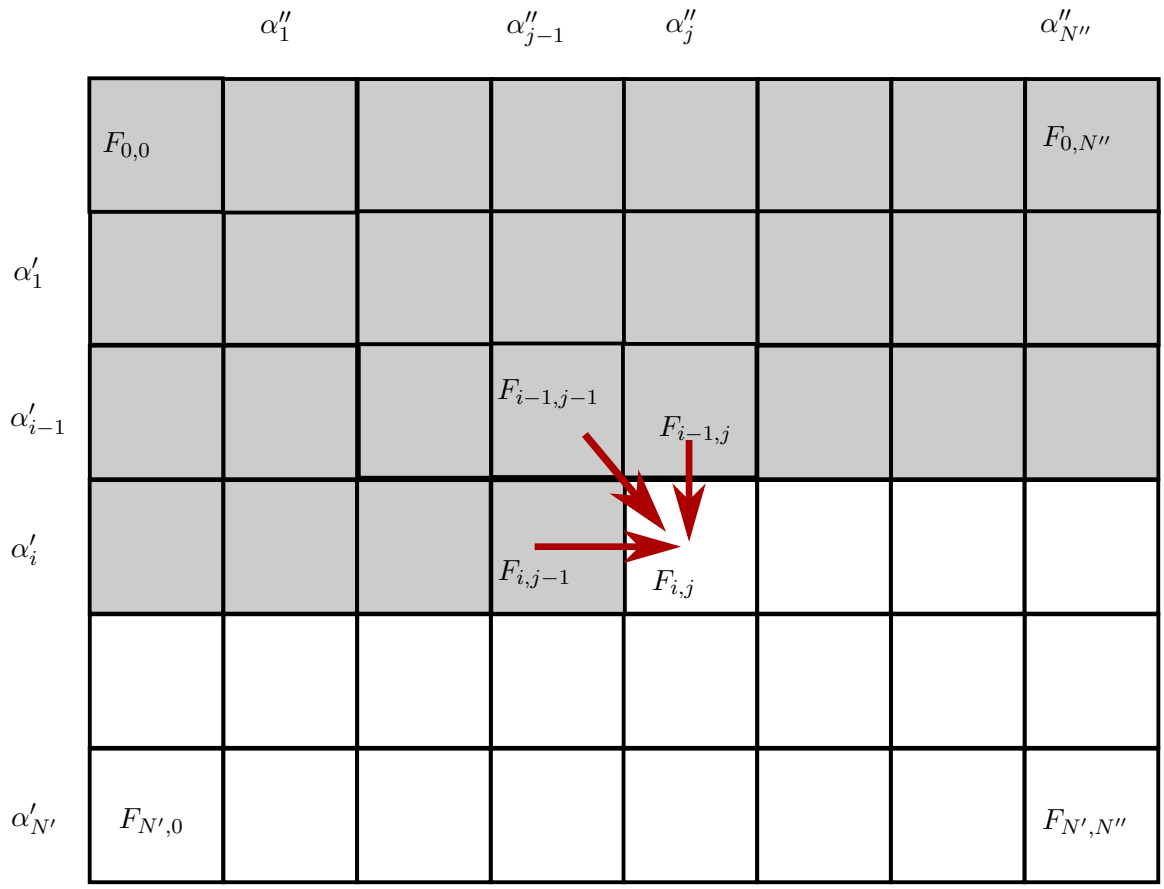

Рис. 3 Схема вычислительного процесса

Для обоих критериев вычисления начинаются с инициализации:

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

а заканчиваются при достижении концов последовательностей: $r_1(\omega', \omega'') = F_{N'N''}$ и $r_2(\omega', \omega'') = \sqrt{F_{N'N''}}$.

Вычислительный процесс такого вида удобно представлять при помощи таблицы парных соответствий (рис. 3).

Вычислительный процесс заключается в последовательном прохождении всех ячеек таблицы, начиная с левой верхней и заканчивая правой нижней, осуществляя рекуррентные вычисления неполных значений несходства $F_{i,j}$, выбирая и запоминая оптимальное перемещение в соответствующую ячейку (по горизонтали, по вертикали или по диагонали). При этом перемещение по горизонтали соответствует вставке пропуска в последовательность ω', перемещение по вертикали — вставке пропуска в последовательность ω'' и продвижение по диагонали — сравнению элементов последовательностей, стоящих на пересечении соответствующих строки и столбца.

Запомненные для каждой ячейки направления оптимальных перемещений могут быть использованы на обратном ходе процедуры, начинаемом с последней ячейки с координатами (N', N''), для восстановления оптимального пути, однозначно определяющего выравнивание пары последовательностей.

5 Экспериментальное исследование

5.1 Исходные данные

В качестве базы для экспериментального исследования использовались аминокислотные последовательности вирусов простого герпеса из базы данных VIDA (Virus Database at University College London) [36], разделенные специалистами в области молекулярной биологии на три класса на основе лабораторного анализа эволюции вирусов герпеса [37]. Структура исходных данных представлена в табл. 1.

Белки всех трех классов выполняют одну и ту же функцию «Мембранный гликопротеин» (Membrane Glicoprotein), но отличаются друг от друга типом гликопротеинов (например, белки класса 1 являются гликопротеинами-H, а белки класса 2 — гликопротеинами-L). Каждый из рассматриваемых классов включает в себя несколько семейств гомологичных белков (Homologous Protein Families — HPF). Согласно исследованию, проведенному в работе [37], семейства, объединенные в один класс, имеют общего прародителя.

Таблица 1 Исходные данные для анализа последовательностей

Класс	Описание	Гомологические семейства (HPF)	Число белков
1 (109 белков)	Гликопротеин H (glycoprotein H)	12	52
		42	39
		531	18
2 (77 белков)	Гликопротеин L (glycoprotein L)	47	30
		50	32
		114	13
		296	2
3 (48 белков)	Гликопротеин M (glycoprotein M)	20	48

5.2 Сравнение оптимальных выравниваний

Предложенный подход, как и традиционный для молекулярной биологии алгоритм Нидлмана–Вунша, основывается на поиске оптимального глобального парного выравнивания сравниваемых последовательностей.

Найденные оптимальные выравнивания зависят от способа сравнения элементов последовательностей и от используемого значения штрафа на пропуск элементов последовательностей.

Экспериментальное исследование показывает, что при стандартных (использующихся по умолчанию) настройках алгоритма Нидлмана–Вунша и описанном способе построения метрики на множестве биомолекулярных последовательностей со значением штрафа (5) найденные оптимальные выравнивания, как правило, оказываются полностью идентичными либо очень близкими для эволюционно близких последовательностей. С увеличением эволюционного расстояния (увеличением доли точечных мутаций аминокислот) локальные различия между найденными оптимальными выравниваниями могут увеличиваться, однако такие традиционные характеристики качества вырванивания, как общее количество пар сопоставленных друг другу в результате одинаковых (identities) и близких (positives) аминокислот оказываются достаточно близки, что подтверждает осмысленность выравниваний, найденных при помощи предложенного подхода. Примеры оптимальных выравниваний представлены на рис. 4.

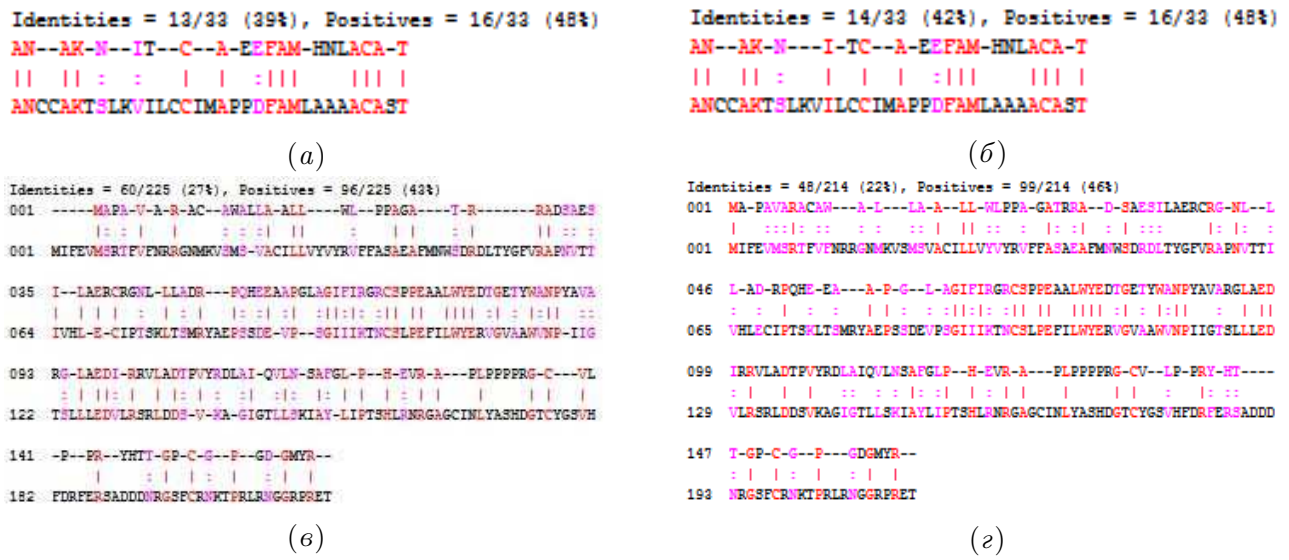

Рис. 4 Примеры оптимальных выравниваний двух пар аминокислотных последовательностей. Выравнивания найдены при помощи предложенного метода (*а* и *в*) и алгоритма Нидлмана–Вунша (*б* и *г*)

5.3 Классификация мембранных гликопротеинов

В данной работе используются 3 базовых способа сравнения последовательностей:

1) мера сходства Нидлмана–Вунша (NW) — $S_1(\omega', \omega'')$;
2) мера сходства Смита–Ватермана (SW) — $S_2(\omega', \omega'')$;
3) предложенная метрика (Metric) — $r(\omega', \omega'')$.

Для обеспечения возможности использования мер сходства Нидлмана–Вунша и Смита–Ватермана совместно с методом опорных векторов (SVM) для каждой из них был выполнен переход в пространство вторичных признаков:

$$K_i(\omega', \omega'') = \left[k_{lt} = \left(S_i^{\langle l \rangle} \right)^T S_i^{\langle t \rangle}, \quad l, t = 1, \ldots, N \right], \quad i = 1, 2.$$

Что касается предложенной метрики, то оказалось, что на рассматриваемом множестве аминокислотных последовательностей она является евклидовой, что позволяет использовать радиальную функцию вида $K_3(\omega', \omega'') = \exp(-\alpha r^2(\omega', \omega''))$. Значение параметра α во всех проведенных экспериментах было установлено равным 0,01.

Следует обратить внимание, что в общем случае предложенный способ построения метрики не гарантирует наличие свойства евклидовости и, соответственно, преобразование $\exp(-\alpha r^2(\omega', \omega''))$ может привести к наличию отрицательных собственных чисел. Однако на практике для $r(\omega', \omega'') = r_2(\omega', \omega'')$ свойство евклидовости обычно выполняется.

Для каждого из трех указанных способов сравнения решались задачи обучения двухклассовому распознаванию каждого из классов (1, 2 и 3) от оставшихся и каждого семейства (hpf 12, 20, 42, 47, 50, 114, 531) от оставшихся, а также задачи обучения попарному распознаванию классов и семейств из табл. 1. Обучение проводилось при помощи SVM.

Качество построенных решающих правил оценивалось по скользящему контролю.

В табл. 2 и 3 приведены проценты ошибок, полученных на скользящем контроле, для случаев, в которых хотя бы два из трех способов сравнения дали различный результат. Жирным выделен лучший результат в каждой строке.

Таблица 2 Проценты ошибок на скользящем контроле при распознавании «один против всех»

Задача	NW	SW	Metric
hpf 12	15,0215	15,0215	**14,5923**
hpf 20	0,4292	**0**	**0**
hpf 42	**0**	0,4292	0,4292
hpf 47	4,721	**0**	**0**
hpf 50	0,4292	**0**	**0**
hpf 114	4,721	0,8584	**0,4292**
hpf 531	**15,0125**	**15,0125**	18,4549
класс 1	0,8584	**0,4292**	**0,4292**
класс 2	0,8584	**0,4292**	**0,4292**
класс 3	0,4292	**0**	**0**

Таблица 3 Проценты ошибок на скользящем контроле при попарном распознавании классов и семейств hpf

Задача	NW	SW	Metric
класс 2 vs класс 3	12,3256	**0**	**0**
hpf 42 vs hpf 47	0,4292	**0**	**0**
hpf 42 vs hpf 114	**0**	1,9231	**0**
hpf 47 vs hpf 114	2,3256	**0**	**0**
hpf 531 vs hpf 12	**48,5714**	51,4286	50,000
hpf 531 vs hpf 42	**1,7544**	3,5088	**1,7544**

Как видно из табл. 2 и 3, предложенный способ сравнения позволяет в большинстве случаев получить наилучший результат, в остальных случаях — близкий к наилучшему, что говорит о его адекватности прикладным задачам анализа аминокислотных последовательностей. Кроме того, очень важным достоинством данного способа сравнения является то, что его применение совместно с SVM требует запоминания и использования при распознавании лишь небольшого количества опорных объектов, а не всех объектов обучающей совокупности, как при использовании традиционных мер сходства Нидлмана–Вунша и Смита–Ватермана, что делает его более эффективным по памяти и скорости работы.

6 Заключение

В данной работе предложен простой способ построения метрики на множестве биомолекулярных последовательностей, основанный, как и традиционные методы, на поиске оптимального парного выравнивания. При соблюдении необременительного условия относительно величины штрафа на пропуск элементов при выравнивании получаемая в результате мера несходства гарантированно обладает свойствами метрики. Доказательства соответствующих теорем приведены в данной статье. Кроме того, в ряде практических случаев (для конкретных конечных множеств последовательностей) данная мера несходства может обладать и свойствами евклидовой метрики, что делает ее применение в сочетании с SVM особенно удобным.

Экспериментальное исследование показало адекватность данного способа сравнения прикладным задачам распознавания аминокислотных последовательностей. Кроме того, очень важным достоинством использования метрики в качестве способа сравнения является то, что его применение совместно с SVM требует запоминания и использования при распознавании лишь небольшого количества опорных объектов, а не всех объектов обучающей совокупности, как при использовании традиционных мер сходства Нидлмана–Вунша и Смита–Ватермана, что делает его более эффективным по памяти и скорости работы.

Приложение 1

Доказательство теоремы 1

Согласно определению (4) и условию теоремы 1 условие неотрицательности $\tilde{\rho}(\alpha', \alpha'') \geqslant 0$ выполняется для всех элементов из расширенного множества $\alpha', \alpha'' \in \tilde{A}$.

Покажем, что для любой тройки элементов $\alpha',\ \alpha'', \alpha''' \in \tilde{A}$ выполняется неравенство треугольника

$$\tilde{\rho}(\alpha', \alpha'') + \tilde{\rho}(\alpha'', \alpha''') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha'', \alpha''' \in \tilde{A}. \tag{10}$$

Рассмотрим все возможные способы вхождения пропуска «$-$» в тройку элементов $\alpha', \alpha'', \alpha''' \in \in \tilde{A}$ и покажем, что для каждого из этих случаевнеравенство треугольника выполняется:

а) α'''«$-$», $\alpha', \alpha'' \neq$«$-$».

Неравенство треугольника (10) в данном случае принимает вид:

$$\tilde{\rho}(\alpha', \alpha'') + \tilde{\rho}(\alpha'', -) \geqslant\ \tilde{\rho}(\alpha', -)\ \forall\ \alpha',\ \alpha'' \in A.$$

Подставив $\tilde{\rho}(\alpha', -) = \tilde{\rho}(\alpha'', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$, получим:

$$\tilde{\rho}(\alpha', \alpha'') + \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'').$$

Очевидно, что это условие выполняется для любых $\alpha',\ \alpha'' \in A$, поскольку $\tilde{\rho}(\alpha', \alpha'')\ \geqslant\ 0$;

б) $\alpha' =$«$-$», $\alpha'', \alpha''' \neq$«$-$». Этот случай абсолютно аналогичен предыдущему. Выполняя аналогичную подстановку $\tilde{\rho}(\alpha'', -) = \tilde{\rho}(\alpha''', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$ в неравенство треугольника, получим:

$$\frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') + \tilde{\rho}(\alpha'', \alpha''') \geqslant \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'').$$

Соответственно, поскольку $\tilde{\rho}(\alpha'', \alpha''') \geqslant 0\ \forall\ \alpha'', \alpha''' \in \tilde{A}$, в данном случае неравенство треугольника тоже выполняется;

в) $\alpha''\ =$«$-$», $\alpha', \alpha'''\ \neq$«$-$». Неравенство треугольника (10) после подстановки $\tilde{\rho}(\alpha', -)\ = \\ = \tilde{\rho}(\alpha''', -) = (1/2) \max\limits_{\eta', \eta'' \in A} \rho(\eta', \eta'')$ примет вид:

$$\frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') + \frac{1}{2} \max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha''' \in A$$

или

$$\max_{\eta', \eta'' \in A} \rho(\eta', \eta'') \geqslant\ \tilde{\rho}(\alpha', \alpha''')\ \forall\ \alpha',\ \alpha''' \in A.$$

Очевидно, что данное неравенство всегда является верным;

г) $\alpha' = \alpha'' =$«$-$», $\alpha''' \neq$«$-$».

В данном случае неравенство треугольника

$$\tilde{\rho}(-,-) + \tilde{\rho}(-,\alpha''') \geqslant \tilde{\rho}(-,\alpha''') \; \forall \; \alpha',\; \alpha'',\alpha''' \in \tilde{A}$$

вырождается в верное равенство:

$$0 + \frac{1}{2} \max_{\eta',\eta'' \in A} \rho(\eta',\eta'') = \frac{1}{2} \max_{\eta',\eta'' \in A} \rho(\eta',\eta'');$$

д) $\alpha' = \alpha'' = \alpha''' =$«$-$». Данный случай является тривиальным, приводя после подстановки к тождеству $0 + 0 = 0$;

е) $\alpha' \neq$«$-$», $\alpha'' \neq$«$-$», $\alpha''' \neq$«$-$».

В данном случае все три элемента принадлежат исходному множеству элементов последовательностей $\alpha',\; \alpha'',\; \alpha''' \in A$, на котором определена метрика $\rho(\alpha',\alpha'') : A \times A \to R$ согласно (2) и, поскольку, согласно (4) $\tilde{\rho}(\alpha',\alpha'') = \rho(\alpha',\alpha'') \; \forall \; \alpha',\alpha'' \in A$, то неравенство треугольника в данном случае тоже выполняется.

Таким образом, для любой тройки элементов $\alpha',\; \alpha'',\alpha''' \in \tilde{A}$ неравенство треугольника выполняется.

Теорема доказана.

Приложение 2

Доказательство теоремы 2

Для доказательства выполнения неравенства треугольника для мер несходства (8) и (9) рассмотрим три последовательности $\omega' = (\alpha'_1, \alpha'_2, \ldots, \alpha'_{N'}) \in \Omega,\; \omega'' = (\alpha''_1, \alpha''_2, \ldots, \alpha''_{N''}) \in \Omega$ и $\omega''' = (\alpha'''_1, \alpha'''_2, \ldots, \alpha'''_{N'''}) \in \Omega$.

Пусть $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''')$ — два оптимальных выравнивания соответствующих пар последовательностей, например следующих:

$$\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'') : \quad \left\{ \begin{array}{cccccc} \hat{\mathbf{w}}_1^{1,2} & \hat{\mathbf{w}}_2^{1,2} & \hat{\mathbf{w}}_3^{1,2} & \hat{\mathbf{w}}_4^{1,2} & \hat{\mathbf{w}}_5^{1,2} & \hat{\mathbf{w}}_6^{1,2} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & - & \alpha'_4 \\ -\alpha''_1 & - & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 \end{array} \right. ;$$

$$\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''') : \quad \left\{ \begin{array}{cccccc} \hat{\mathbf{w}}_1^{2,3} & \hat{\mathbf{w}}_2^{2,3} & \hat{\mathbf{w}}_3^{2,3} & \hat{\mathbf{w}}_4^{2,3} & \hat{\mathbf{w}}_5^{2,3} & \hat{\mathbf{w}}_6^{2,3} \\ \alpha''_1 & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 & - \\ - & \alpha'''_1 & \alpha'''_2 & - & \alpha'''_3 & \alpha'''_4 \end{array} \right. .$$

Очевидно, что два таких выравнивания однозначно определяют третье выравнивание $\mathbf{w}^{1,3} = \mathbf{w}(\omega',\omega''')$, сопоставляющее элементы последовательностей ω' и ω''', а также выравнивание элементов всех трех последовательностей сразу $\mathbf{w}^{1,2,3} = \mathbf{w}(\omega',\omega'',\omega''')$:

$$\mathbf{w}^{1,3} = \mathbf{w}(\omega',\omega''') : \quad \left\{ \begin{array}{cccccc} \mathbf{w}_1^{1,3} & \mathbf{w}_2^{1,3} & \mathbf{w}_3^{1,3} & \mathbf{w}_4^{1,3} & \mathbf{w}_5^{1,3} & \mathbf{w}_6^{1,3} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & \alpha'_4 & - \\ - & - & \alpha'''_1 & \alpha'''_2 & \alpha'''_3 & \alpha'''_4 \end{array} \right. ;$$

$$\mathbf{w}^{1,2,3} = \mathbf{w}(\omega',\omega'',\omega''') : \quad \left\{ \begin{array}{ccccccc} \mathbf{w}_1^{1,2,3} & \mathbf{w}_2^{1,2,3} & \mathbf{w}_3^{1,2,3} & \mathbf{w}_4^{1,2,3} & \mathbf{w}_5^{1,2,3} & \mathbf{w}_6^{1,2,3} & \mathbf{w}_7^{1,2,3} \\ \alpha'_1 & \alpha'_2 & - & \alpha'_3 & - & \alpha'_4 & - \\ \alpha''_1 & - & \alpha''_2 & \alpha''_3 & \alpha''_4 & \alpha''_5 & - \\ - & - & \alpha'''_1 & \alpha'''_2 & - & \alpha'''_3 & \alpha'''_4 \end{array} \right. .$$

Следует обратить внимание, что выравнивания $\mathbf{w}^{1,3}$ и $\mathbf{w}^{1,2,3}$ в отличие от $\hat{\mathbf{w}}^{1,2}$ и $\hat{\mathbf{w}}^{2,3}$ в общем случае не являются оптимальными.

Каждое из рассмотренных выравниваний порождает свои расширенные последовательности в соответствии с (3):

$$
\begin{array}{ll}
\hat{\mathbf{w}}^{1,2}: & \tilde{\omega}'(\hat{\mathbf{w}}^{1,2}) = \{\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{1,2}}\}, \quad \tilde{\omega}''(\hat{\mathbf{w}}^{1,2}) = \{\tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{1,2}}\};\\
\hat{\mathbf{w}}^{2,3}: & \tilde{\omega}''(\hat{\mathbf{w}}^{2,3}) = \{\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{2,3}}\}, \quad \tilde{\omega}'''(\hat{\mathbf{w}}^{2,3}) = \{\tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}, i = 1,\ldots,N_{\hat{\mathbf{w}}^{2,3}}\};\\
\mathbf{w}^{1,3}: & \tilde{\omega}'(\mathbf{w}^{1,3}) = \{\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, i = 1,\ldots,N_{\mathbf{w}^{1,3}}\}, \quad \tilde{\omega}'''(\mathbf{w}^{1,3}) = \{\tilde{\alpha}'''_{\mathbf{w}^{1,3},i}, i = 1,\ldots,N_{\mathbf{w}^{1,3}}\};\\
\mathbf{w}^{1,2,3}: & \left\{\begin{array}{l}
\tilde{\omega}'(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\},\\
\tilde{\omega}''(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\},\\
\tilde{\omega}'''(\mathbf{w}^{1,2,3}) = \{\tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\ ,\ i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}\}.
\end{array}\right.
\end{array}
$$

Для каждой пары последовательностей (ω',ω''),$(\omega'',\omega]''$) и (ω',ω''') рассмотрим векторы одинаковой длины $N_{\mathbf{w}^{1,2,3}}$, составленные из значений метрики (4) между их элементами, поставленными в соответствие выравниванием $\mathbf{w}^{1,2,3}$:

$$
\begin{aligned}
\mathbf{x}^{1,2} &= \mathbf{r}_1(\omega',\omega''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}};\\
\mathbf{x}^{2,3} &= \mathbf{r}_1(\omega'',\omega'''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}''_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}};\\
\mathbf{x}^{1,3} &= \mathbf{r}_1(\omega',\omega'''|\mathbf{w}^{1,2,3}) = [\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,2,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,2,3},i}\right), \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}]^{\mathrm{T}}.
\end{aligned}
$$

Заметим, что, согласно свойству метрики (2), для каждой тройки $i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}$ элементов этих векторов справедливы неравенства:

$$
x_i^{1,2} + x_i^{2,3} \geqslant x_i^{1,3}, \quad i = 1,\ldots,N_{\mathbf{w}^{1,2,3}}.
$$

При этом очевидно, что также будет справедливо неравенство:

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2} + \sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3} \geqslant \sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3}. \tag{11}
$$

Кроме того, нетрудно убедиться, что в этом случае евклидова норма вектора $\mathbf{x}^{1,3}$ не может превосходить сумму евклидовых норм векторов $\mathbf{x}^{1,2}$ и $\mathbf{x}^{2,3}$:

$$
\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{1,2}\right)^2} + \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{2,3}\right)^2} \geqslant \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} \left(x_i^{1,3}\right)^2}. \tag{12}
$$

Заметим, что значения $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2}$ и $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3}$, входящие в неравенство (11), равны, соответственно, несходству последовательностей (ω',ω'') и (ω'',ω'''), определяемому согласно (8) на основе их оптимальных выравниваний $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega',\omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'',\omega''')$, поскольку $\tilde{\rho}(-,-) = 0$:

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,2} = \sum_{i=1}^{N_{\hat{\mathbf{w}}^{1,2}}} \tilde{\rho}\left(\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, \tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}\right) = r_1(\omega',\omega''|\hat{\mathbf{w}}^{1,2}) = \min_{\mathbf{w}\in W_{N'N''}} r_1(\omega',\omega''|\mathbf{w}) = r_1(\omega',\omega'');
$$

$$
\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{2,3} = \sum_{i=1}^{N_{\hat{\mathbf{w}}^{2,3}}} \tilde{\rho}\left(\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, \tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}\right) = r_1(\omega'',\omega'''|\hat{\mathbf{w}}^{2,3}) = \min_{\mathbf{w}\in W_{N''N'''}} r_1(\omega'',\omega'''|\mathbf{w}) = r_1(\omega'',\omega''').
$$

Аналогично, значения $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,2}\right)^2}$ и $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{2,3}\right)^2}$, входящие в неравенство (12), равны, соответственно, несходству последовательностей (ω', ω'') и (ω'', ω'''), определяемому согласно (9) для тех же оптимальных выравниваний $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$:

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,2}\right)^2} = \sqrt{\sum_{i=1}^{N_{\hat{\mathbf{w}}^{1,2}}}\tilde{\rho}^2\left(\tilde{\alpha}'_{\hat{\mathbf{w}}^{1,2},i}, \tilde{\alpha}''_{\hat{\mathbf{w}}^{1,2},i}\right)} = r_2(\omega', \omega''|\hat{\mathbf{w}}^{1,2}) = r_2(\omega', \omega'');$$

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{2,3}\right)^2} = \sqrt{\sum_{i=1}^{N_{\hat{\mathbf{w}}^{2,3}}}\tilde{\rho}^2\left(\tilde{\alpha}''_{\hat{\mathbf{w}}^{2,3},i}, \tilde{\alpha}'''_{\hat{\mathbf{w}}^{2,3},i}\right)} = r_2(\omega'', \omega'''|\hat{\mathbf{w}}^{2,3}) = r_2(\omega'', \omega''').$$

При этом следует заметить, что для пары последовательностей (ω', ω''') оптимальное выравнивание не определено и значения $\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3}$ и $\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,3}\right)^2}$ для них равны определяемым, соответственно, согласно (6) и (7) значениям несходства, условного относительно выравнивания $\mathbf{w}^{1,3}$, порожденного оптимальными выравниваниями $\hat{\mathbf{w}}^{1,2} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$:

$$\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}} x_i^{1,3} = \sum_{i=1}^{N_{\mathbf{w}^{1,3}}}\tilde{\rho}\left(\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,3},i}\right) = r_1(\omega', \omega'''|\mathbf{w}^{1,3});$$

$$\sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,2,3}}}\left(x_i^{1,3}\right)^2} = \sqrt{\sum_{i=1}^{N_{\mathbf{w}^{1,3}}}\tilde{\rho}^2(\tilde{\alpha}'_{\mathbf{w}^{1,3},i}, \tilde{\alpha}'''_{\mathbf{w}^{1,3},i})} = r_2\left(\omega', \omega'''|\mathbf{w}^{1,3}\right).$$

Таким образом, выполняются неравенства:

$$r_1(\omega', \omega'') + r_1(\omega'', \omega''') \geqslant r_1(\omega', \omega'''|\mathbf{w}^{1,3}),$$

$$r_2(\omega', \omega'') + r_2(\omega'', \omega''') \geqslant r_2(\omega', \omega'''|\mathbf{w}^{1,3}).$$

Более того, данные неравенства останутся верными и в случае, если вместо выравнивания $\mathbf{w}^{1,3} = \mathbf{w}(\omega', \omega''')$, порожденного выравниваниями $\hat{\mathbf{w}}^{1,3} = \hat{\mathbf{w}}(\omega', \omega'')$ и $\hat{\mathbf{w}}^{2,3} = \hat{\mathbf{w}}(\omega'', \omega''')$, рассмотреть оптимальное выравнивание $\hat{\mathbf{w}}^{1,3} = \hat{\mathbf{w}}(\omega', \omega''')$ последовательностей ω' и ω'', поскольку оптимальное выравнивание по определению обеспечивает значение несходства, не превышающее значения, вычисленного для любого другого варианта выравнивания последовательностей:

$$r_1(\omega', \omega''') = r_1(\omega', \omega'''|\hat{\mathbf{w}}^{1,3}) \leqslant r_1(\omega', \omega'''|\mathbf{w}^{1,3});$$

$$r_2(\omega', \omega''') = r_2(\omega', \omega'''|\hat{\mathbf{w}}^{1,3}) \leqslant r_2(\omega', \omega'''|\mathbf{w}^{1,3}).$$

Следовательно, для любой тройки последовательностей неравенства треугольников $r_1(\omega', \omega'') + r_1(\omega'', \omega''') \geqslant r_1(\omega', \omega''')$ и $r_2(\omega', \omega'') + r_2(\omega'', \omega''') \geqslant r_2(\omega', \omega''')$ выполняются, и меры несходства, определенные согласно (8) и (9), являются метриками на множестве последовательностей.

Теорема доказана.

Литература

[1] *Needleman S. B., Wunsch C. D.* A general method applicable to the search for similarities in the amino acid sequence of two proteins // J. Mol. Biol., 1970. Vol. 48. No. 3. P. 443–453. doi: 10.1016/0022-2836(70)90057-4.

[2] *Smith T. F., Waterman M. S.* Identification of common molecular subsequences // J. Mol. Biol., 1981. Vol. 147. No. 1. P. 195–197. doi: 10.1016/0022-2836(81)90087-5.

[3] *Zhang Z., Schwartz S., Wagner L., Miller W.* A greedy algorithm for aligning DNA sequences // J. Comput. Biol., 2000. No. 7. P. 203–14. doi: 10.1089/10665270050081478.

[4] *Дурбин Р., Эдди Ш., Крог А., Митчисон Г.* Анализ биологических последовательностей / Пер. с англ. — М.: Ижевск, 2006. 480 с. (*Durbin R., Eddy S., Krogh A., and Mitchison G.* Biological sequence analysis: Probabilistic models of proteins and nucleic acids. — Cambridge Univesrity Press, 1998. 356 p.)

[5] *Vapnik V. N.* Statistical learning theory. — Wiley-Interscience, 1998. 768 p.

[6] *Mottl V. V., Dvoenko S. D., Seredin O. S., Kulikowski C. A., Muchnik I. B.* Alignment scores in a regularized support vector classification method for fold recognition of remote protein families. — Center for Discrete Mathematics and Theoretical Computer Science. Rutgers University, State University of New Jersey, 2001. 33 p.

[7] *Pekalska E., Paclic P., Duin R.* A generalized kernel approach to dissimilarity-based classification // J. Mach. Learn. Res., 2001. Vol. 2. P. 175–211.

[8] *Liao Li, Noble W. S.* Combining pairwise sequence similarity and support vector machines for remote protein homology detection // 6th Annual Conference (International) on Computational Molecular Biology Proceedings, 2002. P. 225–232.

[9] *Schölkopf B., Tsuda K., Vert J.-P.* Kernel methods in computational biology. — MIT Press, 2004. 410 p.

[10] *Ben-Hur A., Ong C. S., Sonnenburg S., Schölkopf B., Rätsch G.* Support vector machines and kernels for computational biology // PLoS Comput. Biol., 2008. Vol. 4. No. 10. P. 1–10.

[11] *Середин О. С.* Линейные методы распознавания образов на множествах объектов произвольной природы, представленных попарными сравнениями. Общий случай // Известия ТулГУ, Серия Естественные науки. — Тула: Изд-во ТулГУ, 2012. Т. 1. С. 141–152.

[12] *Айзерман М. А., Браверман Э. М., Розоноэр Л. И.* Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. 384 с.

[13] *Сулимова В. В.* Потенциальные функции для анализа сигналов и символьных последовательностей разной длины. — Тула, 2009. Дисс. ... канд. наук. 122 с.

[14] *Моттль В. В.* Метрические пространства, допускающие введение линейных операций и скалярного произведения // Докл. РАН, 2003. Т. 388. №3. С. 312–315.

[15] *Leslie C., Eskin E., and Noble W.* The spectrum kernel: A string kernel for SVM protein classification // Pacific Symposium on Biocomputing Proceedings, 2002. P. 564–575.

[16] *Qiu J., Hue M., Ben-Hur A., Vert J.-P., Noble W. S.* A structural alignment kernel for protein structures // Bioinformatics, 2007. Vol. 23. №9. P. 1090–1098.

[17] *Sun L., Ji S., Ye J.* Adaptive diffusion kernel learning from biological networks for protein function prediction // BMC Bioinformatics, 2008. Vol. 9. No. 1. P. 1–14. doi: 10.1186/1471-2105-9-162. `http://www.biomedcentral.com/1471-2105/9/162`.

[18] *Miklós I., Novak A., Satija R., Lyngsø R., Hein J.* Stochastic models of sequence evolution including insertion-deletion events // Stat. Methods Med. Res., 2009. Vol. 18. P. 453–485. `http://ramet.elte.hu/~miklosi/StatAlignReview2008.pdf`.

[19] *Onodera T., Shibuya T.* The gapped spectrum kernel for support vector machines // 9th Conference (International) MLDM Proceedings. — Berlin – Heidelberg – New York: Springer, 2013. P. 1–15. doi: 10.1007/978-3-642-39712-7_1.

[20] *Baisero A., Pokorny F. T., Ek C.H.* On a family of decomposable kernels on sequences // CoRR, 2015. arXiv:/1501.06284.

[21] *Seeger M.* Covariance kernels from baye ian generative models // Stat. Methods Med. Res., 2009. Vol. 18. P. 453–485.

[22] *Абрамов В. И., Середин О. С., Сулимова В. В.* Метод опорных объектов для обучения распознаванию образов в евклидовых метрических пространствах // Международная конференция «Интеллектуализация обработки информации» (ИОИ-9). — Черногория, 2012. С. 5–8.

[23] *Абрамов В. И., Середин О. С., Моттль В. В.* Обучение распознаванию образов по методу опорных объектов в евклидовых метрических пространствах с аффинными операциями // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2013, Вып. 2. Ч. 1. С. 119–136.

[24] *Hein M., Bousquet O., Schölkopf B.* Maximal margin classification for metric spaces // J. Comput. Syst. Sci., 2005. Vol. 71. Iss. 3. P. 333–359.

[25] *Xu W.* Non-euclidean dissimilarity data in pattern recognition. 2012. Ph.D. Thesis.

[26] *Середин О. С., Абрамов В. И., Моттль В. В.* Аффинные операции в псевдоевклидовом линейном пространстве // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2014. Вып. 3. С. 178–196.

[27] *Hancock E. R., Xu E., Wilson R. C.* Pattern recognition with non-Euclidean similarities // Man–Machine Interactions, 2014. Vol. 3. P. 3–15.

[28] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Известия ТулГУ, Естественные науки. — Тула: Изд-во ТулГУ, 2015. Вып. 4. С. 49–66.

[29] *Pekalska E. M.* Dissimilarity representations in pattern recognition. Concepts, theory and applications. 2005. PhD Thesis. 344 p.

[30] *Bellet A., Harbrad A., Sebban M.* A survey on metric learning for feature vectors and structured data // CoRR, 2013. abs/1306.6709. http://arxiv.org/abs/1306.6709.

[31] *Wang J., Sun K., Sha F., Marchand-Maillet S., Kalousis K.* Two-stage metric learning // 31st Conference (International) on Machine Learning Proceedings, Cycle 2. — JMLR.org, 2014. Vol. 32. P. 370–378. http://jmlr.org/proceedings/papers/v32/wangc14.html.

[32] *Xing E. P., Ng A. Y., Jordan M. I., Russel S.* Distance metric learning, with application to clustering with side-information // Advances in neural information processing systems 15 / Eds. S. Becker, S. Thrun, K. Obermayer. — MIT Press, 2003. P. 521–528. http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf.

[33] *Schultz M., Joachims T.* Learning a distance metric from relative comparisons // Advances in neural information processing systems 16 / Eds. S. Thrun, L. K. Saul, B. Schölkopf. — MIT Press, 2004. P. 41–48. http://papers.nips.cc/paper/2366-learning-a-distance-metric-from-relative-comparisons.pdf.

[34] *Wang J., Do H., Woznica A., Kalousis A.* Metric learning with multiple kernels // Advances in neural information processing systems 24 / Eds. J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett. — MIT Press, 2011. P. 1170–1178. http://papers.nips.cc/paper/4399-metric-learning-with-multiple-kernels.pdf.

[35] *Dayhoff M., Schwarts R., Orcutt B.* A model of evolutionary change in proteins // Atlas of protein sequences and structures. — National Biometrical Research Foundation, 1978. Vol. 5. Suppl. 3. P. 345–352.

[36] Virus Database at University College London (VIDA). http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA3/VIDA.html.

[37] *McGeoch D. J., Rixon F. J., Davison A. J.* Topics in herpesvirus genomics and evolution // Virus Res., 2006. No. 117. P. 90–104. doi: 10.1016/j.virusres.2006.01.002.

Поступила в редакцию 31.08.2016

Metrics on the basis of optimal alignment of biomolecular sequences*

***V. V. Sulimova*[1], *O. S. Seredin*[1], *and V. V. Mottl*[2]**
vsulimova@yandex.ru; oseredin@yandex.ru; vmottl@yandex.ru
[1]Tula State University, 92 Lenina Ave., Tula, Russia
[2]Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

Background: It is important for biomolecular sequences analysis to have an appropriate way for comparing them. From the point of view of advanced methods of data analysis, the most preferred way for comparing objects is a dissimilarity measure, possessing metric's properties. From the other side, from the point of view of the molecular biology, it is important to take into account biological features of the compared objects. Besides, the computational effectiveness and the possibility of further using convenient instruments of data analysis are also important. There are a number of ways for comparing biomolecular sequences, though no one of them possess the all required properties.

Methods: This paper proposes a simple enough way for computing metrics for biomolecular sequences. The proposed approach, following traditional ways for biomolecular sequences comparing, is based on finding an optimal pairwise alignment and on the model of mutual changes of amino acids at the process of evolution.

Concluding Remarks: It is proved that the proposed dissimilarity measure is a metric. So, it can be used at the advanced methods of data analysis, saving computational advantages of support vector machine without introducing features of objects or(and) an inner product. The experimental results confirm usability of the proposed metric for membrane glycoprotein classification.

Keywords: *metric; comparing sequences; optimal sequence alignment; biomolecular sequences*

DOI: 10.21469/22233792.2.3.03

References

[1] Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443–453. doi: 10.1016/0022-2836(70)90057-4.

[2] Smith, T. F., and M. S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147(1):195–197. doi: 10.1016/0022-2836(81)90087-5.

[3] Zhang, Z., S. Schwartz, L. Wagner, and W. Miller. 2000. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7:203–14. doi: 10.1089/10665270050081478.

[4] Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press. 356 p.

[5] Vapnik, V. N. *Statistical learning theory*. Wiley-Interscience, 1998. 768 p.

[6] Mottl, V. V., S. D. Dvoenko, O. S. Seredin, C. A. Kulikowski, and I. B. Muchnik. 2001. *Alignment scores in a regularized support vector classification method for fold recognition of remote protein families*. Center for Discrete Mathematics and Theoretical Computer Science. Rutgers University, State University of New Jersey. 33 p.

*The research was supported by the Russian Foundation for Basic Research (grant 15-07-08967).

[7] Pekalska, E., P. Paclic, and R. Duin. 2001. A generalized kernel approach to dissimilarity-based classification. *J. Mach. Learn. Res.* 2:175–211.

[8] Liao, Li, and W. S. Noble. 2002. Combining pairwise sequence similarity and support vector machines for remote protein homology detection // *6th Annual Conference (International) on Computational Molecular Biology Proceedings*. 225–232.

[9] Schölkopf, B., K. Tsuda, and J.-P.Vert. 2004. *Kernel methods in computational biology*. MIT Press. 410 p.

[10] Ben-Hur, A., C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. 2008. Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* 4(10):1–10.

[11] Seredin, O. S. Linear methods of pattern recognition for sets of objeacts of arbitrary kinds, represented by pairwise compariosions. The general case. *Proceedings of Tula State University. Natural sciences ser.* 1:141–152.

[12] Aizerman, M. A., E. M. Braverman, L. I. Rozonoer. 1970. *Metod potentsial'nykh funktsiy v teorii obucheniya mashin* [Potential functions method in machine learning theory]. Moscow: Nauka. 384 p.

[13] Sulimova, V. V. 2009. Potentsial'nye funktsii dlya analiza signalov i simvol'nykh posledovatel'nostey raznoy dliny [Kernel functions for analysis of signals and symbolic sequences of different length]. Tula. PhD Thesis. 122 p.

[14] Mottl, V. V. 2003. Metricheskie prostranstva, dopuskayushchie vvedenie lineynykh operatsiy i skalyarnogo proizvedeniya [Metric spaces admitting linear operations and inner product]. *Dokl. RAS* 388(3):312–315.

[15] Leslie, C., E. Eskin, and W. Noble. 2002. The spectrum kernel: A string kernel for SVM protein classification. *Pacific Symposium on Biocomputing Proceedings*. 564–575.

[16] Qiu, J., M. Hue, A. Ben-Hur, J.-P. Vert, and W. S. Noble. 2007. A structural alignment kernel for protein structures. *Bioinformatics* 23(9):1090–1098.

[17] Sun, L., S. Ji, and J. Ye. 2008. Adaptive diffusion kernel learning from biological networks for protein function prediction. *BMC Bioinformatics* 9(1):1–14. doi: 10.1186/1471-2105-9-162. Available at: `http://www.biomedcentral.com/1471-2105/9/162` (accessed December 27, 2016).

[18] Miklós, I., A. Novák, R. Satija, R. Lyngsø, and J. Hein. 2009. Stochastic models of sequence evolution including insertion-deletion events. *Stat. Methods Med. Res.* 18:453–485. Available at: `http://ramet.elte.hu/~miklosi/StatAlignReview2008.pdf` (accessed December 27, 2016).

[19] Onodera T., and T. Shibuya. 2013. The gapped spectrum kernel for support vector machines. *9th Conference (International) MLDM Proceedings*. Berlin – Heidelberg – New York: Springer. 1–15. doi: 10.1007/978-3-642-39712-7_1.

[20] Baisero, A., F. T. Pokorny, and C. H. Ek. 2015. On a family of decomposable kernels on sequences. *CoRR*. arXiv:/1501.06284.

[21] Seeger, M. 2009. Covariance kernels from baye ian generative models. *Stat. Methods Med. Res.* 18:453–485.

[22] Abramov, V. I., O. S. Seredin, and V. V. Sulimova. 2012. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v evklidovykh metricheskikh prostranstvakh [Method of support objects for pattern recognition in Euclidean metric spaces]. *Conference (International) "Intellectualization of Data Processing"*. Montenegro. 5–8.

[23] Abramov, V. I., O. S. Seredin, and V. V. Mottl. 2013. Obuchenie raspoznavaniyu obrazov po metodu opornykh ob"ektov v evklidovykh metricheskikh prostranstvakh s affinnymi operatsiyami [Pattern recognition training with support vector object method in Euclidean metric spaces with

affine operations]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 2(1):119–136.

[24] Hein, M., O. Bousquet, and B. Schölkopf. 2005. Maximal margin classification for metric spaces. *J. Comput. Syst. Sci.* 71(3):333–359.

[25] Xu, W. 2012. Non-Euclidean dissimilarity data in pattern recognition. Ph.D. Thesis.

[26] Seredin, O. S., V. I. Abramov, and V. V. Mottl. 2014. Affinnye operatsii v psevdoevklidovom lineynom prostranstve [Affine operations in pseudoeuclidean linear space]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 3:178–196.

[27] Hancock, E. R., E. Xu, and R. C. Wilson. 2014. Pattern recognition with non-Euclidean similarities. *Man–Machine Interactions* 3:3–15.

[28] Seredin, O. S., and V. V. Mottl. 2015. Metod opornykh ob"ektov dlya obucheniya raspoznavaniyu obrazov v proizvol'nykh metricheskikh prostranstvakh [Support object method for pattern recognition training in arbitrary metric spaces]. *Proceedings of Tula State University. Natural sciences ser.* Tula: TSU. 4:178–196.

[29] Pekalska, E. M. 2005. Dissimilarity representations in pattern recognition. Concepts, theory and applications. PhD Thesis. 344 p.

[30] Bellet, A., A. Harbrad, and M. Sebban. 2013. A survey on metric learning for feature vectors and structured data. *CoRR.* abs/1306.6709. Available at: `http://arxiv.org/abs/1306.6709` (accessed December 27, 2016).

[31] Wang, J., K. Sun, F. Sha, S. Marchand-Maillet, and K. Kalousis. 2014. Two-stage metric learning. *31st Conference (International) on Machine Learning Proceedings, Cycle 2.* JMLR.org. 32:370–378. Available at: `http://jmlr.org/proceedings/papers/v32/wangc14.html` (accessed December 27, 2016).

[32] Xing, E. P., A. Y. Ng, M. I. Jordan, and S. Russel. 2003. Distance metric learning, with application to clustering with side-information. *Advances in neural information processing systems 15.* Eds. S. Becker, S. Thrun, and K. Obermayer. MIT Press. 521–528. Available at: `http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf` (accessed December 27, 2016).

[33] Schultz, M., and T. Joachims. 2004. Learning a distance metric from relative comparisons. *Advances in neural information processing systems 16.* Eds. S. Thrun, L. K. Saul, and B. Schölkopf. MIT Press. 41–48. Available at: `http://papers.nips.cc/paper/2366-learning-a-distance-metric-from-relative-comparisons.pdf` (accessed December 27, 2016).

[34] Wang, J., H. Do, A. Woznica, and A. Kalousis. 2011. Metric learning with multiple kernels. *Advances in neural information processing systems 24.* Eds. J. Shawe-Taylor, R. S. Zemel, and P. L. Bartlett. MIT Press. 1170–1178. Available at: `http://papers.nips.cc/paper/4399-metric-learning-with-multiple-kernels.pdf` (accessed December 27, 2016).

[35] Dayhoff, M., R. Schwarts, and B. Orcutt. 1978. A model of evolutionary change in proteins. *Atlas of protein sequences and structures.* National Biometrical Research Foundation. 5(3):345–352.

[36] Virus Database at University College London (VIDA). Available at: `http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA3/VIDA.html` (accessed December 27, 2016).

[37] McGeoch, D. J., F. J. Rixon, and A. J. Davison. 2006. Topics in herpesvirus genomics and evolution. *Virus Res.* 117:90–104. doi: 10.1016/j.virusres.2006.01.002.

Received August 31, 2016