DOI 10.37882/2223-2966.2025.05-2.14

СВОЙСТВА МЕТОДА MAHDS С УЧЁТОМ КОРРЕЛЯЦИЙ СИМВОЛОВ В КОНТЕКСТЕ ВЫРАВНИВАНИЯ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

PROPERTIES OF THE MAHDS METHOD CONSIDERING CHARACTER CORRELATIONS IN THE CONTEXT OF AMINO ACID SEQUENCES ALIGNMENT

D. Kostenko
E. Korotkov
M. Korotkova

Summary. One of the most important tasks in bioinformatics is multiple sequence alignment of nucleotide and amino acid sequences. Previously, we proposed a new mathematical method for the alignment of highly divergent sequences (MAHDS). In this study, the method was adapted to align amino acid sequences using information about correlations of neighboring symbols. Using model amino acid sequences, we have investigated some properties of the method that may allow one to choose the most appropriate parameters for its practical application. The data which make it possible to evaluate the effectiveness and convergence of an optimization procedure that maximizes the sequences similarity function in alignment is presented. The properties of distributions of alignment quality indicators at different stages of MAHDS are described.

Keywords: multiple sequence alignment, dynamic programming, objective function optimization, distribution.

Костенко Дмитрий Олегович
Аспирант, младший научный сотрудник, Национальный исследовательский ядерный университет «МИФИ» (Москва); Федеральный исследовательский центр «Фундаментальные основы биотехнологии» РАН (Москва)
dk0stenko@yandex.ru
Коротков Евгений Вадимович
Доктор биологических наук, профессор, Национальный исследовательский ядерный университет «МИФИ» (Москва); Федеральный исследовательский центр «Фундаментальные основы биотехнологии» РАН (Москва)
bioinf@yandex.ru
Короткова Мария Александровна
Кандидат технических наук, доцент, Национальный исследовательский ядерный университет «МИФИ» (Москва)
discretmath@gmail.com

Аннотация. Одной из важнейших задач в биоинформатике является построение множественных выравниваний нуклеотидных и аминокислотных последовательностей. Ранее нами был предложен новый математический метод выравнивания сильно дивергированных последовательностей (MAHDS). В данной работе метод был адаптирован для выравнивания аминокислотных последовательностей с использованием информации о корреляциях соседних символов. На модельных аминокислотных последовательностях мы исследовали некоторые свойства метода, которые могут позволить выбирать наиболее подходящие параметры при его практическом применении. Представлены данные, позволяющие оценить эффективность и сходимость оптимизационной процедуры, максимизирующей функцию сходства последовательностей в выравнивании. Описаны свойства распределений показателей качества выравниваний на разных этапах MAHDS.

Ключевые слова: множественное выравнивание, динамическое программирование, оптимизация целевой функции, распределение.

Введение

Множественное выравнивание символьных последовательностей (MSA) позволяет найти наиболее вероятные мутации (вставки, удаления и замены символов), которые предположительно породили данные последовательности из неизвестного общего предка (в физическом мире такой предок мог и не существовать, это математическая абстракция, обуславливающая сходство последовательностей). Благодаря этому в последовательностях выявляются подобные компоненты, что позволяет аннотировать и предсказывать структуру биологических последовательностей, таких как ДНК/РНК и белки [1].

Было показано, что задача построения MSA является NP-полной [1,2], следовательно не существует алгоритма её решения, имеющего не экспоненциальную сложность. По этой причине на практике для построения MSA используют различные эвристики, позволяющие получить приближённое решение за приемлемое время [1,3–5]. Большинство популярных методов MSA используют разные вариации прогрессивного подхода [6–10]. Его недостатком является то, что даже небольшие неточности в парных выравниваниях накапливаются и существенно влияют на итоговое MSA в случае выравнивания сильно дивергированных последовательностей. Ранее мы разработали собственный метод MAHDS, который избежал данной проблемы, отказавшись от прогрессивного подхода. Метод уже показал себя успешно при выравнивании промоторных последовательностей [11] и белковых семейств с низкой степенью гомологии [12,13].

Ранее использование информации о корреляциях соседних символов в контексте MAHDS было возможно только в отношении нуклеотидных последовательностей. Мы адаптировали метод, чтобы это стало возможно и для аминокислотных последовательностей и провели серию экспериментов на модельных аминокислотных последовательностях, чтобы исследовать некоторые зависимости и распределения показателей, вычисляющихся при использовании метода MAHDS. Полученные данные позволят корректно оценивать выравнивания реальных белковых последовательностей.

Обозначения и постановка задачи выравнивания

Рассмотрим две последовательности символов, принадлежащих алфавиту A. Обозначим эти последовательности S_1 и S_2 и будем индексировать их символы начиная с 1, например, $S_1[1] \in A$ — первый символ последовательности S_1. Функцию, принимающую в качестве аргумента последовательность и возвращающую её длину обозначим L. Тогда чтобы получить парное выравнивание последовательностей S_1 и S_2, необходимо дополнить их специальными символами — гэпами (обычно это символ «–») таким образом, чтобы для получившихся последовательностей $\overline{S_1}$ и $\overline{S_2}$ выполнялось $L(\overline{S_1}) = L(\overline{S_2}) = L_{SA}$, где L_{SA} — длина выравнивания. Для удобства будем кодировать символы алфавита A числами $1,2,\ldots,|A|$. Расширенный алфавит, включающий в себя гэп, обозначим $\overline{A}$, $|\overline{A}| = |A| + 1$, а гэп будем кодировать числом 0. Тогда если $\overline{S_1}[i] = 0$, а $\overline{S_2}[i] \neq 0$, говорят, что в этой позиции у $\overline{S_1}$ делеция (удаление) символа либо у $\overline{S_2}$ вставка символа. Если $\overline{S_1}[i] \neq 0$, $\overline{S_2}[i] \neq 0$ и $\overline{S_1}[i] = \overline{S_2}[i]$, говорят, что символы совпадают, а если при этом $\overline{S_1}[i] \neq \overline{S_2}[i]$, значит в данной позиции замена. Случай, когда $\overline{S_1}[i] = \overline{S_2}[i] = 0$ не имеет смысла.

Матрица замен с размерностью $|A| \times |A|$ устанавливает вес выравнивания символа x с символом y. Обозначим соответствующую весовую функцию $s(x,y)$. Тогда вес выравнивания S_1 и S_2: $F = \sum_{i=1}^{L(\overline{S_1})} s(\overline{S_1}[i], \overline{S_2}[i])$, где s доопределена на гэпах следующим образом. Если $\overline{S_1}[i] = 0$, тогда $s(\overline{S_1}[i], \overline{S_2}[i]) = -d$, где d — величина штрафа за начало разрыва, но если при этом $\overline{S_1}[i-1] = 0$, тогда $s(\overline{S_1}[i], \overline{S_2}[i]) = -e$, где e — величина штрафа за продолжение разрыва (обычно $d > e$); аналогично для $\overline{S_2}$. В оптимальном парном выравнивании гэпы вставляются в S_1 и S_2 таким образом, чтобы максимизировать вес выравнивания F.

В случае множественного выравнивания (MSA) по аналогии с парным каждая последовательность из множества $S = \{S_1, S_2, \ldots, S_n\}$ дополняется гэпами до одинаковой длины: $L_{SA} = L(\overline{S_1}) = L(\overline{S_2}) = \ldots = L(\overline{S_n})$. Произвольное выравнивание SA будем рассматривать как символьную матрицу размерности $n \times L_{SA}$, где $SA[i][j] = \overline{S_i}[j]$.

Подход MAHDS

Основная идея MAHDS состоит в том, чтобы найти формальное представление абстрактного «общего предка» для всех выравниваемых последовательностей S, и построить с ним парные выравнивания каждой последовательности $S_i, i = \overline{1,n}$. Полученные парные выравнивания можно сопоставить, чтобы получить искомое MSA.

Для формального представления «общего предка» в алгоритме MAHDS мы вводим искусственную последовательность SA и позиционно-весовую матрицу PWM. Последовательность SA имеет вид $1,2,\ldots,M$, где M — длина последовательности «общего предка». Таким образом SA содержит только индексы аминокислот «общего предка», а непосредственно аминокислоты, которые находятся в соответствующих позициях изначально не известны. В свою очередь матрица PWM имеет размерность $|A| \times M$. Её строки соответствуют кодам символов исходного алфавита, а столбцы — позициям в SA. Элемент $PWM[i][j]$ несёт следующий смысл: чем больше значение данного элемента, тем больше вероятность, что в j –той позиции «общего предка» находится символ, кодируемый число i.

Процедура преобразования PWM в MSA

Была разработана процедура f, однозначно выстраивающая MSA последовательностей S по заданной PWM.

Для этого строятся парные выравнивания $S_i, i = \overline{1,n}$ с SA, используя двухмерное динамическое программирование подобно алгоритму Нидлмана-Вунша [14], но в качестве матрицы замен используется PWM. При этом вычисляются веса выравниваний, представленные значениями функции сходства в последней строке и последнем столбце матрицы динамического программирования: $F_k = F_k\left(L\left(S_k\right), M\right)$. Для последовательности S_k:

$$F_k(i,j) = \max \begin{cases} F(i-1,j-1) + PWM\left(S_k[i], S_a[j]\right) \\ \begin{cases} F(i-1,j) - e, \text{ если } F(i-1,j) \\ \text{получен из } F(i-2,j) \\ F(i-1,j) - d, \text{ иначе} \end{cases} \\ \begin{cases} F(i,j-1) - e, if F(i,j-1) \\ \text{получен из } \mathrm{F}(i,j-2) \\ F(i,j-1) - d, \text{ иначе} \end{cases} \end{cases} \quad (1)$$

Далее выровненные экземпляры SA будем обозначать $\overline{SA_i}, i = \overline{1,n}$. Для построения MSA мы дополняем последовательности $\overline{S_i}$ гэпами таким образом чтобы в столбцах MSA получилось M столбцов, состоящих из символов, выровненных с символами SA в парных выравниваниях (среди них могут быть и гэпы, соответствующие делециям в последовательностях S). Кроме M основных столбцов в MSA также присутствует $L_{SA} - M$ столбцов, которые соответствуют вставкам в S. Если $\overline{SA_k}[i]$ — гэп, то $\overline{S_k}[i]$ — вставка. Это значит, что необходимо создать столбец, где у всех последовательностей (кроме k-той и других, у которых присутствует вставка в этой позиции) будут гэпы.

Для полученного MSA мы определяем вес F как среднее арифметическое F_k каждого из парных выравниваний.

$$F = \frac{\sum_{k=1}^{n} F_k}{n} \quad (2)$$

Процедура преобразования MSA в PWM

Также была разработана обратная процедура f^{-1}, отображающая MSA в одну из таких PWM, которые при применении f привели бы к построению данной MSA. Для этого из MSA удаляются столбцы, в которых более половины символов — гэпы. Полученное сокращённое выравнивание обозначим MSA'. Каждому столбцу MSA' мы ставим в соответствие символ из SA $(1,2,\ldots,M)$, таким образом M равно количеству столбцов MSA' (если изначально известны $\overline{SA_i}, i = \overline{1,n}$, эти шаги нужно пропустить). Далее мы создаём частотную матрицу FM, которая имеет размерность $|A| \times M$. Номера её строк соответствуют кодам символов алфавита A, а номера столбцов — символам S_a. Величина элемента $FM[i][j]$ равна количеству экземпляров i-той аминокислоты в j-том столбце MSA'.

Элементы FM распределены по биномиальному закону, где количество испытаний Бернулли: $N = \sum_{i=1}^{|A|}\sum_{j=1}^{M} FM[i][j]$, а вероятность успеха: $p[i][j] = \frac{x[i] * y[j]}{N^2}$, $x[i] = \sum_{j=1}^{M} FM[i][j]$, $y[j] = \sum_{i=1}^{|A|} FM[i][j]$ (под успехом подразумевается вероятность того что символ с кодом i попадёт в j-тый столбец MSA'). Известно, что $Bin(n,p) \approx N\left(np, np(1-p)\right)$, где Bin — биномиальное распределение с вышеописанными параметрами, а N — нормальное распределение с математическим ожиданием np и дисперсией $np(1-p)$. Матрицу FM', элементы которой являются аргументами стандартного распределения $N(0,1)$, мы получаем путём следующих преобразований FM.

$$FM'[i][j] = \frac{FM[i][j] - np[i][j]}{\sqrt{np[i][j]\left(1 - p[i][j]\right)}} \quad (3)$$

Полученную на данном этапе матрицу FM' нельзя рассматривать как PWM, так как PWM используется в качестве матрицы замен при построении парных выравниваний. При фиксированных значениях параметров штрафа d и e, масштабирование значений в ячейках PWM приведёт к построению разных выравниваний и получению разных значений веса (меры подобия) F. Более того, даже если, масштабировать вместе с элементами PWM и параметры штрафа за разрыв, то, хотя выравнивания и будут одинаковыми, значения F будут пропорционально изменяться. Для обеспечения сравнимости этих значений мы нормализуем PWM, накладывая следующие ограничения:

$$R^2 = \sum_{i=1}^{|A|}\sum_{j=1}^{M} PWM[i][j]^2 \quad (4)$$

$$K_d = \sum_{i=1}^{|A|}\sum_{j=1}^{M} PWM[i][j] * p[i][j] \quad (5)$$

Здесь $p[i][j] = p_s[i] * p_a[j]$; $p_s[i]$ это вероятность появления символа i в S, $p_a[j]$ — вероятность появления j в SA (в данном случае $\forall j : t[j] = 1 / M$). K_d это фиксированный параметр, определяющий желаемое математическое элементов PWM. R^2 — квадрат длины радиус-вектора PWM, если рассматривать её как точку в пространстве размерности $|A| * M$. R^2 ограничивает порядок величин элементов PWM. Мы не можем использовать R^2 как фиксированный параметр, так как M не яв-

ляется константой даже при фиксированных S. Вместо этого в качестве параметра вводится масштабирующий множитель R_m: $R^2 = R_m|A|M$. Способ преобразования FM для удовлетворения условий (4) и (5) представлен в статье [15]. Все PWM, использующиеся в алгоритме MAHDS, должны удовлетворять условиям (4) и (5) с заданными параметрами K_d и R_m.

Оптимизация PWM

Для того чтобы найти такую PWM, которая для данных S позволит построить MSA с наибольшим значением функции сходства F при фиксированном M мы применяем итеративную процедуру. Сначала мы генерируем множество Q случайных максимально удалённых друг от друга (по Евклидовому расстоянию) матриц размерности $|A| \times M$. По умолчанию $|Q| = 400$. Это стартовые точки для дальнейшей оптимизации. Матрицы вошедшие в Q нормализуются, чтобы удовлетворять условиям (4) и (5), после чего интерпретируются как PWM.

Далее для каждой PWM из множества Q применяется процедура f: строятся парные выравнивания и оценивается F. По полученным парным выравниваниям восстанавливается PWM (процедура f^{-1}), которая не является идентичной предыдущей (она в большей степени подстроена под множество S). Такие преобразования выполняются итеративно до тех пор, пока значение F возрастает (то есть выполняется оптимизация PWM, при которой целевой функцией является функция сходства). Среди оптимизированных матриц из множества Q выбирается PWM с наибольшим значением F. Эта PWM с помощью процедуры f позволяет построить наилучшее MSA при данном фиксированном значении M.

Оценка статистической значимости

При построении MSA изначально неизвестно, какая длина S_a (значение M) является оптимальной. Величина F зависит от количества столбцов выравнивания и поэтому не может использоваться для корректного сравнения качества MSA, полученных при разных значениях M. В таких случаях показательной является статистическая значимость Z, которая является мерой неслучайности выравнивания [16].

Для оценивания Z выравнивания, полученного с помощью заданной PWM, используется метод Монте-Карло. В ходе оценивания генерируется R_n (по умолчанию $R_n = 100$) множеств случайных последовательностей. Они получаются путём перемешивания символов в рамках отдельных последовательностей из S. Для каждого из сгенерированных множеств подсчитывается F по формулам (1) и (2) с использованием заданной PWM. Для подсчёта Z используется формула $Z = \frac{F - \tilde{m}}{\tilde{\sigma}}$, где $\tilde{m}$ — оценка математического ожидания F, $\tilde{\sigma}$ — оценка среднеквадратичного отклонения F. Значение Z, большее, чем пороговое Z_t, свидетельствует о том, что выравнивание, полученное с помощью данной PWM, является статистически значимым. Z_t определяется по правилу 3 сигма при выравнивании случайных последовательностей [12].

Для выбора наиболее подходящего M мы перебираем варианты близкие к средней длине последовательностей S. Множество предполагаемых значений M обозначим M_s. Оно может быть задано, например, так: $M_s.= \{M|\left(\bar{l} - \Delta \le M \le \bar{l} + \Delta\right)\&\left(M \bmod s = 0\right)\}$, где $\bar{l}$ — средняя длина последовательности в S, Δ — параметр, определяющий нижнюю и верхнюю грани, s — шаг. Для каждого значения M из M_{set} при помощи итеративной процедуры находится PWM, обеспечивающая наибольшее F и оценивается Z. При помощи PWM, обеспечившей наибольшее Z, строится итоговое MSA.

Кроме непосредственно построения MSA, наш подход позволяет также оценивать статистическую значимость произвольных MSA вне зависимости от метода, которым они были получены. Для этого достаточно выполнить процедуру f^{-1} и оценить значение Z для PWM данного MSA.

Модификации MAHDS

В реальных биологических последовательностях скоррелированные мутации сразу в нескольких соседних позициях происходят чаще чем одиночные мутации. Мы разработали модификацию MAHDS, в которой учитываются корреляции символов, что позволяет находить более корректные MSA в некоторых случаях [11]. В этой вариации MAHDS вместо одиночных символов мы работаем с упорядоченными парами символов, и PWM в таком случае имеет размерность $|A|^2 \times (M - 1)$. Все шаги алгоритма при этом остаются прежними. Количество элементов PWM при большом значении $|A|$ существенно возрастает, что делает итеративную процедуру оптимизации PWM не эффективной из-за разреженности частотных матриц FM. Особенно это актуально для аминокислотных последовательностей, для которых $|A| = 20$. Решением может быть сужение исходного алфавита до 5 символов, обозначающих химические свойства аминокислот: неполярные, полярные незаряженные. ароматические, заряженные отрицательно, заряженные положительно. При таком подходе теряется часть информации о последовательностях, однако появляется возможность учитывать корреляции.

Также можно модифицировать процедуру оценки Z по методу Монте-Карло. При выравнивании множеств случайных последовательностей с использованием за-

данной PWM может применяться итеративная процедура, подстраивающая PWM под эти последовательности. Это увеличивает объём вычислений, однако позволяет получать более стабильные и воспроизводимые результаты. Величина Z_t в таком случае становится меньше. Оценку Z с итеративной оптимизацией PWM мы ранее применяли в [12,13] и применяем в данной работе.

Результаты и обсуждение

Мы сгенерировали 100 множеств по 100 случайных последовательностей. Множества состоят из 40 последовательностей по 60 символов, 10 по 90, 10 по 110, 40 по 140 ($\bar{l} = 100$). Символы появляются в последовательностях с вероятностями характерными для белков живых организмов в среднем [17]. Мы построили выравнивание каждого из множеств и оценили статистическую значимость этих выравниваний, чтобы определить пороговое значение Z_t, корректное при выравнивании аминокислотных последовательностей с учётом корреляций. Гистограмма плотности распределения величины Z для MSA модельных последовательностей представлена ниже (см. рис. 1).

Полученное эмпирическое распределение близко к $N(1.19, 2.76)$. По критерию согласия Колмогорова-Смирнова величина $p = 0.47$ (критерий имеет правостороннюю критическую область). Также по рисунку 1 видно, что ни для одного множества последовательностей оцененная статистическая значимость MSA не превысила $m + 3\sigma$, значит мы действительно можем применять Z_t рассчитанную по правилу 3 сигма: $Z_t = 6.17$.

Если при оценке статистической значимости PWM по методу Монте-Карло выравниваний не оптимизировать PWM под множества случайным образом перемешанных последовательностей, получится гистограмма следующего вида.

При сравнении данного распределения с $N(29.03, 24.3)$ по критерию согласия Колмогорова-Смирнова величина $p = 0.39$, что свидетельствует о меньшей схожести данного распределения с нормальным, чем на рисунке 1. При этом для обобщённого распределения экстремальных значений величина $p = 0.65$. Все значения также как и на предыдущем рисунке не превышают $m + 3\sigma$, поэтому можно считать $Z_t = 43.82$.

Для выравнивания одного случайно взятого множества модельных последовательностей мы построили график изменения значений F в ходе оптимизации итеративной процедурой каждой из 400 инициализирующих матриц.

Как видно на рисунке 3, наибольшие значения F достигаются не менее чем через 10 итераций. При этом итеративная процедура достигает локального максимума в среднем за 7.91 итерацию. Поэтому величина $|Q|$ действительно должна быть достаточно большой (не менее 400), чтобы среди инициализирующих матриц нашлись те, которые могут достичь высоких значений F. В среднем разница между F на первой и последней итерациях составляет 84.84. Сильное возрастание F на второй итерации, вероятнее всего, связано с тем, что после первой итерации PWM создаётся из частотной матрицы

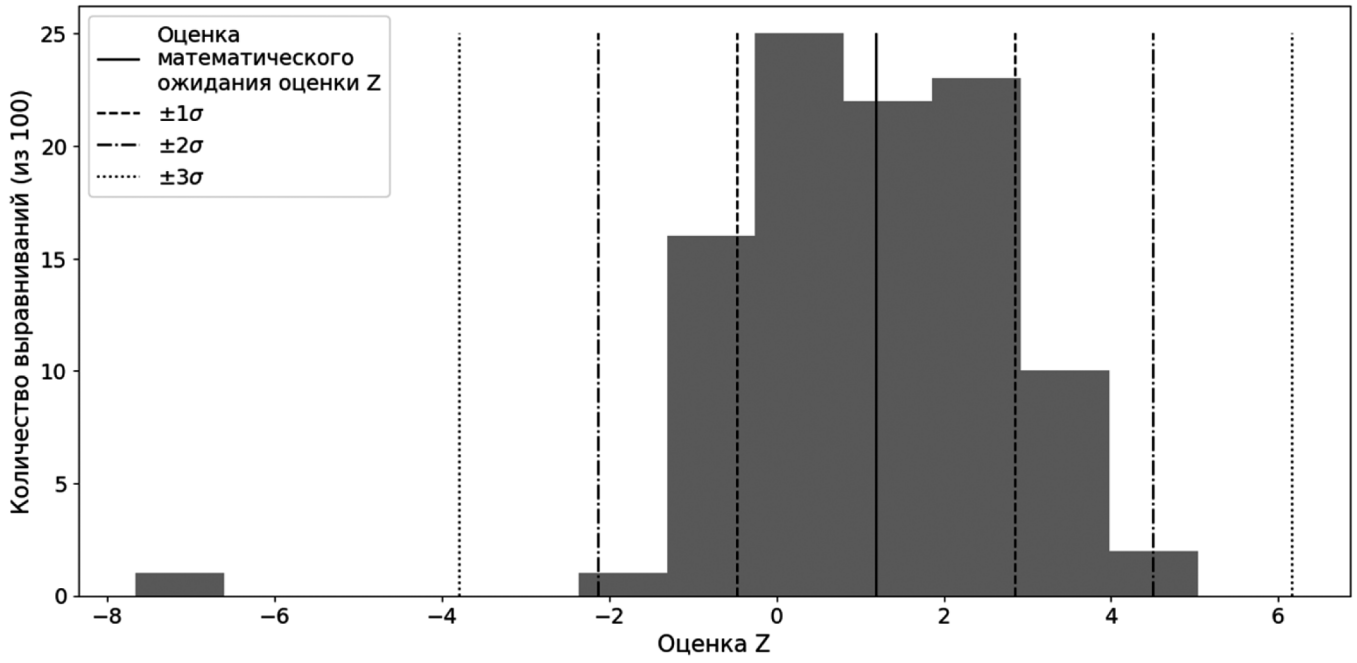

Рис. 1. Эмпирическая плотность распределения величины Z для MSA случайных аминокислотных последовательностей

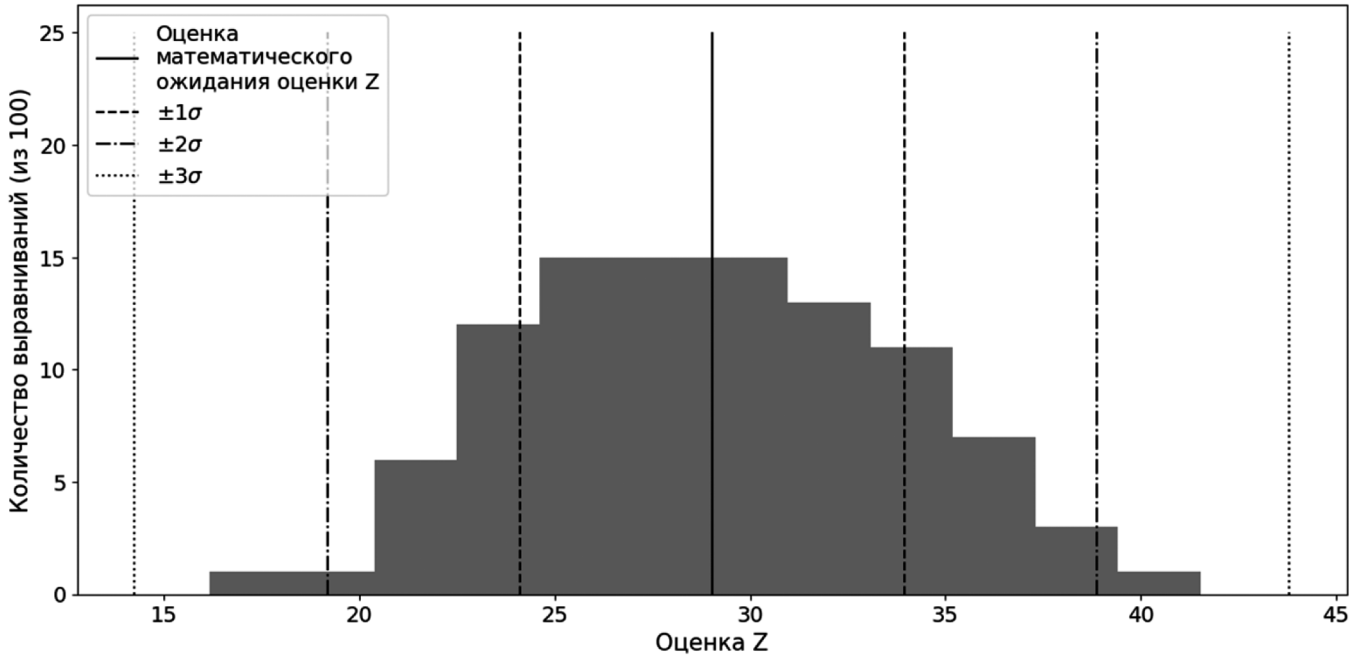

Рис. 2. Эмпирическая плотность распределения величины Z для MSA случайных аминокислотных последовательностей (без оптимизации PWM при оценивании Z)

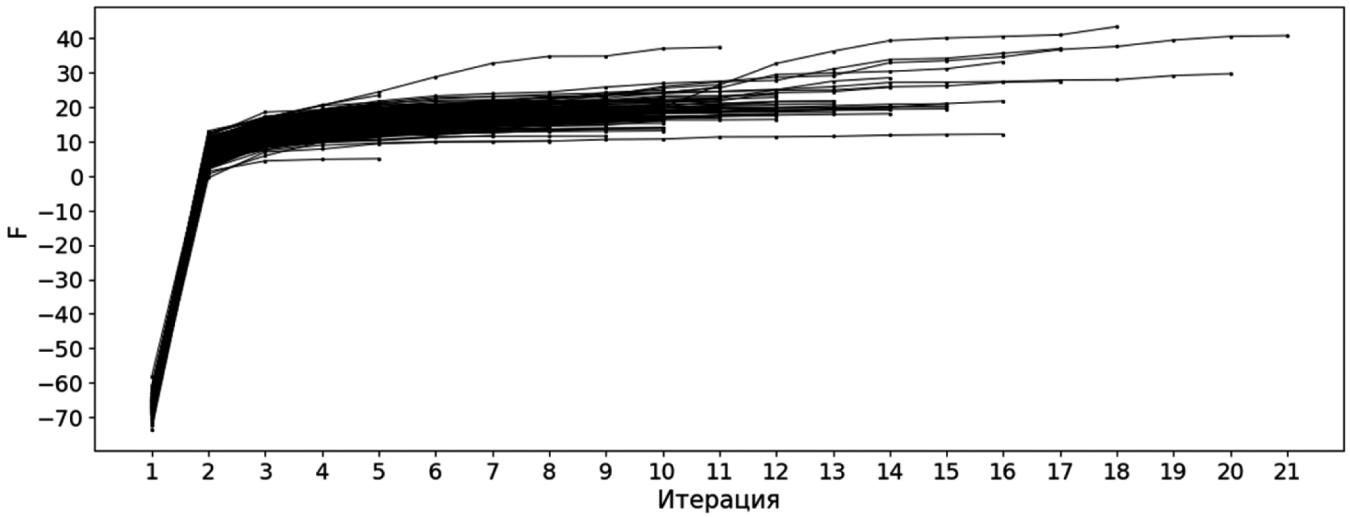

Рис. 3. Зависимость значений функции подобия F от номера итерации при оптимизации каждой из 400 случайных матриц

выравнивания, а не из нормализованной случайной матрицы. Также мы построили гистограммы распределения F для матриц множества Q до оптимизации и после.

Изначально близкое к $N(-66.34, 4.45)$ распределение (по критерию согласия Колмогорова-Смирнова значение $p=0.47$) после оптимизации преображается в иное распределение, не являющееся нормальным или обобщённым распределением экстремальных значений (значение $p \leq 0.0001$). Наличие на гистограмме справа значений F, существенно превышающих математическое ожидание, свидетельствует о том, что некоторые PWM были хорошо оптимизированы под выравниваемые последовательности и их применение позволит построить значимое MSA.

Для выравнивания того же самого множества последовательностей мы построили график изменения значений F в ходе подстройки итеративной процедурой PWM под каждое из 100 множеств последовательностей со случайно перемешанными символами при оценивании статистической значимости выравнивания по методу Монте-Карло.

В данном случае итеративная процедура достигает локального максимума в среднем за 7.39 итераций. При этом средняя разница между величинами F на пер-

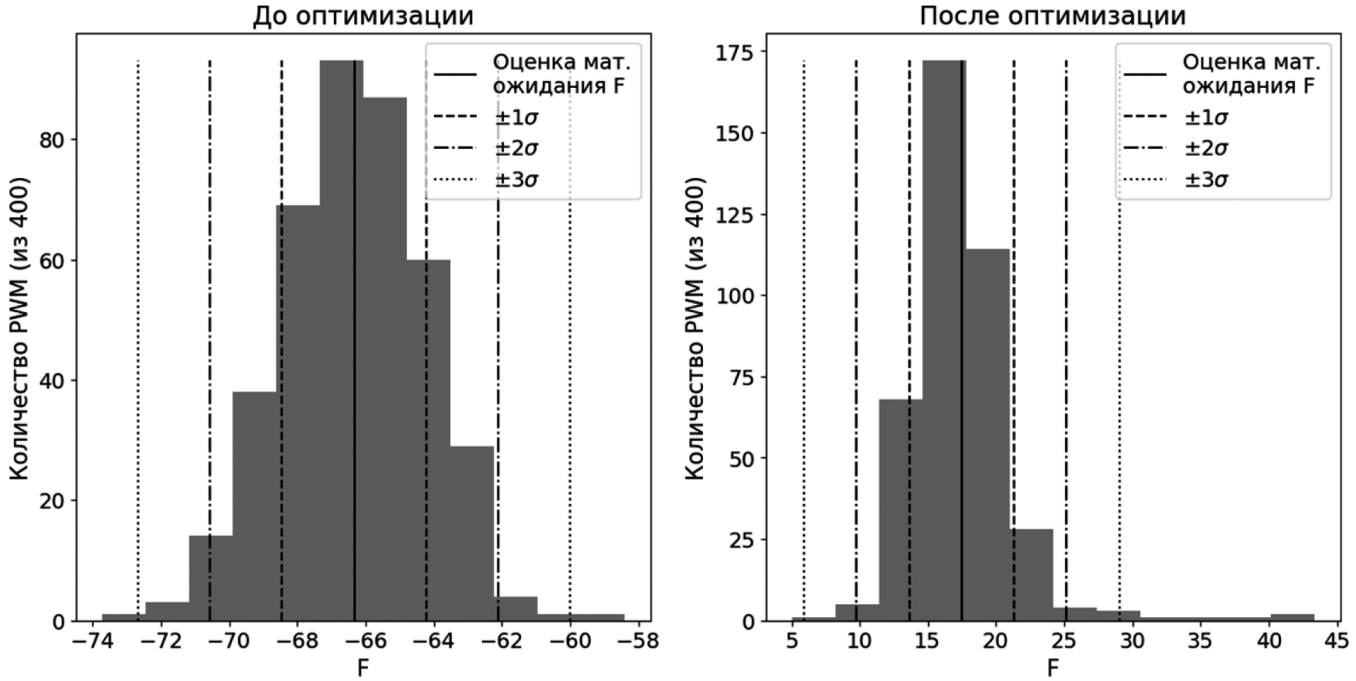

Рис. 4. Гистограммы распределения значений *F* до и после применения итеративной процедуры

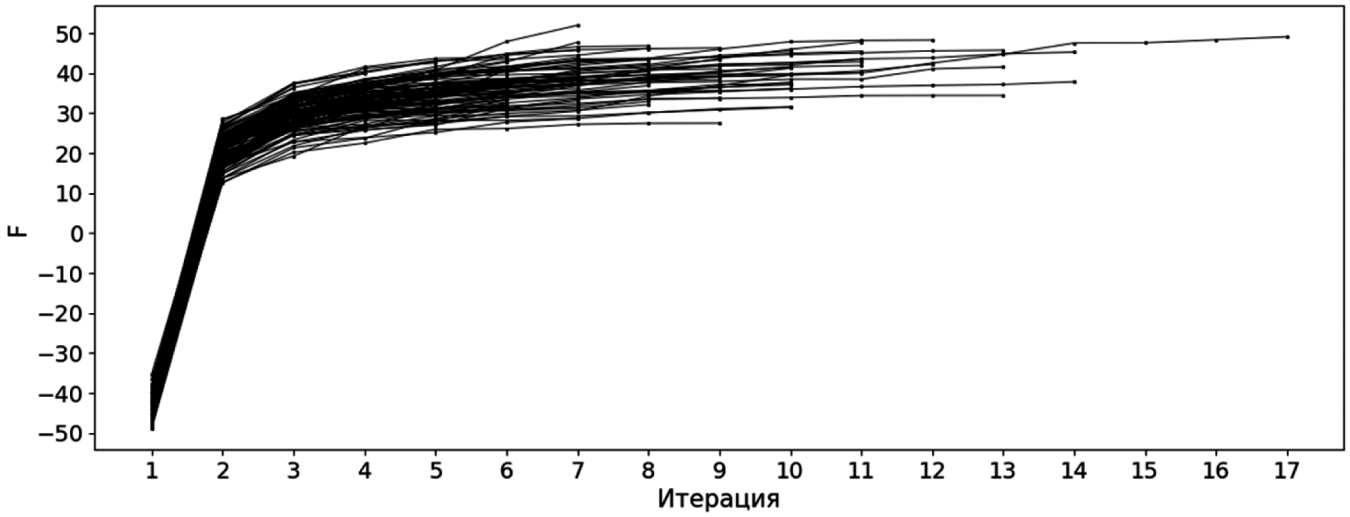

Рис. 5. Зависимость значений функции подобия *F* от номера итерации при оптимизации *PWM* под перемешанные последовательности в ходе оценки *Z*

вой и последней итерациях составляет 70.42, что существенно меньше, чем в ходе поиска оптимальной PWM для выравнивания *S*. Однако на первой итерации на рисунке 5 в сравнении с рисунком 3 величины *F* в среднем больше, так как при оценке статистической значимости изначально PWM хоть и не адаптирована под случайные последовательности, но по крайней мере отражает их аминокислотный состав.

Также мы построили гистограммы распределения *F* для выравнивания случайно перемешанных последовательностей с помощью заданной PWM до её оптимизации и после.

В данном случае, в отличие от рисунка 4, вид распределения не меняется существенно. До оптимизации эмпирическое распределение близко к $N(-42.49, 8.29)$ (по критерию согласия Колмогорова-Смирнова значение $p = 0.54$), после оптимизации — к $N(38.03, 29.48)$ (по критерию согласия Колмогорова-Смирнова значение $p = 0.79$). Для корректной оценки *Z* необходимо, чтобы *F* было распределено по нормальному закону. Подстройка PWM позволяет получить распределение более близкое к нормальному, поэтому применение итерационной процедуры на этапе оценки *Z* желательно.

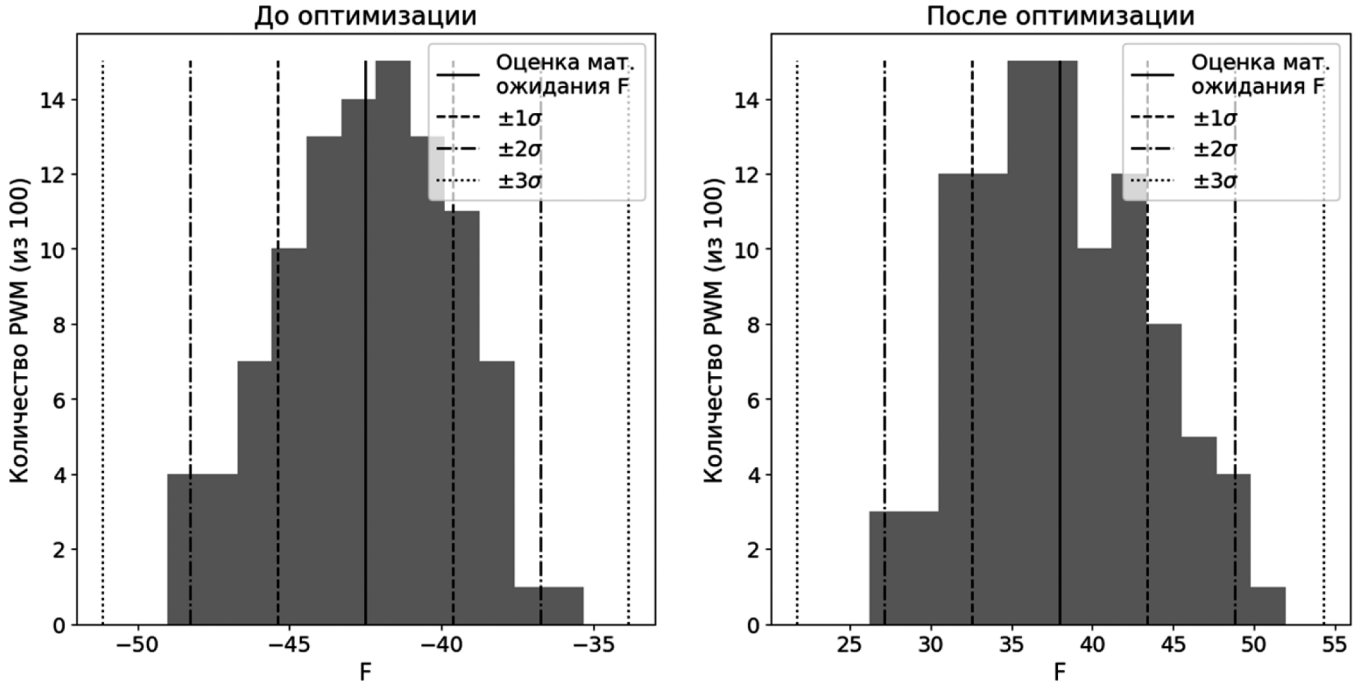

Рис. 6. Гистограммы распределения значений *F* до и после применения итеративной процедуры в ходе оценивания статистической значимости *PWM*

Заключение

Были исследованы свойства MAHDS, проявляющиеся при построении и оценивании выравниваний аминокислотных последовательностей с учётом корреляций соседних символов. В этом контексте для оптимизации *PWM* итеративной процедурой необходимо использовать упрощённый алфавит (вместо 20 аминокислот символы соответствуют 5 классам аминокислот по их химическим свойствам). Были рассчитаны пороговые величины значимости. Если при оценке значимости по методу Монте-Карло *PWM* не оптимизируется, *MSA* может считаться отличным от случайного при $Z > 43.82$. Если *PWM* с помощью итеративной процедуры подстраивается под множества последовательностей со случайно перемешанными символами при оценке значимости, отличное от случайного выравнивание должно иметь $Z > 6.17$. Точность и воспроизводимость оценки во втором случае лучше, однако такой подход требует затраты дополнительных вычислительных ресурсов.

ЛИТЕРАТУРА

1. Chatzou M. et al. Multiple sequence alignment modeling: methods and applications // Briefings in Bioinformatics. 2016. Vol. 17, № 6. P. 1009–1023.
2. Wang L., Jiang T. On the complexity of multiple sequence alignment.: 4 // J Comput Biol. United States, 1994. Vol. 1, № 4. P. 337–348.
3. Gotoh O. Heuristic Alignment Methods // Multiple Sequence Alignment Methods / ed. Russell D.J. Totowa, NJ: Humana Press, 2014. P. 29–43.
4. Chowdhury B., Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm // Genomics. 2017. Vol. 109, № 5. P. 419–431.
5. Feng D.F., Doolittle R.F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees // J Mol Evol. 1987. Vol. 25, № 4. P. 351–360.
6. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Res. 2004. Vol. 32, № 5. P. 1792–1797.
7. Notredame C., Higgins D.G., Heringa J. T-coffee: A novel method for fast and accurate multiple sequence alignment: 1 // Journal of Molecular Biology. 2000. Vol. 302, № 1. P. 205–217.
8. Higgins D.G., Sharp P.M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer // Gene. 1988. Vol. 73, № 1. P. 237–244.
9. Lassmann T., Sonnhammer E.L. Kalign — an accurate and fast multiple sequence alignment algorithm // BMC Bioinformatics. 2005. Vol. 6, № 1. P. 298.
10. Katoh K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform // Nucleic Acids Res. 2002. Vol. 30, № 14. P. 3059–3066.
11. Korotkov E.V. et al. Multiple Alignment of Promoter Sequences from the Arabidopsis thaliana L. Genome: 2 // Genes. 2021. Vol. 12, № 2. P. 135.
12. Korotkov E.V., Kostenko D.O. Application of the MAHDS Method for Multiple Alignment of Highly Diverged Amino Acid Sequences: 7 // International Journal of Molecular Sciences. 2022. Vol. 23, № 7.
13. Kostenko D., Korotkova M., Korotkov E. Multiple Alignments of Protein Families with Weak Sequence Similarity Within the Family // Symmetry. 2025. Vol. 17, № 3.
14. Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins: 3 // Journal of Molecular Biology. Academic Press, 1970. Vol. 48, № 3. P. 443–453.
15. Pugacheva V., Korotkov A., Korotkov E. Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming: 5 // Statistical Applications in Genetics and Molecular Biology. 2016. Vol. 15, № 5.
16. Comet J.P. et al. Significance of Z-value statistics of Smith-Waterman scores for protein alignments: 3–4 // Comput Chem. 1999. Vol. 23, № 3–4. P. 317–331.
17. Kozlowski L.P. Proteome-pI: proteome isoelectric point database // Nucleic Acids Research. 2016. Vol. 45, № D1. P. D1112–D1116.

© Костенко Дмитрий Олегович (dk0stenko@yandex.ru); Коротков Евгений Вадимович (bioinf@yandex.ru);
Короткова Мария Александровна (discretmath@gmail.com)

Журнал «Современная наука: актуальные проблемы теории и практики»