МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2013, том 47, № 2, с. 347–355

БИОИНФОРМАТИКА

УДК 577.322.4

ВЫЧИСЛИТЕЛЬНЫЕ МЕТОДЫ ПРЕДСКАЗАНИЯ ВТОРИЧНОЙ СТРУКТУРЫ МЕМБРАННЫХ БЕЛКОВ ПО ИХ АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

© 2013 г. М. Н. Симакова[1]*, Н. Н. Симаков[2]

[1]Institute of Complex Systems, ICS-5: Molecular Biophysics, Research Center-Juelich, 52425 Juelich, Germany

[2]Ярославский государственный технический университет, 150023, Ярославль, Россия

Поступила в редакцию 25.04.2012 г.
Принята к печати 17.08.2012 г.

Структура мембранных белков интересна тем, что обуславливает их функциональные свойства, важные для медицины и фармакологии. Особенность и органическое свойство политопических мембранных белков – повторяемость трансмембранных участков, состоящих из аминокислот гидрофобной группы. Упорядоченную повторяемость (периодичность) можно выявить методом Фурье, применяя его к цифровому образу символьной последовательности аминокислот в белке. В данной работе это сделано для 24 трансмембранных белков, для 14 из них – успешно. Если повторяемость трансмембранных участков непериодическая, то для ее выявления можно использовать другой метод – метод многократного (4–5 раз) усреднения функции гидрофобности белка в пределах перемещаемого вдоль последовательности "окна" шириной 9–11 а.о. Этот новый метод применен к тем же 24 трансмембранным белкам (успешно для 19 из них) и показал большую, чем метод Фурье, пригодность для прогнозирования вторичной структуры такого рода белков и соответствующих ей функциональных свойств.

***Ключевые слова*: мембранные белки, вторичная структура, трансмембранные участки, повторяемость, периодичность, метод Фурье, метод перемещаемого "окна".**

COMPUTATIONAL METHODS FOR PREDICTION OF STRUCTURE OF MEMBRANE PROTEINS USING THEIR AMINO ACIDS SEQUENCES, *by M. N. Simakova[1],*, N. N. Simakov[2]* ([1] Institute of Complex Systems, ICS-5: Molecular Biophysics, Research Center-Juelich, 52425 Juelich, Germany; *e-mail: m_simakova@mail.ru; [2]Yaroslavl State Technical University, Yaroslavl, 150023 Russia). The structure of membrane proteins is interesting because of their functional properties that are important to medicine and pharmacology. The feature and an organic property of polytopic membrane proteins is the repetition of transmembrane regions consisting of hydrophobic amino acids. The ordered repetition – periodicity – can be identified by the Fourier method, applied to a digital image of symbolic sequence of amino acids in a protein. In this work it was carried out for the 24 transmembrane proteins, for 14 of them successfully. If the repetition of transmembrane regions is ordered insufficiently – non-periodic, then a different method is supposed to use for its detection – the method of multiple (4–5 times) averaging of function of hydrophobicity of the protein within a "window" with width 9–11 aa moved along the sequence. This new method was applied to the same 24 transmembrane proteins (for 19 of them successfully) and it was shown to be more suitable (than the Fourier method) for predicting of the secondary structure of such proteins and functional properties corresponding to it.**

***Keywords*: membrane proteins, secondary structure, transmembrane regions, repetition, periodicity, Fourier method, method of "moving window".**

DOI: 10.7868/S0026898413010151

ОСОБЕННОСТИ СТРУКТУРЫ И ФУНКЦИИ МЕМБРАННЫХ БЕЛКОВ

Исследования мембранных белков играют очень важную роль, как для развития биологической науки, так и для фармакологии и медицины. Мембранные белки составляют примерно 25% всех белков, кодируемых геномом человека. Они отвечают за многие функции клеток, в частности: обеспечивают избирательный обмен веществ между клеткой и окружающей ее средой, поддер-

Принятые сокращения: ТМУ – трансмембранный участок, домен белковой последовательности; TPD – топологический домен; $N_{TMУ}$ – число трансмембранных участков в белковой последовательности; а.о. – аминокислотные остатки.

* Эл. почта: m_simakova@mail.ru

живают разность электрических потенциалов внутри клетки и снаружи, обеспечивают передачу электрических сигналов в клетку и из нее, участвуют в производстве и переносе энергии в организме. Поэтому действие более половины фармацевтических препаратов направлено именно на мембранные белки.

Функции, выполняемые белком, определяются его пространственной трехмерной структурой. Определение этой структуры белка представляет собой очень сложную проблему. Основным экспериментальным методом детального исследования структуры мембранных белков является рентгеноструктурный анализ [1, 2]. Этот метод непосредственно применяют к кристаллам белка, которые необходимо предварительно получить, используя очень сложную и трудоемкую технологию. Она включает в себя следующие основные стадии: создание плазмидной конструкции, содержащей ген, кодирующий данный белок; экспрессию белка в клетках реципиента; разрушение клеток и отделение клеточных мембран, солюбилизацию белка в присутствии соответствующего детергента; очистку и концентрирование белка; реконституцию белка и его кристаллизацию. Наиболее критичными для достижения конечной цели являются стадия очистки из-за значительных потерь белка (до 40% на каждом из нескольких этапов) и стадия его кристаллизации из-за неопределенности необходимых условий ее проведения для данного конкретного белка. В силу указанных трудностей структура мембранных белков установлена к настоящему времени лишь для очень ограниченной их части (<4%).

С учетом перечисленных обстоятельств актуальной альтернативой экспериментальному методу рентгеновской кристаллографии при исследовании структуры мембранных белков стали методы компьютерного моделирования, основанные на использовании данных о первичной структуре белка – символьной последовательности его аминокислот [3–8]. Так, например, в работе [7] исследовали периодичность расположения аминокислот в фибриллярных белках бактериофага T4 методом преобразования Фурье цифрового образа символьной последовательности аминокислот белка. В частности, изучена повторяемость расположения аминокислотных остатков (а.о.), принадлежащих к определенным группам, например, гидрофобным или гидрофильным. В численном эксперименте для нескольких исследованных белков выявлена повторяемость с характерными малыми периодами $T = 8$, 10, 15 а.о. и большими $T = 46$, 55, 215, 256 а.о. Последние делят всю последовательность белка на 4 или 6 равных частей и могут быть признаками его третичной структуры [3]. Результаты численного моделирования учтены в успешном эксперименте, направленном на получение рекомбинантной формы белка gp37 посредством создания химерного белка [8].

Цель данной работы – выяснение применимости к исследованию вторичной структуры мембранных белков двух частных методов компьютерного моделирования: известного метода с использованием преобразования Фурье и усовершенствованного нами метода “скользящего окна” и сравнение их между собой и с другими известными методами.

ИССЛЕДОВАНИЕ ПЕРИОДИЧНОСТИ РАСПОЛОЖЕНИЯ АМИНОКИСЛОТ ГИДРОФОБНОЙ ГРУППЫ В ПОСЛЕДОВАТЕЛЬНОСТЯХ МЕМБРАННЫХ БЕЛКОВ МЕТОДОМ ФУРЬЕ

Из многообразия всех мембранных белков значительный интерес вызывает группа интегральных трансмембранных белков, имеющих несколько гидрофобных участков, насквозь пронизывающих мембрану и совместно выполняющих функцию транспортных каналов для различных веществ. Очевидная особенность этих белков состоит в том, что свойство повторяемости (возможно, периодической) в расположении гидрофобных аминокислот им органически присуще. Если отмеченная повторяемость периодическая, то она как глобальное свойство и признак вторичной структуры белка может быть выявлена известным методом с использованием преобразования Фурье, как это сделано в случае фибриллярных белков [7].

При исследовании периодичности белковой последовательности 20 составляющих ее аминокислот делят на несколько (2–4) характерных групп по степени их гидрофобности (или гидрофильности), среди которых всегда присутствует гидрофобная группа. Группа гидрофильных аминокислот может быть одна или поделена на 2–3 подгруппы, в зависимости от знака заряда [9]. Это деление на группы в известной мере условно, так как его проводят по разным критериям и характеристикам.

В работах [7, 9] к первой группе относили гидрофобные аминокислоты, обозначаемые символами: A, F, G, I, L, V, W, Y. Остальные аминокислоты в работе [7] относили ко второй группе. В отдельных расчетах деление на группы проводили иначе. Затем символьную последовательность заменяли ее математическим образом $f(k) = f(x)$ на дискретном множестве $k = 1, 2, ..., L$ мест аминокислот в последовательности (или непрерывном множестве $0 < x < L$), присваивая M элементам, стоящим на местах k_m аминокислот интересую-

щей группы ($m = 1, 2, ..., M$) значения 1, а остальным – значение 0 по формуле:

$$f(x) = f(k) = F(k_m) = \sum_{m=1}^{M} \delta(x - k_m), \quad (1)$$

где $\delta(x)$ – обобщенная функция Дирака [10].

Спектр Фурье для функции (1) можно вычислить по формулам

$$I(\omega_n) = c(\omega_n)^2 = c_n^2 = a_n^2 + b_n^2, \quad (2)$$

$$a_n = \frac{2}{L}\sum_{m=1}^{M}\cos(\omega_n k_m),\ b_n = \frac{2}{L}\sum_{m=1}^{M}\sin(\omega_n k_m), \quad (3)$$

где $I = c_n^2$ – интенсивность гармоники с частотой ω_n [11], a_n и b_n – коэффициенты ряда Фурье

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos(\omega_n x) + b_n \sin(\omega_n x)) \cong$$
$$\cong \frac{a_0}{2} + \sum_{n=1}^{n_{\max}}(a_n \cos(\omega_n x) + b_n \sin(\omega_n x)), \quad (4)$$

представляющего функцию (1). Для ряда Фурье (4) набор частот спектра (2) – дискретный:

$$\omega_n = 2\pi n/L, \quad n = 0, 1, 2, ..., L. \quad (5)$$

Как указано в замечаниях об особенностях Фурье-спектров в статье [7], если на какой-то частоте $\omega_n = 2\pi n/L$ (периоде $T_n = 2\pi/\omega_n = L/n$, $n = 1, 2, ..., L$) обнаруживается достаточно интенсивный пик $I(T_n)$, то и на кратных ей больших частотах $\omega_{n\times p} = p \times \omega_n$ при $p = 2, 3, 4...$ (периодах $T_{n\times p} = T_n/p < T_n$) также могут быть обнаружены пики $I(T_{n\times p})$, сравнимой с $I(T_n)$ – несколько большей или меньшей – интенсивности. При этом все такие пики порождаются одной и той же периодичностью с основным – наибольшим – периодом T_n, и пики с меньшими периодами $T_{n\times p} = T_n/p$ можно не принимать во внимание. Если интенсивность $I(\omega_n)$ гармоник в спектре нормирована так, что $\langle I \rangle = 1$, то вероятность случайного превышения пиком уровней интенсивности $I(\omega_n) = 2, 3, 4$ составляет соответственно 16, 2.3, 0.13%.

Описанный здесь алгоритм [7] вместе с реализующей его компьютерной программой сначала применили к исследованию периодичности 6 трансмембранных белков. В нижеследующем перечне этих белков приведено название по-русски, по-английски, сокращенное обозначение, код белка и название организма в соответствии с источником [12]:

1) бактериородопсин, bacteriorhodopsin, bR, **P02945**, *Halobacterium salinarium*;

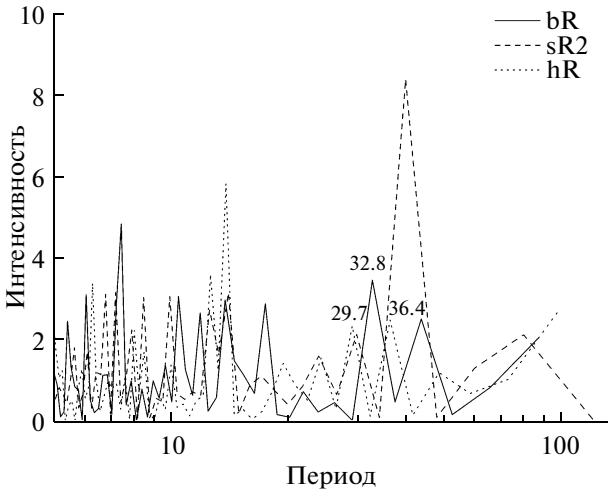

Рис. 1. Фурье-спектры для гидрофобной группы аминокислот в символьных последовательностях трансмембранных белков: bR, sR2, hR.

2) галородопсин, halorhodopsin, hR, **P15647**, *Natronomonas pharaonis*;

3) сенсорный родопсин 2, sensory rhodopsin-2, sR2, **P42196**, *Natronomonas pharaonis*;

4) коннексин-32, connexin-32, Cx32, **P08034**, *Homo sapiens*;

5) зависимый от циклического нуклеотида калиевый канал, cyclic nucleotide-gated potassium channel **mll3241**, CNG (обозначение, сокращенное авторами данной статьи, так как общепринятого нет), **Q98GN8**, *Mesorhizobium loti*;

6) метаботропный глутаматный рецептор 6, metabotropic glutamate receptor 6, GRM6, **O15303**, *Homo sapiens*.

Некоторые результаты расчета спектров Фурье представлены на рис. 1. Условием нормировки интенсивностей $I(\omega_n)$ гармоник в спектре было равенство $\langle I \rangle = 1$.

В спектре Фурье для последовательности bR (рис. 1) четко проявился период $T = 33$, который, согласно данным источника [12], в среднем равен общему числу членов в трансмембранном участке (ТМУ) и примыкающем к нему топологическом домене (TPD). Тогда число ТМУ (иначе говоря, число альфа-спиралей в цепи) можно приближенно определить по предлагаемой формуле

$$N_{\text{ТМУ}} = L/T - 1 = n - 1, \quad (6)$$

где L – вся длина цепи, T – главный период последовательности, определенный из ее спектра Фурье для аминокислот гидрофобной группы по пику с максимальной интенсивностью $I(T)$ среди пиков с большими значениями периодов T. Он равен средней длине участков повторения элементов цепи. Вычитание единицы в формуле (6) учитывает, что число TPD на единицу больше,

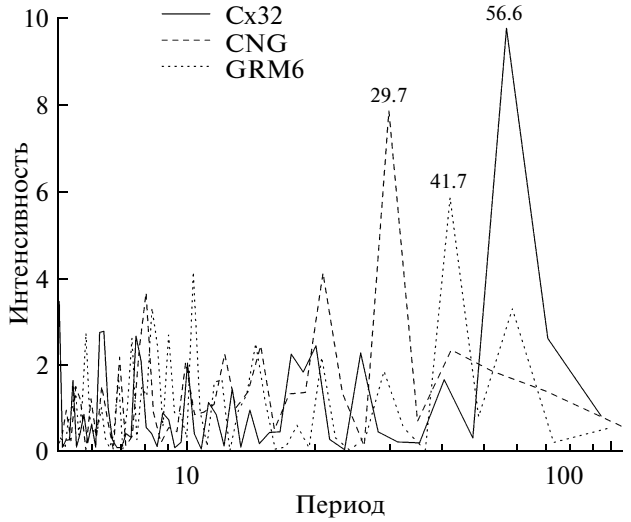

Рис. 2. Фурье-спектры для гидрофобной группы аминокислот в символьных последовательностях трансмембранных белков: Cx32, CNG, GRM6.

чем число ТМУ. Так, для последовательности bR (длиной $L = 262$ а.о.) проявился пик с периодом $T = 33$ и относительной интенсивностью $I(T) = 3.5$, для которого по формуле (6) находим $N_{\text{ТМУ}} = 7$, что совпадает с известным числом альфа-спиралей в структуре этого белка [12].

В спектре последовательности белка sR2 на рис. 1 очевиден пик с частотой $n = 6$, периодом $T_6 = L/n = 39.8$ и очень большой интенсивностью $I(T_6) = 8.38$. Но частота этого пика оказывается двукратной к частоте $n = 3$ другого, менее интенсивного пика с периодом $T_3 = L/n = 79.7$, слишком большим и потому маловероятным для ТМУ. Поэтому с учетом вышеприведенного замечания оба этих кратных по частоте пика можно исключить из рассмотрения. Следующим по интенсивности $I(T) = 2.12$ среди пиков с большим периодом $20 < T < 80$ является пик с периодом $T = 29.7$ и частотой $n = L/T \approx 8$. Для него по формуле (6) вычисляем число $N_{\text{ТМУ}} = 7$, что опять совпало с числом ТМУ в структуре белка, описанной в работе [12]. Заметим, что у этой цепи длиной $L = 239$ а.о. есть 7 ТМУ различной длины (22, 22, 22, 23, 28, 28, 33 а.о.), и 8 TPD имеют очень разные длины (3, 8, 14, 3, 4, 4, 8, 17 а.о.).

В спектре последовательности hR имеется пик с относительно большой интенсивностью $I(T_8) = 2.48$, большим периодом $T_8 = 36.4$ и частотой $n = 8$, для которого $N_{\text{ТМУ}} = 7$, что снова совпадает с числом ТМУ в предсказываемой по аналогии ("similarity" – [12]) структуре этого белка. В спектре Фурье присутствует также пик, сравнимый с вышеуказанным по интенсивности $I = 2.33$ и периоду $T = 29.1$. Это, по-видимому, обусловлено существенными нарушениями периодичности белковой последовательности hR по сравнению с bR и sR2. По данным о ее структуре, длины ТМУ примерно одинаковы (26, 24, 22, 24, 23, 24, 29 а.о.), а длины TPD (30, 6, 34, 3, 3, 12, 9, 22 а.о.) очень сильно (до 10 раз) отличаются [12]. В этом состоит особенность этого белка.

Стоит подчеркнуть, что три рассмотренные белковые последовательности (bR, sR2, hR) содержат в своей структуре по 7 альфа-спиралей, и метод преобразования Фурье это надежно выявляет.

В Фурье-спектре последовательности белка Cx32 (рис. 2) проявился пик с периодом $T = 56.6$ и очень большой относительной интенсивностью $I(T) = 9.8$, для которого по формуле (6) нашли $N_{\text{ТМУ}} = 4$, что в точности совпало с числом альфа-спиралей в вероятной структуре данного белка. И это удалось, несмотря на то, что, по литературным данным [12], при приблизительно равных длинах (23, 20, 23, 23 а.о.) ТМУ топологические домены могут значительно различаться по длине (22, 30, 35, 38, 69 а.о.), особенно "хвостовой" домен.

Белок CNG содержит 6 ТМУ примерно одинаковой длины (18, 23, 20, 18, 21, 25 а.о.) [12]. Тогда как длины его TPD сильно отличаются (12, 8, 13, 0, 17, 11, 5, 145 а.о.), особенно четвертого (с длиной, равной нулю, т.е. вовсе отсутствующего) и последнего – концевого. Кроме того, между пятым и шестым ТМУ расположен необычный участок длиной 19 а.о., примерно, как ТМУ, в котором гидрофобные и гидрофильные аминокислоты представлены почти поровну.

Чтобы упростить расчет Фурье-спектра символьной последовательности белка CNG и повысить достоверность результата, мы сначала попробовали "отрезать" хвостовой TPD. В полученном для оставшейся части цепи (длиной $L = 210$ а.о.) спектре (рис. 2) наблюдается пик с самой большой интенсивностью $I(T) = 7.85$, периодом $T \approx 30$ и частотой $n = L/T = 7$, для которого по формуле (6) число альфа-спиралей $N_{\text{ТМУ}} = 6$, что соответствует известным данным о структуре белка CNG [12]. При расчете спектра Фурье (на рис. 2 не показан) для полной последовательности белка CNG (с длиной цепи $L = 355$ без обрезания) проявился пик с наибольшей интенсивностью $I(T) = 5.23$ и почти тем же, что и у обрезанной цепи, периодом $T = 29.6$, но другой частотой $n = 12$. Откуда следует значение числа $N_{\text{ТМУ}} = 11$. Это можно понять так: к шести периодам, укладывающимся на части цепи без "хвостового" домена добавили пять периодов, которые могли бы уложиться на "хвостовом" TPD ($145/30 = 4.8 \approx 5$).

Следовательно, даже если число альфа-спиралей $N_{\text{ТМУ}}$ в сложной цепи по формуле (6) определено неверно, то основной период T чередования ТМУ все равно вычисляется правильно. В последовательности белка CNG основной период $T \approx 30$,

примерно, равен общим длинам пар ТМУ вместе с предшествующим TPD, которые составляют (30, 31, 33, ..., 30, 30 а.о.). Отметим, что здесь пропущены особые участки цепи.

Белок GRM6 имеет особенно сложную структуру [12]. Она предположительно содержит сигнальный пептид длиной 24 а.о., очень длинный (561 а.о.) первый TPD, 7 ТМУ с близкими длинами (23, 21, 19, 21, 22, 23, 26 а.о.), между которыми находятся 6 TPD и в конце цепи еще один. Эти топологические домены тоже имеют разные длины (14, 11, 24, 30, 13, 13, 32 а.о.). Сначала для упрощения расчета спектра мы "отрезали" передний кусок длиной (585 а.о.), состоящий из сигнального пептида и первого TPD. Предварительная оценка среднего периода для остальной части длиной $L = 292$ а.о. дала $\langle T\rangle = L/7 = 41.7$. В полученном для нее спектре Фурье (рис. 2) очевидно, что пик с самой большой интенсивностью $I(T) = 5.84$ имеет тот же период $T = 41.7$ и частоту $n = L/T = 7$. Для него по формуле (6) получаем число $N_{\text{ТМУ}} = 6$, что оказалось на единицу меньше числа ТМУ в предполагаемой структуре. Можно сказать, что здесь метод Фурье предсказал число ТМУ с точностью до (минус) единицы.

Затем мы получили спектр для всей, необрезанной, цепи длиной $L = 877$ а.о. В спектре Фурье (на рис. 2 не показан) наблюдали ряд пиков с кратными частотами и слишком большим и потому маловероятным для ТМУ основным периодом $T > 60$, за что эти пики можно было сразу исключить из рассмотрения. Среди пиков, подходящих по периоду ($T \approx 40$) интересен пик достаточно большой интенсивности $I(T) = 2.5$ с периодом $T = 41.7$, тем же, что установлен в спектре Фурье, полученном для "обрезанной" цепи. Для этого периода частота $n = 21$, и по формуле (6) получили число $N_{\text{ТМУ}} = n - 1 = 20$, что значительно отличается от предполагаемого числа ТМУ в структуре, равного 7.

Следовательно, даже для всей длинной последовательности метод Фурье позволяет правильно определить, если не число ТМУ, то период их следования в трансмембранной части цепи.

Таким образом, данное исследование периодичности трансмембранных белков методом Фурье показало следующее. Если некоторые элементы, части последовательности повторяются практически периодически, то и период повторения T, и число $N_{\text{ТМУ}}$ в цепи определяются верно, как для белков bR, sR2, hR и Cx32. Если же периодичность есть лишь на части последовательности, а на другой ее части периодичности нет, то основной период T следования ТМУ в трансмембранной части цепи, как "обрезанной", так и "необрезанной", определяется по спектру Фурье верно, а вычисленное по его значению число ТМУ в "необрезанной" цепи может оказаться неверным.

Чтобы преодолеть этот недостаток, предлагаем использовать другой метод компьютерного анализа символьной последовательности белка, суть которого состоит в следующем.

КОМПЬЮТЕРНЫЙ АНАЛИЗ РАСПОЛОЖЕНИЯ ТМУ БЕЛКА МЕТОДОМ МНОГОКРАТНОГО УСРЕДНЕНИЯ ФУНКЦИИ ГИДРОФОБНОСТИ ВНУТРИ ПЕРЕМЕЩАЕМОГО ВДОЛЬ ЦЕПИ "ОКНА"

Поскольку ТМУ состоят в большей степени из гидрофобных аминокислот [13], понятно, что средняя по такому участку величина гидрофобности, задаваемая в последовательности белка некоторой функцией $f(k) = H_N[i(k)]$ от номера k аминокислоты в цепи, должна быть выше, чем в примыкающих к нему с обеих сторон гидрофильных TPD. Причем, это локальное свойство совершенно не зависит от периодичности расположения характерных ТМУ и TPD в последовательности аминокислот белка. Здесь $i(k) = 1, 2, ..., 20$ – это номер той из известного ряда 20 аминокислот (табл. 1), которая стоит в последовательности белка на месте с номером k.

Впервые эту идею реализовали авторы работы [13], которые использовали усреднение функции $f(k)$ в пределах перемещаемого вдоль последовательности аминокислот сегмента – "окна" – шириной $d = 5, 7, 9, 11, 13$ а.о. Результат усреднения присваивали члену новой числовой последовательности $f_1(k)$ с номером k, соответствующим текущему положению средней точки сегмента.

Используемая в этом методе шкала гидрофобности $H_N(i)$ может быть задана многими способами (табл. 1) в зависимости от характеризующей это свойство и физически измеряемой величины [13–17]. В работах [13–15] в качестве меры гидрофобности использовали изменение величины свободной энергии боковых групп аминокислот при их переносе из гидрофобной среды в воду. В работах [16, 18] мера (шкала) гидрофобности аминокислот определена в виде функции $H_4(i) = 1 - \langle A\rangle/A^0$ (табл. 1) по величинам площади $A^0(i)$ поверхности аминокислотного остатка, доступной для растворителя в стандартном состоянии, и средней площади $\langle A(i)\rangle$, доступной в свернутом состоянии белка. В работе [16] установлена корреляция между величиной свободной энергии и площадью поверхности доступной для растворителя.

В данной работе сначала использовали ту же простейшую функцию гидрофобности (1), что и в вышеописанном методе преобразования Фурье. Процедура усреднения функции $f(k)$ по шкале $H_N(i)$ отлична от той, которую использовали в ра-

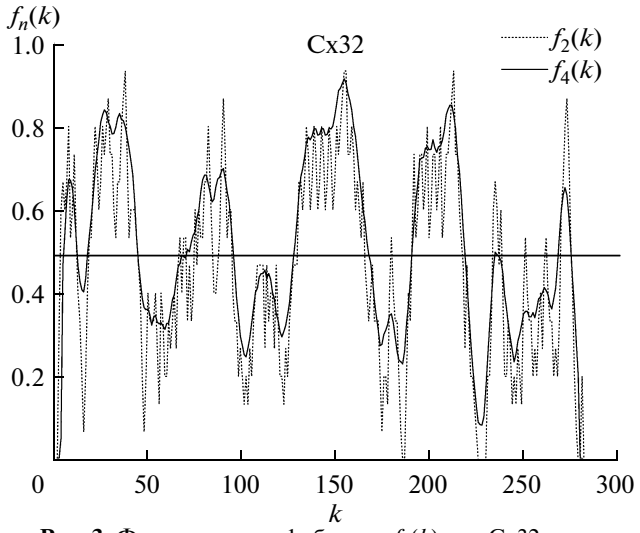

Рис. 3. Функции гидрофобности $f_n(k)$ для Cx32 после усреднения при $n = 2$ и $n = 4$ по шкале $H_1(i)$ в табл.1.

боте [13]. В данном исследовании усреднение проводилось не один, а несколько раз по алгоритму

$$f_n(k) = \frac{1}{2n+1}\sum_{-n}^{+n} f_{n-1}(k), \qquad (7)$$

$$n = 1, 2, \ldots, 5, \quad f_0(k) = f(k),$$

где каждое новое усреднение производили над предыдущей функцией по окну большей ширины $d = 2n + 1$: первое усреднение по трем элементам, второе – по пяти и т.д. Наилучший результат, на наш взгляд, достигался при $n = 4$, усреднении по окну шириной $d = 9$ а.о. (иногда при $n = 5$, $d = 11$ а.о.).

На рис. 3 представлены результаты усреднения функции гидрофобности для последовательности белка Cx32 по самой простой и грубой шкале $H_1(i)$ в табл. 1. Очевидно, что если исключить из рассмотрения два узких пика на краях графика функции $f_4(k)$, то оставшиеся 4 широких пика, превышающие средний уровень $u = \langle f_n(k)\rangle = 0.49$, как раз будут соответствовать четырем ТМУ в предполагаемой структуре этого белка. В графике функции $f_2(k)$ второй ТМУ еще не разрешен, на его месте и в других местах присутствуют несколько узких пиков.

В случае последовательности белка bR обработка с тем же набором гидрофобных аминокислот не позволила правильно определить на среднем уровне $\langle f_4(k)\rangle$ все 7 его трансмембранных участков: пара последних слилась в один, другие два (3 и 5) оказались разделенными на два участка. Чтобы обработать описанным методом функции $f(k)$ этого белка, больше подошла предложенная в работах [16, 18] гидрофобная группа, которая включала в себя аминокислоты: C, F, I, L, M, V, W. На рис. 4 представлены результаты, полученные для bR-последовательности с использованием этой гидрофобной группы и шкалы гидрофобности $H_2(i)$ в табл. 1. Очевидно, в отличие

Таблица 1. Шкалы гидрофобности $H_N(i)$

i	Код	Название	$H_1(i)$, [9]	$H_2(i)$, [16]	$H_3(i)$, [16]	$H_4(i)$, [16, 18]	$H_5(i)$, [17]	$H_6(i)$, [15]
1	A	Alanine	1	0	0	0.74	0.62	1.60
2	C	Cysteine	0	1	1	0.91	0.29	2.00
3	D	Aspartic acid	0	0	–1	0.62	–0.90	–9.20
4	E	Glutamic acid	0	0	–1	0.62	–0.74	–8.20
5	F	Phenylalanine	1	1	1	0.88	1.19	3.70
6	G	Glycine	1	0	–1	0.72	0.48	1.00
7	H	Histidine	0	0	0	0.78	–0.40	–3.00
8	I	Isoleucine	1	1	1	0.88	1.38	3.10
9	K	Lysine	0	0	–1	0.52	–1.50	–8.80
10	L	Leucine	1	1	1	0.85	1.06	2.80
11	M	Methionine	0	1	1	0.85	0.64	3.40
12	N	Asparagine	0	0	–1	0.63	–0.78	–4.80
13	P	Proline	0	0	–1	0.64	0.12	–0.20
14	Q	Glutam	0	0	–1	0.62	–0.85	–4.10
15	R	Arginine	0	0	–1	0.64	–2.53	–12.3
16	S	Serine	0	0	–1	0.66	–0.18	0.60
17	T	Threonine	0	0	–1	0.70	–0.05	1.20
18	V	Valine	1	1	1	0.86	1.08	2.60
19	W	Tryptophan	1	1	1	0.85	0.81	1.90
20	Y	Tyrosine	1	0	0	0.76	0.26	–0.70

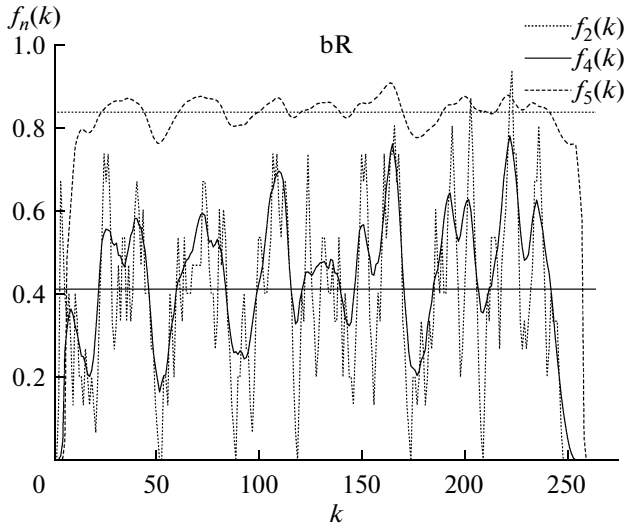

Рис. 4. Функции гидрофобности $f_n(k)$ для bR после усреднения при $n = 2$ и $n = 4$ по шкале $H_2(i)$ и усреднения при $n = 5$ по шкале $H_4(i)$ в табл. 1.

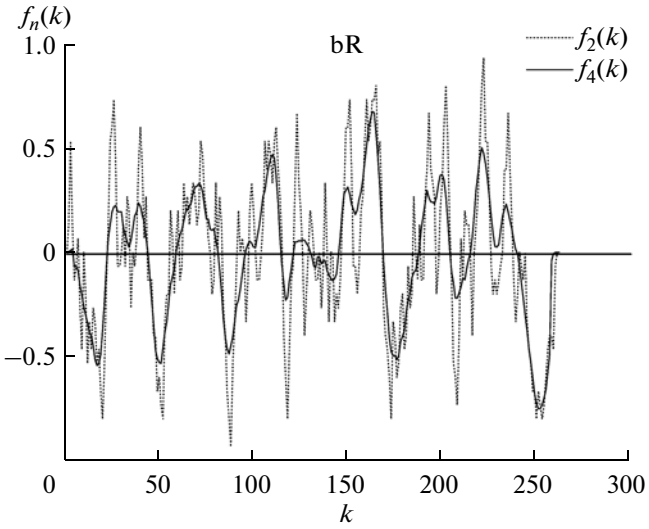

Рис. 5. Функции гидрофобности $f_n(k)$ для bR после усреднения при $n = 2$ и $n = 4$ по шкале $H_3(i)$ в табл. 1.

от функции $f_2(k)$ на графике, соответствующем $f_4(k)$, выше среднего уровня $u = 0.41$ разрешены все известные для структуры bR 7 ТМУ [12].

Затем функцию $f(k)$ несколько усложнили. В соответствии с работой [18] использовали деление 20 аминокислот по величине их гидрофобности на 3 группы: гидрофобные – C, F, I, L, M, V, W – всего 7; гидрофильные – D, E, G, K, N, P, Q, R, S, T – всего 10; нейтральные – A, H, Y – всего 3. Гидрофильным аминокислотам в шкале $H_N(i)$ присвоили значение –1, гидрофобным – значение +1, нейтральным – значение 0. Так получили грубую шкалу $H_3(i)$ в табл. 1. На рис. 5 приведены графики рассчитанных с использованием этой шкалы функций $f_2(k)$ и $f_4(k)$ для bR. На графике, соответствующем $f_4(k)$, выше среднего уровня $u = –0.01$ определены все 7 ТМУ [12].

На рис. 4 представлена также функция $f_5(k)$ после усреднения по более точной шкале гидрофобности $H_4(i)$, приведенной в работах [16, 18]. Очевидно, что на среднем уровне $\langle f(k)\rangle = 0.76$ не разрешен ни один ТМУ, а выше некоторого уровня $f_5(k) = 0.83$, граничного для гидрофобной и нейтральной групп аминокислот, разрешены все те же 7 ТМУ.

По пересечению графика функции $f_n(k)$ с прямой среднего уровня $u = \langle f(k)\rangle$ определяли границы ТМУ различных белков, которые приведены в табл. 2. В той же таблице для сравнения приведены границы ТМУ из работы [12]. С учетом погрешностей $\Delta k \approx d/2 \approx 5$ определения границ ТМУ можно признать хорошим согласие результатов расчета положения границ ТМУ и данных работы [12].

Отметим, что обработка с многократным (4–5 раз) усреднением функции гидрофобности $f_n(k)$ при использовании разных шкал, грубых $H_1(i)$–$H_3(i)$ или более точных $H_4(i)$–$H_6(i)$, дает для границ ТМУ отличающиеся значения. Иногда эти различия незначительны, а иногда существенны. Так, для белка GRM6 применение шкал $H_3(i)$ и $H_4(i)$ выявило только 6 ТМУ вместо 7, как и преобразование Фурье. А применение шкал $H_5(i)$ и $H_6(i)$ определило все 7 ТМУ, предполагаемых в его структуре. Напротив, для белка bR расчет по шкалам $H_2(i)$–$H_4(i)$ показал все 7 ТМУ, а по шкале $H_5(i)$ – только 6 ТМУ.

В работах [19–21] метод “скользящего окна” с однократным усреднением значений плотности окружения, или гидрофобности, или заряда аминокислотных остатков в последовательностях глобулярных и “нативно-развернутых” белков применяли в программе FoldAmyloid [22] для предсказания агрегационных и амилоидогенных участков. Эти участки так же, как и ТМУ в мембранных белках, обогащены гидрофобными остатками. Поэтому указанная программа может быть использована для предсказания трансмембранных α-спиралей.

Попытка применить программу FoldAmyloid с окном шириной 11 а.о. к вышеуказанным шести белкам, исследованным в данной работе, оказалось отчасти удачной. В случае четырех белков (sR2, hR, Cx32, GRM6) границы ТМУ определены практически правильно, почти также как в табл. 2. Результаты хуже для двух белков: у bR четвертый ТМУ участок вообще не определен, а первый и седьмой разделены на 2 части каждый, у CNG первый ТМУ не обнаружен.

В целом, границы ТМУ в указанных шести белках с использованием программы FoldAmyloid определены, на наш взгляд, не лучше, чем предлагаемым методом многократного усреднения гидрофобности по каждый раз изменяющейся ширине скользящего окна. Заметим, что в данной работе установлена как оптимальная ширина окна 9–11 а.о. Поиск ТМУ приводит к гораздо худшим результатам в обоих сравниваемых методах при ширине окна 5 а.о., рекомендованной для программы FoldAmyloid в работе [21].

При дополнительном тестировании и сравнении между собой методы, алгоритмы которых изложены в данной работе, кроме шести вышеописанных белков, применили еще к 18 белкам, вторичная структура которых, экспериментально определенная или предсказанная другими методами, представлена на сайте [12]. Это белки из интересовавших нас и рассмотренных выше групп bR, sR, hR и connexin с кодами из следующего ряда: O93740, P42197, P71411, P25964, P33743, O93743, P16102, Q48314, O93741, O93742, P33742, P28230, O75712, Q6PEY0, P28235, P28234, Q96KN9, P23242.

Результаты тестирования при исследовании в общей сложности 24 белков показали, что с помощью метода с использованием преобразования Фурье число трансмембранных участков $N_{ТМУ}$ правильно определено по формуле (6) для 14 белков, а в 10 случаях – на 1–2 ТМУ больше (или меньше), чем иными методами.

Тестирование другого предложенного в данной работе и выраженного формулой (7) метода (каскадного усреднения функции гидрофобности в пределах скользящего окна) показало, что из 24 исследованных этим новым методом белков число трансмембранных участков правильно определено для 19 белков, а в пяти случаях выявлено на 1 ТМУ больше, чем иными методами. Кроме того, для 19 белков, с верно определенным числом ТМУ, разногласие с результатами других методов в положении границ ТМУ не превосходит ошибок их определения ($\Delta k < 6$) у 182 границ из 224, т.е. в 81% случаев.

ПРИМЕНИМОСТЬ ПРЕДЛОЖЕННЫХ КОМПЬЮТЕРНЫХ МЕТОДОВ ПРЕДСКАЗАНИЯ СТРУКТУРЫ МЕМБРАННЫХ БЕЛКОВ ПО ИХ АМИНОКИСЛОТНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ

Повторяемость расположения аминокислот гидрофобной (и/или гидрофильной) группы в белковых последовательностях – один из признаков вторичной и третичной структуры белка. Это верно для фибриллярных, глобулярных и мембранных белков.

Если повторяемость периодическая, то эту периодичность можно выявить, применяя известный метод преобразования Фурье к цифровому образу символьной последовательности аминокислот белка. В данной работе этот метод применили к 24 трансмембранным белкам. Число ТМУ, определяемое по формуле (6) через основной период повторяемости аминокислот гидрофобной группы в белковой последовательности, правильно выявлено для 58% исследованных белков. Невысокое значение этой доли обусловлено отклонением от периодичности в последовательностях значительной части белков. Оценить местоположение и протяженность гидрофобных участков

Таблица 2. Сравнение с известными данными сайта [12] для границ ТМУ, вычисленных при обработке функций гидрофобности $f_n(k)$ при $n = 4$ и $n = 5$** по шкалам $H_3(i)$ и $H_5(i)$ для шести различных мембранных белков

Код белка	Источник данных	Номер и границы трансмембранных и интрамембранного* участков						
		1	2	3	4	5	6	7
bR	[12]	24–42	57–75	92–109	121–140	148–167	186–204	217–236
	$H_3(i)$	23–44	59–81	96–114	122–130	146–168	188–204	216–239
sR2	[12]	4–25	34–55	70–91	95–117	122–149	154–181	190–222
	$H_3(i)$	5–23	40–62	74–91	99–116	126–142	165–182	192–215
hR	[12]	31–56	63–86	121–142	146–169	173–195	208–231	241–269
	$H_3(i)$	39–55	69–77	132–140	154–164	175–197	215–240	251–265
Cx32	[12]	23–45	76–95	131–153	192–214			
	$H_3(i)$	22–41	68–96	130–169	190–216			
GRM6	[12]	586–608	623–643	655–673	698–718	749–770	784–806	820–845
	$H_5(i)$**	591–606	625–649	655–666	692–719	748–769	784–808	818–845
CNG	[12]	13–30	39–61	75–94	95–112	130–150	162–180*	186–210
	$H_3(i)$	4–28	41–59	77–90	99–107	133–149	165–172	191–206

этим методом можно лишь очень грубо, с точностью до половины основного периода повторяемости гидрофобных аминокислот в последовательности белка.

Из двух рассмотренных в данной работе методов более предпочтительным оказался метод многократного (4–5 раз) усреднения функции гидрофобности $f(k)$ белка внутри "окна" шириной 9–11 а.о., перемещаемого вдоль аминокислотной последовательности. Число ТМУ правильно предсказано этим методом для 79% исследованных белков. Доля правильно предсказанных границ участков составляет 81%.

Шкала и функция гидрофобности могут быть заданы многими способами – известно более 30. Сравнение различных шкал и функций гидрофобности, выполненное в данной работе, показало, что определяемое с их использованием число и расположение трансмембранных участков часто практически одинаковы, даже для очень простых (грубых) шкал, например, $H_2(i)$ и $H_3(i)$ (см. табл. 1). Но иногда для данного белка одна из шкал может оказаться предпочтительнее из-за того, что она позволяет лучше определить близкорасположенные ТМУ.

Сравнение результатов, полученных в данной работе путем компьютерного анализа аминокислотных последовательностей 24 белков, с известными данными об их структуре показало применимость обоих предложенных методов, как по отдельности, так и совместно, для предсказания признаков неизвестной вторичной структуры мембранных белков и обусловленных этими признаками функциональных свойств этих белков.

Ценность обоих методов – простота, быстрота и экономичность применения.

СПИСОК ЛИТЕРАТУРЫ

1. Kendrew J., Bodo G., Dintzis H., Parrish R., Wyckoff H., Phillips D. 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature*. **181**, 662–666.
2. Deisenhofer J., Epp O., Miki K., Huber R., Michel H. 1985. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature*. **318**, 618–624.
3. Коротков Е.В., Короткова М.А., Френкель Ф.Е., Кудряшов Н.А. 2003. Информационная концепция поиска периодичности в символьных последовательностях. *Молекуляр. биология*. **37**, 436–451.
4. McLachlan A.D., Stewart M. 1976. The 14-fold periodicity in α-tropomyosin and the interaction with actin. *J. Mol. Biol*. **103**, 271–298.
5. Макеев В.Ю., Туманян В.Г. 1994. О связи методов внутренних гомологий, корреляционных функций и Фурье-анализа при поиске периодичностей в первичных структурах биополимеров. *Биофизика*. **39**, 294–297.
6. Лобзин В.В., Чечеткин В.Р. 2000. Порядок и корреляции в геномных последовательностях ДНК. Спектральный подход. *Усп. физ. наук*. **170**, 57–81.
7. Симакова М.Н., Симаков Н.Н. 2005. Исследование периодичности расположения аминокислот в фибриллярных белках бактериофага Т4. *Молекуляр. биология*. **39**, 321–329.
8. Чупров-Неточин Р.Н., Файзулина Н.М., Сыкилинда Н.Н., Симакова М.Н., Месянжинов В.В., Мирошников К.А. 2010. Бета-спиральный домен бактериофага Т4 управляет укладкой фрагмента длинных фибрилл в составе химерного белка. *Биоорган. химия*. **36**, 193–199.
9. Остерман Л.А. 2002. *Методы исследования белков и нуклеиновых кислот*. М.: МЦНМО.
10. Корн Г., Корн Т. 1977. *Справочник по математике для научных работников и инженеров*. М.: Наука.
11. Кудрявцев Л.Д. 1973. *Математический анализ, т. 2*. М.: Высшая школа.
12. Сайт: http://www.uniprot.org
13. Kyte J., Doolittle R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol*. **157**,105–132.
14. Frommel C. 1984.The apolar surface area of amino acids and its empirical correlation with hydrophobic free energy. *J. Theor. Biol*. **111**, 247–260.
15. Engelman D.M, Steitz T.A., Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem*. **15**, 321–353.
16. Rose G.D., Geselowitz A.R., Lesser G.J., Lee R.H., Zehfus M.H. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science*. **229**, 834–838.
17. Eisenberg D., Schwarz E., Komaromy M., Wall R. 1984. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol*. **179**, 125–142.
18. Lesser G.J., Lee R.H., Zehfus M.H., Rose G.D. 1987. Hydrophobic interactions in proteins. In: *Protein Engineering*. Eds Oxender D.L., Fox C.F., N.Y.: Alan R. Liss, pp. 175–179.
19. Галзитская О.В., Гарбузинский С.А., Лобанов М.Ю. 2006. Предсказание нативно-развернутых участков белковой цепи. *Молекуляр. биология*. **40**, 341–348.
20. Галзитская О.В., Гарбузинский С.А., Лобанов М.Ю. 2006. Поиск амилоидогенных участков белковой цепи. *Молекуляр. биология*. **40**, 910–918.
21. Garbuzynskiy S.O., Lobanov M.Yu., Galzitskaya O.V. 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*. **26**, 326–332.
22. Сайт: http://bioinfo.protres.ru/fold-amyloid/oga.cgi