БИОФИЗИКА, 2015, том 60, вып. 6, с. 1057–1068

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.212.2

РАЗРАБОТКА МАТЕМАТИЧЕСКОГО МЕТОДА ДЛЯ ПОИСКА СКРЫТОЙ ПЕРИОДИЧНОСТИ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ БЕЛКОВ С УЧЕТОМ ДЕЛЕЦИЙ И ВСТАВОК

© 2015 г. Е.В. Коротков* **, М.А. Короткова**

**Федеральный Центр «Фундаментальные основы биотехнологии» РАН, 117312, Москва, просп. 60-тия Октября, 7/1;*

***Национальный исследовательский ядерный университет (МИФИ), 105409, Москва, Каширское шоссе, 31*

E-mal: bioinf@yandex.ru

Поступила в редакцию 30.06.15 г.

Разработан математический метод для поиска скрытой периодичности в аминокислотных последовательностях белков и в других символьных последовательностях с использованием динамического программирования и случайных периодических матриц. Метод позволяет находить скрытую периодичность со вставками и делециями символов в заранее неизвестных местах. Разработанный метод был применен для поиска периодичности в аминокислотных последовательностях некоторых белков и в обменном курсе евро/доллар, начиная с 2001 года. Показано присутствие длинной периодичности со вставками и делециями аминокислот с длиной периода, равной семи аминокислотам в белках, содержащих суперспиральные области, а также наличие периодичности длиной в шесть и пять аминокислот и более длинными периодами. Показано также существование периодичности, равной шести и семи суткам, а также 24 и 25 часам в проанализированных финансовых рядах, которую можно обнаружить только со вставками и делециями. Обсуждаются причины возникновения скрытой периодичности со вставками и делециями символов в аминокислотных последовательностях белков и в финансовых рядах.

Ключевые слова: скрытая периодичность, динамическое программирование, евро, доллар.

В условиях больших успехов по секвенированию разнообразных геномов и нарастающего накопления информации о полных геномах многих видов особое значение приобретают разработка и применение математических методов для изучения аминокислотных и нуклеотидных последовательностей [1]. Без этих методов бóльшая часть известных последовательностей оснований ДНК хранится в компьютерных банках данных без существенного использования. Особенно это касается геномов эукариот. Одной из задач при развитии новых математических методов является нахождение новых математических закономерностей организации нуклеотидных и аминокислотных последовательностей и выяснение связи этих закономерностей с известными биологическими функциями. Таким образом, разработка новых математических методов позволяет сделать структурную аннотацию тех участков геномов и аминокислотных последовательностей, которые в настоящее время никак не охарактеризованы. Эти же исследования позволяют затем связать те или иные структурные закономерности строения последовательностей с их биологическими свойствами. В результате таких исследований развиваются новые методы структурной и функциональной аннотации нуклеотидных и аминокислотных последовательностей.

Одной из структурных закономерностей последовательностей, широко представленных в аминокислотных ДНК, является периодичность, как явная, так и скрытая [2,3]. Данная работа направлена на развитие нового математического подхода к поиску скрытой периодичности аминокислотных последовательностей, который позволит провести обширную аннотацию не охарактеризованных последовательностей различных белков. Под скрытой периодичностью понимается такая периодичность, где было накоплено достаточно большое количество замен аминокислот. Число таких замен зависит от числа периодов в изучаемой последовательности и может колебаться от одной до двух замен на аминокислоту [31].

В математических подходах, развитых в настоящее время для поиска периодичностей в символьных и числовых последовательностях, имеется существенный пробел. Спектральные подходы позволяют находить достаточно «размытую» периодичность последовательности без вставок или же делеций. К числу спектральных методов можно отнести преобразование Фурье, вейвлет-преобразование, информационное разложение и некоторые другие методы [2,5-10]. Однако эти подходы имеют существенное ограничение – они не позволяют обнаруживать периодичность со вставками и делециями символов. С другой стороны, методы, основанные на динамическом программировании, позволяют относительно точно находить вставки или же делеции символов. Однако эти методы не могут обнаруживать значительно «размытую» или скрытую периодичность, где статистическая значимость подобия между любыми двумя периодами невелика. Это связано с тем, что методы динамического программирования используют фиксированную весовую матрицу, от вида которой зависит вид выравнивания [11,12]. Поэтому в данной работе мы хотим разработать математический метод, позволяющий учесть данный пробел и находить скрытую периодичность аминокислотной последовательности в условиях присутствия вставок или же делеций аминокислот (в заранее неизвестных позициях) и при отсутствии заданной матрицы весов.

Любую периодичность последовательности S длиной N можно характеризовать либо частотной [3], либо созданной по ней позиционно-весовой матрицей M [13]. Признаки строк такой матрицы – аминокислоты, а признаками столбцов являются позиции периода. Элементы этой матрицы $m(i,j)$ содержат вес, который имеет аминокислота i в позиции j периода. Позиции периода меняются от 1 до n. Введем также последовательность Seq_1 длиной N, которая будет представлять собой искусственную периодическую последовательность $1,2,\ldots,n,1,2,\ldots,n,\ldots$ Здесь числа будут рассматриваться как символы и им будут соответствовать столбцы j в матрице M. Любой периодичности с периодом, равным n последовательности S, будет соответствовать своя весовая матрица $M(20,n)$.

Задача формулируется следующим образом. Нам дана последовательность S длиной в N символов. Необходимо найти такую оптимальную весовую матрицу M_0, где локальное выравнивание последовательностей S и Seq_1 будет иметь наибольшую статистическую значимость. Под статистической значимостью понимается вероятность P того, что $F_r > F_{\max}$, где $F_{\max}$ – максимальный вес локального выравнивания последовательностей S и Seq_1 с использованием матрицы M_0. Это означает, что ищется такая матрица M_0, которая будет давать наименьшую вероятность P. Здесь F_r – максимальный вес локального выравнивания случайно перемешанной последовательности S и последовательности Seq_1 с использованием найденной для них оптимальной матрицы M_r. Всегда можно задать пороговый уровень вероятности P_0, и если вероятность $P(F_r > F_{\max})$ будет меньше P_0, то найденное локальное выравнивание последовательностей S и Seq_1 с использованием весовой матрицы M_0 можно считать статистически значимым.

Процедура локального выравнивания аминокислотной последовательности S и искусственной периодической последовательности Seq_1 относительно известной весовой матрицы была предложена ранее [12]. Остается каким-либо способом найти оптимальную весовую матрицу M_0. Данная работа направлена на разработку математического подхода для поиска матрицы M_0, а также для метода оценки вероятности P. Для поиска оптимальной весовой матрицы используется оптимизационный алгоритм и динамическое программирование. Для оценки вероятности P используется метод Монте-Карло. Целью работы было также показать на нескольких примерах, что в аминокислотных последовательностях существует скрытая периодичность со вставками и делециями аминокислот. Мы не ставили себе задачу проанализировать все известные аминокислотные последовательности, так как разработанный метод требует достаточно больших компьютерных ресурсов. Разработанный алгоритм применен для поиска скрытой периодичности со вставками и делециями в аминокислотных последовательностях небольшого количества белков. Показано присутствие скрытой периодичности со вставками и делециями в аминокислотных последовательностях таких белков, для которых присутствие скрытой периодичности ранее не было известно. Одновременно с этим удалось показать, что разработанный подход применим к числовым последовательностям, если их перекодировать в символы (20 символов). Мы показали это на примере курса евро/доллар, где нам удалось обнаружить периодичность со вставками и делециями длиной в семь суток и в 24 и 25 часов.

МЕТОДЫ И АЛГОРИТМЫ

Для поиска периодичности со вставками и делециями аминокислот мы использовали ал-

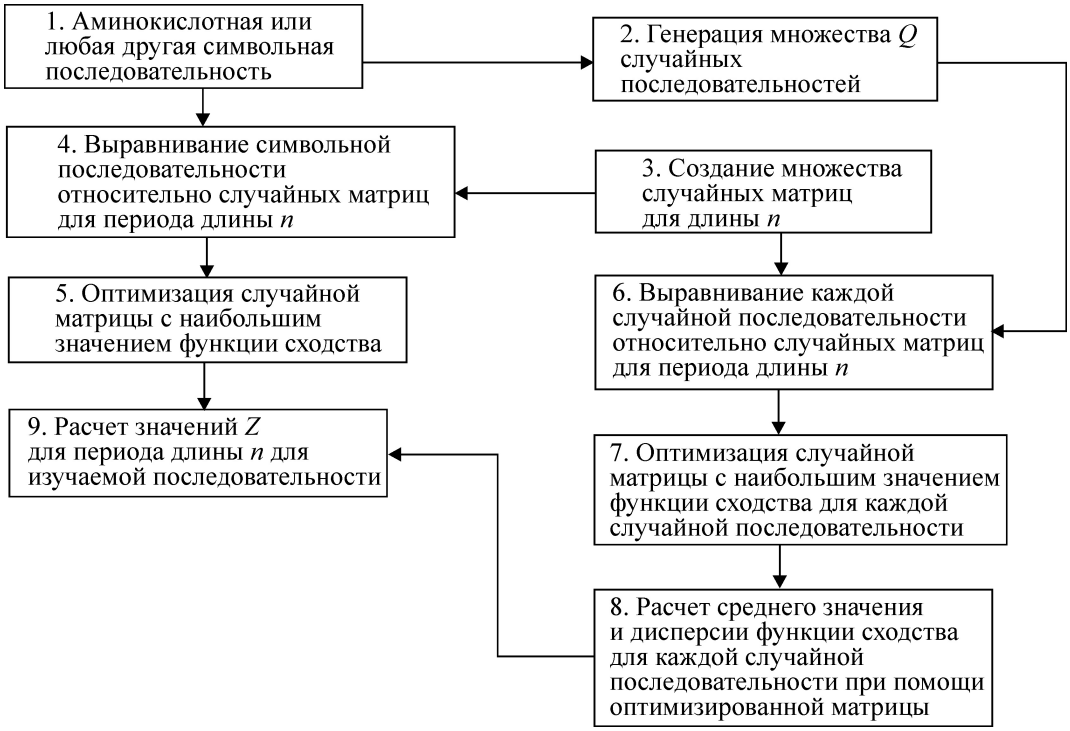

Рис. 1. Схема математического алгоритма, применяемого для расчета статистической значимости Z периода длиной n в изучаемой последовательности.

горитм, представленный на рис. 1. По этому алгоритму вначале мы генерируем множество случайных матриц (рис. 1, пункт 3) размером $20\times n$, где n есть длина периода, а 20 представляет собой размер алфавита изучаемой аминокислотной последовательности. Затем эти матрицы используются для построения локального выравнивания изучаемой последовательности относительно каждой из созданных случайных матриц (рис. 1, пункт 4). Для построения выравнивания и определения функции сходства F используется динамическое программирование. Матрицы трансформировались таким образом, чтобы распределение функции сходства F для всех случайных последовательностей было подобным для всех полученных случайных матриц. Выбиралась такая случайная матрица, которая имела наибольшее значение функции сходства F со случайными последовательностями. Затем эта случайная матрица оптимизировалась с целью достижения наибольшего значения функции сходства $\max F$ (рис. 1, пункт 5). После этого создавалось множество случайных последовательностей (рис. 1, пункт 2). Для каждой последовательности из этого множества рассчитывалось значение $\max F$, что позволило определить для $\max F$ среднее значение и дисперсию (рис 1, пункты 6, 7 и 8). Алгоритм был применен для периодов n длиной от 2 до 100, и для каждой длины периода мы рассчитывали соответствующее значение Z. В результате работы алгоритма мы получаем зависимость Z от n, которую обозначим как $Z(n)$.

Следует отметить, что в данном исследовании мы использовали метод динамического программирования, позволяющий находить локальное выравнивание. Это означает, что границы района, где получается $\max F$, могут отличаться от начала и конца изучаемой последовательности. Это означает, что значения $Z(n)$ для различных n могут быть получены для различных фрагментов изучаемой последовательности. Мы отдельно будем приводить границы фрагментов, для которых были получены значимые значения $Z(n)$. Ниже рассмотрим более детально каждый из этапов алгоритма, показанного на рис. 1.

Генерация множества случайных последовательностей. Было создано множество Q случайных последовательностей путем случайного перемешивания последовательности S (рис. 1, пункт 2). Объем множества Q составлял 200 последовательностей. Для получения одной случайной аминокислотной последовательности мы генерировали случайную числовую последовательность длины N датчиком случайных чисел. После этого мы упорядочивали случайную числовую последовательность по возрас-

танию с запоминанием созданных перестановок. Затем произведенные перестановки применялись для перемешивания последовательности S, и в результате такого перемешивания создавалась случайная аминокислотная последовательность из множества Q.

Создание множества случайных матриц для длины *n*. Мы использовали случайные матрицы, которые имеют размерность 20×n, где n есть длина периода (рис. 1, пункт 3). Каждую матрицу можно рассматривать как точку в пространстве 20×n. Нам нужно было создать множество случайных матриц W, расстояние между которыми в пространстве 20×n было не меньше определенного значения. Для расчета различий между двумя матрицами $m_1(i,j)$ и $m_2(i,j)$ мы использовали информационную меру [14]:

$$I_j(M_1,M_2) = \sum_{i=1}^{20} m_1(i,j)\ln(m_1(i,j)) + \sum_{i=1}^{20} m_2(i,j)\ln(m_2(i,j)) - \sum_{i=1}^{20}(m_1(i,j)+m_2(i,j))\ln(m_1(i,j)+m_2(i,j)) + (s_1(j)+s_2(j))\ln(s_1(j)+s_2(j)) - s_1(j)\ln(s_1(j)) - s_2(j)\ln(s_2(j)), \quad (1)$$

где $s_k(j) = \sum_{i=1}^{20} m_k(i,j)$. $2I_j$ имеет асимптотическое $\chi^2(df)$-распределение с df = 19 [20]. Затем мы рассчитывали:

$$I(M_1,M_2) = \sum_{j=1}^{n} I_j(M_1,M_2). \quad (2)$$

Здесь $2I(M_1,M_2)$ имеет асимптотическое $\chi^2(df)$-распределение и df равно 19n, так как $I_1(M_1,M_2)$, $I_2(M_1,M_2)$,…, $I_{n-1}(M_1,M_2)$ независимы друг от друга и $I_n(M_1,M_2)$ полностью определено [14]. Затем мы использовали аппроксимацию для нормального распределения:

$$x(M_1,M_2) = \sqrt{4I(M_1,M_2)} - \sqrt{2df - 1}. \quad (3)$$

Мы получили величину $x(M_1,M_2) \sim N(0,1)$, где $N(0,1)$ есть аргумент стандартного нормального распределения. $N(0,1)$ очень удобно использовать как меру различия между матрицами $m_1(i,j)$ и $m_2(i,j)$. Вероятность $p = P(x > x(M_1,M_2))$ показывает вероятность того, что различия между матрицами $m_1(i,j)$ и $m_2(i,j)$ обусловлены случайными факторами. Чем больше $N(0,1)$, тем больше различие между матрицами $m_1(i,j)$ и $m_2(i,j)$. Мы выбрали различие $x(M_1,M_2)$ между матрицами не меньше чем 1,0. Алгоритм генерации матриц был следующий. Каждый элемент матрицы $m(i,j)$, i = 1,…,20, j = 1,…,n случайно заполнялся с равной вероятностью либо 0, либо 1. Затем матрица сравнивалась со всеми матрицами, которые уже вошли во множество W. Если хотя бы с одной матрицей это различие было меньше, чем L = 1,0, то созданная матрица не включалась в множество W. Если же для всех матриц из множества W различие было бóльшим, чем L = 1,0, то матрица включалась во множество W. Всего было создано 10^8 таких матриц для каждой длины периода n. Эти матрицы были использованы для построения выравниваний созданных последовательностей относительно матриц из множества W.

Выравнивание аминокислотной последовательности относительно случайных матриц для периода длины *n*. Для поиска периодичности в последовательности S со вставками и делециями мы проводили выравнивание последовательности S относительно модифицированных матриц m' из множества W, способ модификации матриц описан ниже (рис. 1, пункт 4). Для построения выравнивания мы заполняли матрицу для функции сходства F с использованием модифицированной матрицы $m'(i,j)$:

$$F(i,j) = \max\begin{Bmatrix} 0 \\ F(i-1,j-1) + m'(s(i),k) \\ F(i,j-1) - d \\ F(i-1,j) - d \end{Bmatrix}, \quad (4)$$

где $s(i)$ – элемент последовательности S, d – цена за вставку или делецию аминокислоты в последовательности S. Здесь i и j менялись от 1 до n; $k = j - n*\mathrm{int}((j - 0{,}1)/n)$. Это означает, что индексу j всегда соответствует столбец матрицы с номером k. Матрица F имеет размерность N на N, где N – длина последовательности S.

Одновременно с построением функции сходства F заполнялась матрица обратных переходов F' такой же размерности, как и матрица F. Каждый элемент матрицы $F'(i,j)$ содержит номер элемента матрицы F, для которого достигается максимум в формуле (4). После заполнения матриц F и F' мы определяли максимальный элемент max$F$ в матрице F и его координаты (i_m, j_m). Затем мы по матрице обратных переходов F создавали выравнивание последовательности S относительно последовательности индексов k, как это было уже описано ранее [15]. Созданному выравниванию соответ-

ствует путь в матрице F от точки (i_m, j_m) до точки (i_0, j_0) с использованием матрицы обратных переходов. В точке (i_0, j_0) значения функции F первый раз обращаются в ноль. Эта точка является началом выравнивания.

Первой последовательностью выравнивания является последовательность номеров столбцов k, а последовательность S является второй последовательностью в выравнивании. Каждому столбцу k в выравнивании будет сопоставлена аминокислота последовательности S или же знак *, который показывает, что данному столбцу не сопоставлена ни одна аминокислота последовательности S. Аналогично, каждой аминокислоте последовательности S сопоставлен определенный столбец k или же знак *, который показывает, что данной аминокислоте не сопоставлен ни один столбец k.

Для построения выравниваний последовательностей, созданных как описано в предыдущих разделах, использовались модифицированные матрицы m' из множества W. Сначала для матриц m рассчитывалось значения A и B как:

$$A = \sum_{i=1}^{20} \sum_{j=1}^{n} m(i,j)^2, \quad (5)$$

$$B = \sum_{i=1}^{20} \sum_{j=1}^{n} m(i,j)p(i)f(j), \quad (6)$$

где $f(j) = 1/n$; $p(i) = n(i)/N$, $n(i)$ – количество аминокислот типа i в последовательности S, N – общее число аминокислот последовательности S. Для проведения выравнивания нам нужны матрицы m', которые удовлетворяют двум условиям. Первое условие состоит в том, что значения A для матриц m' с одной и той же длиной периода n были бы одинаковы и равны $200 \times n$. Второе условие состоит в том, чтобы функции распределения значений max$F$ были бы максимально похожими для всех матриц с n столбцами. Такое распределение для каждой матрицы можно построить, если проводить выравнивание каждой матрицы с n столбцами со случайными последовательностями из множества Q. Мы подбирали такое значение константы B для каждой матрицы, которое обеспечивало максимальную идентичность функции распределения max$F$ с одним и тем же значением n.

Приведенные выше два условия позволяют нам заменить матрицы m на матрицы m', которые этим условиям удовлетворяют. Уравнение (5) есть уравнение поверхности шара в пространстве $20 \times n$, а уравнение (6) – уравнение плоскости. Если матрица m' удовлетворяет этим условиям, то она лежит на окружности, образованной пересечением поверхности шара (уравнение 5) плоскостью (уравнение 6). Матрица m рассматривалась как точка в пространстве $20 \times n$, и от этой точки мы брали ближайшую к ней точку, которая лежит на окружности, образованной пересечением поверхности шара плоскостью. Координаты этой точки будут задавать искомую матрицу m'. Несложно написать несколько простых уравнений для определения координат этой точки (значения матрицы m') по координатам матрицы m с использованием уравнений (5), (6). Фактически это означает, что если мы задали константы A, B, матрицу m и посчитали $p(i)$ для последовательности S, то мы однозначно определяем матрицу m' (если есть пересечение поверхности шара плоскостью). Далее задача состояла в том, чтобы подобрать для каждой матрицы с n столбцами константу B, которая бы обеспечивала максимальную идентичность функции распределения max$F$ на множестве последовательностей Q.

Одновременно с расчетом функции распределения max$F$ на множестве последовательностей Q мы рассчитывали также среднюю длину случайного выравнивания $\overline{l}$ как разность (i_m – i_0), где i_m – координата max$F$ в последовательности S, а i_0 – координата, где F = 0,0 при построении выравнивания (координата начала выравнивания в последовательности S). Мы выбрали среднюю длину случайного выравнивания равной $N/5$ аминокислот. Это значение обеспечивает наилучшее определение границ выравнивания относительно реальных границ на модельных последовательностях.

Выбор константы B осуществлялся итеративно. Константа B, которая обеспечивает $\overline{l}$ около $N/5$ аминокислот, заведомо лежит в интервале от $B_1 = 0$ до $B_2 = -15{,}0$. У нас не было зарегистрировано матриц, которые не удовлетворяли бы этому требованию. Затем мы брали середину этого интервала. Если $\overline{l}$ была больше, чем $N/5$, то устанавливаем $B_1 = -5{,}0$, а если $\overline{l}$ была меньше, чем $N/5$, то устанавливаем $B_2 = -5{,}0$, и процесс определения $\overline{l}$ повторялся. При достижении значения $\overline{l} = N/5 \pm 60$ процесс выбора константы B останавливался.

Оптимизация случайной матрицы с наибольшим значением функции сходства. Для всех матриц из множества W определялась такая модифицированная матрица max(m'), которая имела наибольшее значение функции сходства F, а также строилось выравнивание и определялись координаты выравнивания (рис. 1,

пункт 5). Однако несмотря на то, что мы использовали предельно большое количество матриц, матрица max(m') может иметь не самое большое значение max$F$, которое возможно для последовательности S и для данной длины периода n. Это означает, что наибольшее значение может быть достигнуто для матрицы, которая лежит на некотором расстоянии от матрицы max(m'), меньшем, чем выбранный нами порог различия матриц L. Поэтому мы создали примерно 10^7 матриц, которые имели отличие от матрицы max(m') L меньшее, чем заданное в предыдущем разделе (L = 1,0), но всегда большее, чем 0,0. Это означает, что различие L с матрицей max(m') находились в интервале (0,0–1,0). Эти матрицы были также применены, как указано выше. В качестве финальной была выбрана матрица, которая имела наибольшее значение max$F$.

Расчет значения Z для периода длины n для изучаемой последовательности. Вышеописанные процедуры были применены для всех случайных последовательностей из множества Q. В результате для каждой длины n и аминокислотной последовательности S были определены среднее значение max$\overline{F}$ и дисперсия D(max$F$) для величины max$F$ после проведения оптимизации. Это позволило в качестве меры периодичности взять величину:

$$Z(n) = \frac{\max F(n) - \max\overline{F}(n)}{D(\max F(n))^{0,5}}. \qquad (7)$$

В результате нами были построены зависимости $Z(n)$ для нескольких аминокислотных последовательностей S (рис. 1, пункты 6–9).

Создание символьной последовательности из числовой последовательности. В данной работе для демонстрации мощности метода мы также исследовали периодичность числовой последовательности курса евро/доллар. В качестве изучаемой числовой последовательности выбирались последовательности курсов, где открытие и закрытие свечи было разделено интервалом в 24 ч и в 1 ч. Пусть $x_1(i)$ – курс на момент открытия свечи, а $x_2(i)$ – курс на момент закрытия свечи. В качестве первой последовательности A_1 рассчитывали разницу $s(i) = x_2(i) - x_1(i)$, где $x_1(i)$ и $x_2(i)$ разделены сутками. Начало свечи приходится на 0.00 часов, а конец свечи на 24.00 часов московского времени. Мы перекодировали числовые последовательности A_1 длиной N в символьную последовательность S_1 с алфавитом из 20 букв (рис. 1, пункт 1) следующим образом. Для перекодировки символьной последовательности мы определяли минимальный и максимальный элемент последовательности A_1, после чего этот интервал разбивался на 20 таких интервалов, чтобы число элементов ряда в каждом интервале было равно приблизительно N/20. Каждому интервалу была поставлена в соответствие буква латинского алфавита. Если числовой ряд содержал множество одинаковых значений, границы интервалов варьировались нами таким образом, чтобы все одинаковые значения ряда кодировалась одним и тем же символом. Последовательность S_1 была получена для данных с 16.02.2001 по 02.09.2014 и всего содержала 4522 суток.

В качестве последовательности A_2 рассчитывали разницу $s(i) = x_2(i) - x_1(i)$, где $x_1(i)$ и $x_2(i)$ разделены одним часом. Это означает, что начало свечи приходится на начало каждого часа, а конец свечи на конец каждого часа. Перевод последовательности A_2 в символьную последовательность осуществлялся так же, как и для последовательности A_1. При переводе этой последовательности в символьную последовательность мы получили последовательность S_2. Последовательность S_2 была получена для данных с 01.01.2014 по 02.09.2014, она содержала 4275 часов. Числовые данные были взяты нами с сайта http://finam.ru, время московское. Полученная перекодировка показана в табл. 1 и 2.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Выбор порогового значения для Z_0. На рис. 2 показано применение разработанного подхода к случайной последовательности и к последовательности S_3, полученной из последовательности S_3' = (EFKLWNMSTWRYLQKLWQ SMETMQ)$_{16}$. В последовательность S_3' случайным образом было внесено 75% случайных замен, и после этого был применен разработанный нами подход. Результаты анализа случайной последовательности показывают, что значения $Z(n)$ колеблются в районе 4,0 ± 2,6. Только для длин периодов больше 70 заметен небольшой тренд к возрастанию значений Z для случайной последовательности. Из этих результатов можно сделать вывод, что представляют интерес периоды со значениями Z > 7,0. Результаты изучения последовательности S_3 показывают, что разработанный подход оказался способным выявить сильно размытую периодичность и определить оптимальную весовую матрицу для обнаружения периодичности в 24 символа. Видна также периодичность в 48 символов и в 72 символа. Однако из-за возможности вставок или же делеций большие значения

Таблица 1. Кодировка последовательности A_1 для получения последовательности S_1

K	N	I	M	T	R	S	L	Y	F
–0,03700	–0,01140	–0,00780	–0,00550	–0,00390	–0,00260	–0,00160	0,00100	0,00040	–0,00010
–0,01140	–0,00780	–0,00550	–0,00390,	–0,00260	–0,00160	–0,00100	–0,00040	–0,00010	0,00020
C	W	P	H	Q	V	A	D	E	G
0,00020	0,00060	0,00100	0,00150	0,00230	0,00320	0,00460	0,00610	0,00820	0,01140
0,00060	0,00100	0,00150	0,00230	0,00320	0,00460	0,00610	0,00820	0,01140	5,00000

Примечание. Если значение элемента последовательности A_1 попадает в указанный интервал, то оно кодируется символом, указанным в верхней строке таблицы.

Таблица 2. Кодировка последовательности A_2 для получения последовательности S_2

K	N	I	M	T	R	S	L	Y	F
–0,00880	–0,00120	–0,00080	–0,00060	–0,00040	–0,00030	–0,00020	0,00020	–0,00010	0,00000
–0,00120	–0,00080	–0,00060	–0,00040,	–0,00030	–0,00020	–0,00020	–0,00010	0,00000	0,00000
C	W	P	H	Q	V	A	D	E	G
0,00000	0,00010	0,00010	0,00020	0,00020	0,00030	0,00040	0,00060	0,00080	0,00110
0,00010	0,00010	0,00020	0,00020	0,00030	0,00040	0,00060	0,00080	0,00110	0,00100

Примечание. Если значение элемента последовательности A_2 попадает в указанный интервал, то оно кодируется символом, указанным в верхней строке таблицы.

Z получили также периоды с длиной, близкой к 48 и 72 символам. Всегда можно за счет вставок или же делеций выровнять последовательность S_3 так, чтобы она имела периодичность, близкую к 48 или 72 символам. Однако это приводит к некоторому падению значений Z, что связано со штрафами за вставки или делеции. Поэтому график в районе максимумов около 48 и 72 символов имеет вид пологой горы.

Затем мы анализировали 300 случайно выбранных из банка данных Swiss-prot [16] аминокислотных последовательностей и для каждой из них рассчитали спектр $Z(n)$. В процессе выбора мы исключили из множества последовательности с любыми видами уже известных аминокислотных повторов или повторяющихся доменов [17]. Выбор числа последовательностей, равным 300, был связан с большой вычислительной сложностью разработанного алгоритма. Анализ 300 последовательностей потребовал около шести месяцев расчетов на компьютерном кластере с 10 процессорами AMD FX-8350. Кроме того, мы не ставили себе задачу проанализировать весь банк данных Swiss-prot, что потребовало бы очень больших компьютерных ресурсов. Мы имели задачу показать, что в аминокислотных последовательностях существует периодичность с большим количеством замен аминокислот, где присутствуют также вставки и делеции аминокислот. Эту периодичность можно обнаружить только с помощью разработанного нами подхода и невозможно выявить другими методами. Для решения данной задачи нам хватило 300 аминокислотных последовательностей. В результате была обнаружена 71 последовательность, для которых $Z(n) > 7{,}0$. Одновременно с этим мы перемешали все 300 последовательности, как было описано в разделе «Генерация множества случайных последовательностей» для множества Q. В этом множестве случайных последовательностей мы не обнаружили последовательностей, для которых $Z(n) > 7{,}0$.

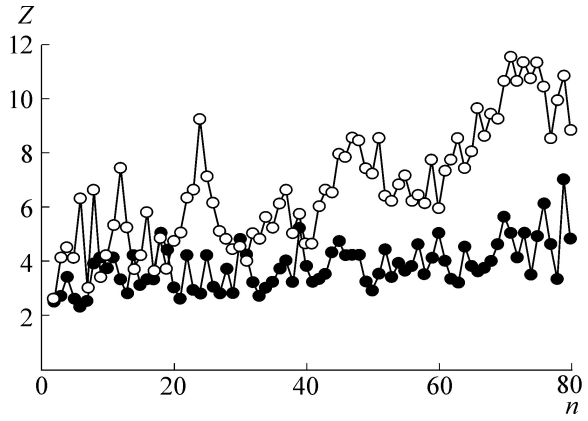

Рис. 2. Темными кружками показаны $Z(n)$ для случайной последовательности. Светлыми кружками показаны $Z(n)$ для последовательности с периодом в 24 символа и общей длиной в 384 символа с 75% случайных замен. Из рисунка видно, что, несмотря на большое количество случайных замен, разработанный математический метод смог заметить периодичность при $n = 24$.

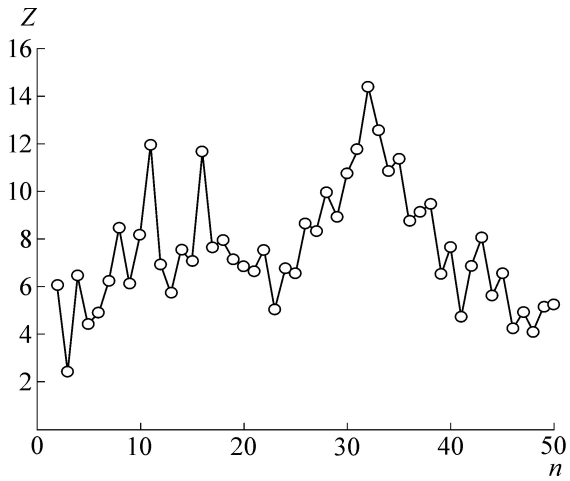

Рис. 3. Спектр $Z(n)$, полученный для последовательности комплексной субъединицы RsxC, участвующей в транспорте электронов (последовательность Q0THJ8).

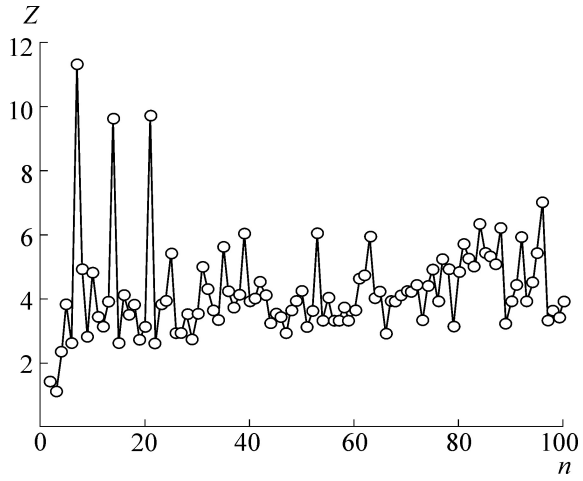

Рис. 4. Спектр $Z(n)$, полученный для последовательности A4GSN8, которая содержит белок, являющийся компонентом ядерной поры.

Рассмотрим четыре типичных примера скрытой периодичности со вставками и делециями. Первый пример относится к последовательности комплексной субъединицы RsxC, участвующей в транспорте электронов (последовательность Q0THJ8). Эта субъединица выделена из бактерии *E. coli*. В этой последовательности [18] на участке от 525 до 752 аминокислоты наблюдается периодичность длиной в 32 аминокислоты с $Z(32) = 14{,}3$. График для $Z(n)$ показан на рис. 3. Периодичность можно заметить только в присутствии пяти делеций и вставок различной длины в различных позициях последовательности. Пологая структура максимума на рис. 3 связана с возможностью создать статистически значимое выравнивание аминокислотной последовательности со вставками или делециями для длин периодов, близких к 32. В этом случае можно сделать несколько вставок и получить достаточно большое значение Z для длин периодов, близких к 29. На рис. 3 можно также заметить, что в данной последовательности выделяются минорные максимумы для длин в 11 и 16 аминокислот, которые «наводятся» основным периодом в 32 аминокислоты.

На рис. 4 показан второй пример спектра $Z(n)$ для последовательности A4GSN8, которая содержит белок, являющийся компонентом ядерной поры [19]. Участок с 28 по 2077 аминокислоту содержит периодичность длиной в 7 аминокислот, которую можно заметить только с делециями и вставками. $Z(7)$ для этого участка имеет максимальное значение и оно равно 11,6. Этот участок содержит семь протяженных суперспиральных (coiled coil) областей. Выравнивание содержит 23 делеции и вставки различной длины, т.е. средняя длина между вставками составляет ~100 аминокислот. Для суперспиральной области характерна периодичность длиной в семь аминокислот [20] вида *HPPHCPC,* где позиции периода обозначаются как *abcdefg*. Здесь *H* показывает положение гидрофобных аминокислот, *C* представляет, обычно, заряженные аминокислоты, а *P* представляет полярные (и, следовательно, гидрофильные) аминокислоты. Позиции семи аминокислотных мотивов обычно обозначаются строчными буквами от *a* до *g*. Эти аминокислотные мотивы являются основой для создания глобулярных белков, в частности лейциновой «застежки», которые имеют лейцин преимущественно в положении *d* в седьмом аминокислотном повторе. Как видно из табл. 3, наблюдаемая в последовательности Q1D823 периодичность отличается от периодичности, характерной для суперспиральных областей. Только столбец 7 таблицы немного похож на позицию *a* этого повтора. Остальные позиции отличаются от наблюдаемого ранее мотива длиной в семь аминокислот [20]. Мы может предполагать, что существуют различные повторы длиной в семь аминокислот, которые могут образовывать суперспиральные области. Вероятно также, что такое отличие связано со вставками или же делециями аминокислот. Полученная нами матрица, вероятно, может быть использована для поиска протяженных областей, имеющих суперспиральную структуру в различных белках.

На рис. 5 представлен третий пример, где показана периодичность участка из последовательности O-ацил-гликозамин-N-ацетилтрансферазы (Q31B90) с 11-й по 301-ю аминокислоту

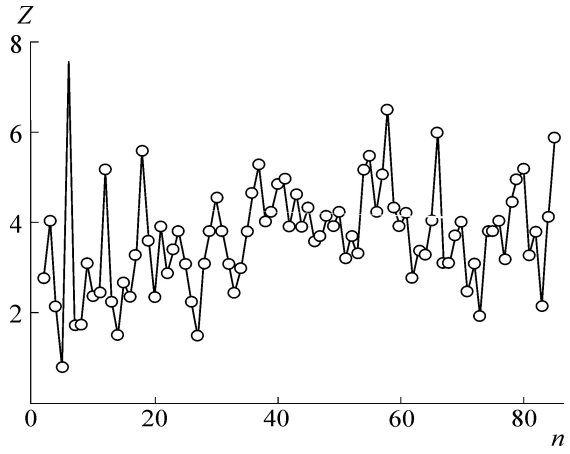

Рис. 5. Спектр $Z(n)$, полученный для последовательности О-ацил-гликозамин-N-ацетилтрансферазы. (Q31B90).

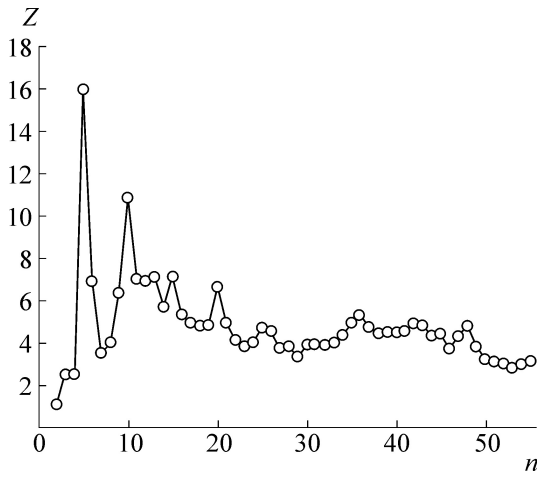

Рис. 6. Спектр $Z(n)$, полученный для последовательности Q46397, которая содержит белок, подобный гистону H1.

[21]. Это почти полная последовательность данного белка. Участок имеет периодичность длиной шесть аминокислот и $Z(6) = 7{,}8$. Этот период может быть образован в аминокислотных последовательностях 3_{10}-спиралью [22,23]. Можно предполагать, что данная аминокислотная последовательность может на всем своем протяжении переходить в 3_{10}-спираль. В этом случае полученная нами матрица может быть использована для выявления потенциальных 3_{10}-спиралей.

На рис. 6 показан спектр $Z(n)$ для аминокислотной последовательности Q46397 [24] на участке со 2-й по 154-ю аминокислоту. Эта последовательность содержит белок, подобный гистону H1. Совершенно отчетливо виден период длиной в пять аминокислот. Участок захватывает бóльшую часть аминокислотной последовательности, он был обнаружен только в присутствии трех вставок или делеций аминокислот размером в одну аминокислоту. Без вставок и делеций выявить эту периодичность невозможно.

Возникает закономерный вопрос о роли наблюдаемой периодичности в структуре и функции белков. Можно выдвинуть два предположения относительно функциональной роли обнаруженной нами периодичности в белках. Во-первых, найденная нами периодичность является неким свойством, обеспечивающим образование определенных вторичных структур [25]. Такое предположение уже было высказано для аминокислотных повторов, которые были обнаружены ранее [17,26]. В нашем случае это периоды в семь и шесть аминокислот, которые могут участвовать в образовании α-спиралей. Можно предполагать, что периодичность в пять аминокислот может быть ответственна за образование *pi*-спиралей в белках. С учетом возможности вставок и делеций при построении $Z(n)$ период в 4,1 аминокислоты может трансформироваться в период, составляющий пять

Таблица 3. Весовая матрица для аминокислот, с помощью которой была обнаружена периодичность длиной в семь аминокислот в последовательности A4GSN8, которая содержит белок, являющийся компонентом ядерной поры

	1	2	3	4	5	6	7
K	0,3	2,7	0,6	–2,3	–4,2	2,1	–1,0
N	–0,8	1,3	–2,4	2,5	1,3	–1,3	–2,9
I	–1,8	–3,3	–2,8	2,7	–0,9	–0,9	2,1
M	–1,7	1,8	0,2	–0,4	–1,7	–2,9	2,4
T	–2,7	2,3	0,1	–2,7	1,3	2,2	–1,9
R	2,4	–2,4	1,8	–2,7	1,0	–0,5	–0,1
S	1,9	–0,8	–3,6	–3,0	2,9	1,9	–2,0
L	–3,4	–6,7	–3,1	3,1	–5,2	–4,0	3,4
Y	–0,1	–3,3	–1,7	2,4	–0,9	–2,5	2,0
F	1,6	–1,0	0,8	1,9	–1,9	–1,0	–1,0
C	–1,0	–2,2	–2,2	2,3	–2,2	–2,2	2,3
W	–1,6	–1,6	–1,6	2,0	0,0	1,6	0,0
P	0,8	1,8	1,4	–0,5	–1,7	–1,1	–0,5
H	0,4	–1,1	–1,1	–1,1	1,9	1,1	–0,4
Q	2,8	2,2	0,6	–3,5	–3,5	–0,2	–3,1
V	–0,3	–2,9	–0,8	2,8	–1,2	–2,5	0,5
A	–5,9	2,4	0,9	0,6	–0,7	–1,4	2,5
D	1,0	–0,3	2,4	–5,2	2,4	1,9	–4,7
E	2,5	1,5	3,0	–6,7	2,4	2,4	–8,1
G	–1,4	0,4	–1,4	–2,0	1,8	2,1	–0,8

Примечание. Периодичность показана на рис. 4.

Таблица 4. Весовая матрица для аминокислот, с помощью которой была обнаружена периодичность длиной в шесть аминокислот в последовательности O-ацил-гликозамин N-ацетилтрансферазы (Q31B90)

	1	2	3	4	5	6
K	1,7	3,6	–1,7	–3,0	0,8	–1,7
N	–18	–4,6	–4,6	1,0	1,0	4,1
I	–3,1	–0,4	4,6	–5,8	–4,0	–3,1
M	–1,3	–1,3	2,4	2,4	–1,3	–1,3
T	3,0	1,1	–2,8	–2,8	1,1	–0,9
R	–2,1	2,5	–2,1	–2,1	0,5	2,5
S	–0,0	–2,7	–2,7	2,5	1,3	2,5
L	–2,1	–3,6	2,9	0,9	–2,1	2,9
Y	1,7	1,7	–1,6	1,7	–1,6	–1,6
F	–3,0	2,4	0,7	–3,0	2,9	–1,1
C	3,7	–0,3	–2,5	–2,5	–2,5	–2,5
W	–1,0	–1,0	–1,0	–1,0	–1,0	2,9
P	2,8	–2,5	–2,5	–0,3	1,9	–0,3
H	–0,3	–0,3	–2,5	2,8	1,9	–2,5
Q	–1,9	2,8	–1,9	–1,9	–1,9	2,8
V	3,8	2,9	–0,9	–3,3	–3,3	–4,5
A	1,1	–2,8	–1,5	2,8	2,8	–2,8
D	–3,5	–0,4	–0,4	3,2	1,2	–1,9
E	2,6	2,6	–0,1	–3,4	1,5	–3,4
G	–4,9	0,6	–4,9	3,8	0,6	0,6

Применчание. Периодичность показана на рис. 5.

Таблица 5. Весовая матрица для аминокислот, с помощью которой была обнаружена периодичность длиной в семь аминокислот в последовательности Q46397, которая содержит белок, подобный гистону H1

	1	2	3	4	5
K	7,2	–5,8	–5,8	–5,8	6,2
N	0,0	0,0	0,0	0,0	0,0
I	0,0	0,0	0,0	0,0	0,0
M	0,1	0,1	0,1	0,1	0,1
T	–4,1	5,3	0,3	–2,7	–4,1
R	–3,8	–2,2	–3,8	–3,8	5,3
S	–1,0	–1,0	3,4	–1,0	–1,0
L	3,3	–1,2	–1,2	–1,2	–1,2
Y	0,0	0,0	0,0	0,0	0,0
F	0,0	0,0	0,0	0,0	0,0
C	0,1	0,1	0,1	0,1	0,1
W	0,0	0,0	0,0	0,0	0,0
P	–3,1	3,7	3,7	–3,1	–3,1
H	0,0	0,0	0,0	0,0	0,0
Q	–1,2	–1,2	–1,2	3,3	–1,2
V	–4,5	–2,1	5,9	1,7	–4,6
A	–5,2	3,9	–0,2	6,7	–5,3
D	0,0	0,0	0,0	0,0	0,0
E	0,0	0,0	0,0	0,0	0,0
G	–1,1	3,3	–1,1	–1,1	–1,1

Применчание. Периодичность показана на рис. 6

аминокислот [27]. Периодичность длиной в пять аминокислот может быть также отражениием способности аминокислотной последовательности принимать форму $2,2_7$-спирали, где на один период приходится 2,2 аминокислоты. В этом случае два периода могут образовывать период длиной в пять аминокислот, регистрируемый нашим методом с учетом возможности вставок или делеций.

Во-вторых, найденная нами периодичность может отражать определенную пространственную повторяемость в составе 3D-структуры. Для известных повторов это можно наблюдать для последовательностей, известных как «цинковые пальцы» (Zn-finger domains) [28], для Ig-домена [29] и для металлопротеазного домена [30]. В работе [17] проведена структурная классификация белков на основе длин наблюдаемой периодичности. Происхождение множественных тандемных повторов в белках может быть связано с процессами множественных тандемных дупликаций в ДНК [31]. Это может приводить к образованию новых белков [32]. Дальнейшая эволюция и накопление мутаций (замены аминокислот и делеции и вставки аминокислот) могут привести к созданию скрытой периодичности со множеством замен аминокислот и вставок или же делеций аминокислот. Именно такую периодичность мы можем обнаруживать в настоящей работе.

Далее мы применили разработанный подход к поиску периодичности для символьных последовательностей S_1 и S_2, полученных из числовых последовательностей A_1 и A_2 (раздел «Создание символьной последовательности из числовой последовательности»). Спектр $Z(n)$, полученный для этих последовательностей, показан на рис. 7 и 8 соответственно. Из рис. 7 видно, что последовательность S_1 имеет два периода. Первый период равен 6 сут, он имеет наибольшее значение Z с 565-х по 2009-е сутки последовательности S_1 и был найден в присутствии 11 делеций и вставок различного размера. Второй период равен семи суткам, он имеет наибольшее значение Z с 1993-х по 4383-е сутки и найден с делециями и вставками различного размера общей численностью 12 штук. На рис. 7 также выделяются кратные периоды, равные

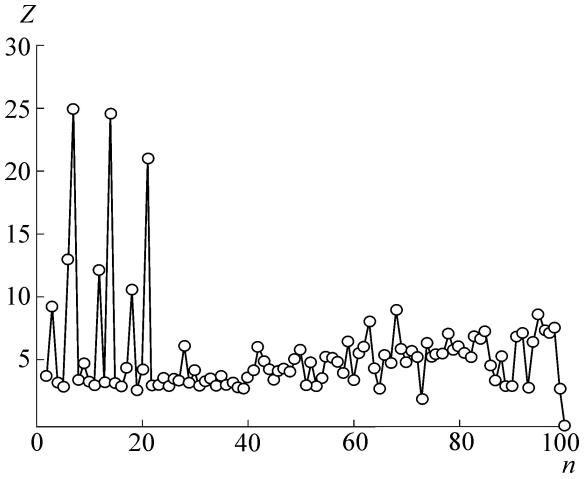

Рис. 7. Спектр $Z(n)$, полученный для последовательности S_1.

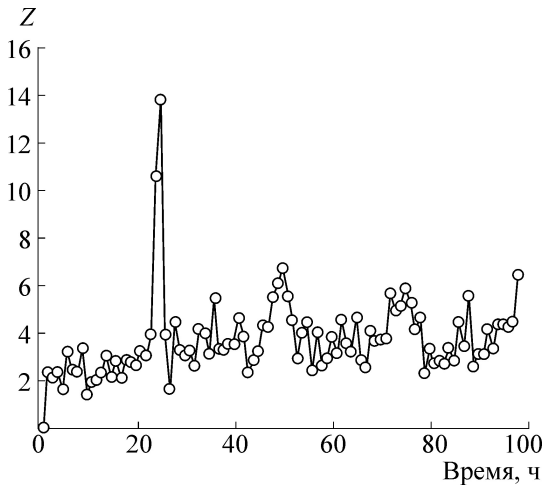

Рис. 8. Спектр $Z(n)$ полученный для последовательности S_2.

12-ти и 14-ти суткам. Данные по значениям $Z(n)$ для последовательности S_2 показаны на рис. 8. Для последовательности S_2 наблюдается периодичность в 24 и 25 ч. При этом обе периодичности наблюдаются почти для одних и тех же данных. Период в 24 ч наблюдается с 29-го по 4165-й ч последовательности S_2, он найден в присутствии 46 делеций и вставок различного размера. Период в 25 ч наблюдается с 29-го по 4224-й час последовательности S_2 и найден в присутствии 52 делеций и вставок различного размера. Период в 25 ч наиболее статистически выражен ($Z \approx 14{,}0$), тогда как для длины периода в 24 ч $Z \approx 10{,}5$. Можно предполагать, что построение значимых выравниваний сразу для двух периодов связано с возможностью создания вставок или же делеций символов в последовательности S_2.

Из проделанной работы возникает закономерный вопрос о природе найденной периодичности в курсе евро/доллар. На валютные курсы оказывают влияние различные периодические процессы. Самому человеку присущи ритмы различной частоты. Например, в работе [33] проведена классификация биоритмов человека по длине периода. Выделено несколько групп ритмов. 1. Низкочастотная группа – циркасептанные (период 7 ± 3 сут), циркадисептанные (период 14 ± 3 сут), циркавигинтанные (период 20 ± 3 сут), циркатригинтанные (период 30 ± 3 сут), цирканнуальные ритмы (период 12 ± 2 месяцев). 2. Среднечастотная группа – циркадные ритмы (период 20–28 ч), инфрадианные (период 28–72 ч). 3. Высокочастотная зона – ультрадианные ритмы (период до 20 ч). Поэтому можно предполагать, что наблюдаемая периодичность валютного курса в шесть и семь суток является отражением низкочастотных ритмов человека. Периодичность в 24 и 25 ч отражает влияние среднечастотной группы ритмов человека на валютный курс.

Мы также строили последовательность S_2 для других интервалов времени, но такой же длины. В этих случаях спектр периодичности для периодов в 24 и 25 ч был аналогичным. Интересно также попытаться объяснить присутствие вставок или же делеций символов, без которых выделить периодичность невозможно. Вероятно, существует большая нестабильность в поведении больших людских масс, и эта нестабильность может создавать сдвиги фазы периодичности, что будет восприниматься нашим методом как вставки или же делеции. Также можно предполагать, что те или иные события общественной жизни могут быть причиной сдвига фазы. Для более точного рассмотрения этого вопроса требуется отдельное изучение по поиску корреляций между сдвигами фазы в последовательностях S_1 и S_2 и событиями общественной жизни и другими факторами как социальной, так и физической природы.

Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (грант 04-00164-2014).

СПИСОК ЛИТЕРАТУРЫ

1. R. Ekblom and J. B. W. Wolf, Evol. Appl. **7** (9), 1026 (2014).
2. E. V. Korotkov, M. A. Korotkova, and N. A. Kudryashov, Phys. Lett. Sect. A Gen. At. Solid State Phys. **312** (3-4), 198 (2003).
3. E. V. Korotkov, M. A. Korotkova, and N. A. Kudryashov, Mol. Biol. (Mosk). **37** (3), 436 (2003).

4. V. Afreixo, P. J. S. G. Ferreira, and D. Santos, Digit. Signal Process **14** (6), 523 (2004).
5. G. I. Kravatskaya, Y. V Kravatsky, V. R. Chechetkin, and V. G. Tumanyan, Genomics **98** (3), 223 (2011).
6. V. V. Lobzin, Uspekhi Fiz. Nauk **170** (1), 57 (2000)
7. T. Meng, A. T. Soliman, M.-L. Shyu, et al., IEEE/ACM Trans. Comput. Biol. Bioinform. **10** (6), 1442 (2013).
8. M. de Sousa Vieira, Phys. Rev. E. Stat. Phys. Plasmas. Fluids. Relat. Interdiscip. Topics **60** (5 Pt B), 5932 (1999).
9. Y. M. Suvorova, M. A. Korotkova, and E. V. Korotkov, Comput. Biol. Chem. **53** (Pt A), 43 (2014).
10. S. Tiwari, S. Ramachandran, A. Bhattacharya, et al., Comput. Appl. Biosci. CABIOS **13** (3), 263 (1997).
11. A. Agrawal and X. Huang, IEEE/ACM Trans. Comput. Biol. Bioinform. **8** (1), 194 (2009).
12. T. F. Smith and M. S. Waterman, J. Mol. Biol. **147**, 195 (1954).
13. A. Shelenkov, K. Skryabin, and E. Korotkov, DNA Res. **13** (3), 89 (2006).
14. *Information Theory and Statistics*, Ed. by S. Kullback (Dover publications., New York, 1997).
15. V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, Algorithms Mol. Biol. **6** (1), 25 (2011).
16. B. Boeckmann, A. Bairoch, R. Apweiler, et al., Nucl. Acids Res. **31** (1), 365 (2003).
17. A. V. Kajava, J. Struct. Biol. **179** (3), 279 (2012).
18. B. Hochhut, C. Wilde, G. Balling, et al., Mol. Microbiol. **61** (3), 584 (2006).
19. K. Tamura, Y. Fukao, M. Iwamoto, et al., Plant Cell **22** (12), 4084 (2010).
20. J. M. Mason and K. M. Arndt, Chembiochem **5** (2), 170 (2004).
21. A. Copeland, S. Lucas, A. Lapidus, et al., Submitt. to EMBL/GenBank/DDBJ databases. US DOE Jt. Genome Inst. (2005)
22. P. Enkhbayar, K. Hikichi, M. Osaki, et al., Proteins **64** (3), 691 (2006).
23. P. Fábián, V. S. Chauhan, and S. Pongor, Biochim. Biophys. Acta **1208** (1), 89 (1994)
24. T. J. Brickman, C. E. Barry, and T. Hackstadt, J. Bacteriol. **175** (14), 4274 (1993).
25. K. K. Jernigan and S. R. Bordenstein, PeerJ **3**, e732c (2015).
26. J. Jorda, B. Xue, V. N. Uversky, and A. V Kajava, FEBS J. **277** (12), 2673 (2010).
27. M. N. Fodje and S. Al-Karadaghi, Protein Eng. Des. Sel. **15** (5), 353 (2002).
28. M. S. Lee, G. P. Gippert, K. V Soman, et al., Science **245** (4918), 635 (1989).
29. M. R. Sawaya, W. M. Wojtowicz, I. Andre, et al., Cell **134** (6), 1007 (2008).
30. P. A. Elkins, Y. Sen Ho, W. W. Smith, et al., Acta Crystallogr. D. Biol. Crystallogr. **58** (Pt 7), 1182 (2002).
31. A. De Grassi and F. D. Ciccarelli, Genome Biol. **10** (12), R137c (2009).
32. A. K. Björklund, D. Ekman, and A. Elofsson, PLoS Comput. Biol. **2** (8), e114c (2006).
33. F. Halberg, Annu. Rev. Physiol. **31** (1), 675 (1969).

Developing the Mathematical Method in order to Search for Latent Periodicity in Protein Amino Acid Sequences with Deletions and Insertions

E.V. Korotkov* and M.A. Korotkova**

**Federal Center "Fundamental Bases of Biotechnology", Russian Academy of Sciences, prosp. 60-letiya Oktyabrya 7/1, Moscow, 117312 Russia*

***National Research Nuclear University, Moscow Engineering Physics Institute, Kashirskoe shosse 31, Moscow, 105409 Russia*

A mathematical method was developed in order to search for latent periodicity in protein amino acid and other symbolical sequences using the dynamic programming and random matrixes. The method permits detection of the latent periodicity with insertions and deletions in the previously unknown positions. The developed method was applied to search for the periodicity in the amino acid sequences of some proteins and the periodicity in EUR/USD exchange rate since 2001. The presence of the long period length with insertions and deletions in amino acid sequences was shown. The period length of 7 amino acids was found in proteins containing supercoiled areas (coiled coil), the period length of 6 and 5 and more amino acids was also demonstrated. The existence of the period length of 6 and 7 days as well as 24 and 25 hours in the analyzed financial time series, which can be detected with insertions and deletions only, is revealed. The reasons of the occurence of the latent periodicity with insertions and deletions in the amino acid sequences and financial time series are discussed.

Key words: latent periodicity, dynamic programming, amino acid sequences, euro, dollar