БИОФИЗИКА, 2020, том 65, № 6, с. 1065–1071

МОЛЕКУЛЯРНАЯ БИОФИЗИКА

УДК 577.3

УРОВНИ ИЕРАРХИЧЕСКОЙ ОРГАНИЗАЦИИ БЕЛКОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ. АНАЛИЗ ЭНТРОПИЙНЫХ ХАРАКТЕРИСТИК

© 2020 г. А.Н. Некрасов*, Ю.П. Козмин*, С.В. Козырев, Н.Г. Есипова***, Р.Х. Зиганшин*, А.А. Анашкина*****

**Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН, 117997, Москва, ГСП-7, ул. Миклухо-Маклая, 16/10*
E-mail: alexei_nekrasov@mail.ru

***Математический институт им. В.А. Стеклова РАН, 119991, Москва, ул. Губкина, 8*

****Институт молекулярной биологии им. В.А. Энгельгардта РАН, 119991, Москва, ул. Вавилова, 32*
E-mail: anastasya.anashkina@gmail.com

Поступила в редакцию 18.03.2020 г.
После доработки 18.03.2020 г.
Принята к публикации 15.09.2020 г.

Исследованы 24647 негомологичных белковых последовательностей. Для каждой из последовательностей построен профиль встречаемости пептапептидов и в каждом из этих профилей специальным математическим методом выявлены иерархически организованные элементы различных размеров. Исследованы корреляции между этими иерархическими элементами и показано, что в исследованном наборе белковых последовательностей существуют 11 уровней организации белков с элементами размером от 7 до 56 аминокислотных остатков. Высказано предположение, что выявленные уровни организации соответствуют различным по топологии элементам супер-вторичной структуры.

Ключевые слова: белковые последовательности, иерархическая организация, энтропия.

DOI: 10.31857/S0006302920060046

Наблюдаемое многообразие свойств белков и их функциональные характеристики обеспечиваются пространственным расположением составляющих эти белки аминокислотных остатков. Обладая различными физико-химическими свойствами, аминокислотные остатки обеспечивают существование ряда уникальных свойств белков. Одним из этих свойств является способность к самоорганизации – фолдингу. При этом для нативного фолдинга большинству белков достаточно «правильного расположения» аминокислотных остатков в первичной структуре белка, т.е. пространственная структура таких белков полностью определяется последовательностью аминокислотных остатков.

Наличие иерархической организации в пространственной структуре белков было выявлено разными авторами с привлечением различных подходов, таких как расчет локальной плотности упаковки атомов в структуре [1–3] и энергии взаимодействия внутри белковой глобулы [4]. Выявляемые иерархические элементы белков рассматривались как элементы фолдинга белков на разных стадиях [5–7]. В процессе анализа баз данных последовательностей белков был найден ряд повторяющихся элементов (паттернов) [8–10]. Такие паттерны использовались для предсказания пространственной организации и функции белков [11–13]. Однако в отличие от структурных элементов, полученных при анализе пространственных структур, паттерны, выявляемые в первичных структурах белков, не имеют установленной иерархической организации.

Ранее в работе [14] мы показали, что в белковых последовательностях наиболее низкий уровень энтропии Шеннона[15] наблюдается внутри блоков из пяти аминокислотных остатков. Подход, при котором первичная структура белков рассматривается как система из перекрывающихся или последовательно расположенных блоков из пяти аминокислотных остатков, был ранее использован для создания структурного алфавита белков [16], исследования топологически устойчивых элементов низшего уровня организации пространственной структуры белков [17], описания фолдинга белковых молекул [18, 19]. Взяв за единицу последовательности белка фрагмент из

Сокращение: ЭЛИС – элементы информационной структуры.

пяти аминокислотных остатков, мы предложили метод выявления иерархически организованных структур в белковых последовательностях (метод АНИС) [20, 21]. Этот метод выделяет в последовательности белков древовидные иерархические структуры (графы). На ряде примеров было показано, что отдельно стоящие графы соответствуют структурным доменам [22]. Впоследствии выявляемые иерархические элементы были использованы нами для дизайна белковых молекул [23–26] и для исследования механизмов функционирования белков [22, 27, 28]. В работах [23–26] экспериментально было показано, что удаление отдельно стоящих графов из нативной последовательности белка приводит к минимальным нарушениям фолдинга рекомбинантного белка, а в работах [22, 27, 28] с помощью анализа таких иерархических элементов в структуре белка были предложены механистические модели функционирования белковых молекулярных машин.

Применение метода АНИС к большому числу природных белковых последовательностей привело нас к наблюдению, что существуют характерные размеры фрагментов последовательностей, при которых древовидные графы разделяются на более мелкие иерархические элементы. Такие фрагменты меньше структурных доменов, однако больше структурного алфавита, предложенного в работе [16]. Настоящая работа посвящена анализу размеров таких выявляемых иерархически организованных элементов белковой последовательности.

ФРАГМЕНТЫ И ИНФОРМАЦИОННАЯ СТРУКТУРА БЕЛКА

Длина структурного элемента, оптимальная для описания белковых последовательностей, может быть оценена следующим образом. Пусть x и y – суть дискретные случайные величины, принимающие конечное множество значений. В рассматриваемом случае x и y – случайно выбираемые аминокислотные остатки, расположенные в белковой цепи на некотором расстоянии друг от друга. Тогда эти остатки можно охарактеризовать вероятностями $p(x)$, $p(y)$ выпадения тех или иных аминокислот, совместным распределением вероятностей $p(x, y)$ для пар аминокислот (зависящим от расстояния между x и y вдоль цепи) и взаимной информацией (тоже зависящей от расстояния между x, y):

$$I(x,y) = \sum_{x.y=1}^{20} p(x.y) \log \frac{p(x,y)}{p(x)p(y)}. \quad (1)$$

В работе [14] было показано, что взаимная информация довольно быстро уменьшается при увеличении расстояния больше пяти между аминокислотными остатками. Таким образом, рассмотрение фрагментов длины пять является оптимальным для учета большей части корреляций, существующих в последовательностях белков.

Рассмотрим фрагменты длиной в пять остатков (мы рассматриваем всевозможные перекрывающиеся фрагменты, т. е. соседние фрагменты пересекаются по четырем остаткам). Каждому фрагменту I, рассматриваемому как последовательность остатков длиной пять, мы сопоставим частоту $\phi(I)$ его встречаемости в базе данных негомологичных белковых последовательностей. Для получения более надежной статистики мы рассматриваем усредненную частоту

$$\Phi(I) = \sum_{J:d(I,J)\le\delta} \phi(J), \quad (2)$$

где усреднение идет по последовательностям J, удаленным от I на расстояние не более δ. В наших работах [20, 21] было использовано расстояние δ в метрике Хэмминга, равное единице, т. е. усреднение проводили по последовательностям, отличающимся от данной заменой не более одного аминокислотного остатка.

Далее последовательности белка $I = i_1 ... i_N$ сопоставляется последовательность фрагментов I_j длиной пять, нумеруемых центральными остатками в этих фрагментах, т. е. $j = 3, ..., N–2$. Проанализировав встречаемость пентапептидов в базе данных белковых последовательностей NRDB, получим значение встречаемости j-го пентапептида в белке I. При анализе встречаемости эквивалентными считаются пентапептиды, отличающиеся по расстоянию Хэмминга на единицу, т. е. встречаемостью пентапептида мы называем сумму встречаемости конкретного пентапептида и всех пентапептидов, отличающихся на один аминокислотный остаток. На исследуемой белковой последовательности мы строим функцию, в которой каждый аминокислотный остаток характеризуется суммой $f_I(j) = \Phi(I_j)$ встречаемостей всех пентапептидов, в которые он вошел.

Рассмотрим гауссову функцию на вещественной оси, предложенную в работах [20, 21]:

$$g(x,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad (3)$$

и введем сглаженное распределение как свертку гауссовой функции с частотами встречаемости фрагментов в белке I:

$$F_I(x) = \sum_j f_I(j) g(x - j, \sigma). \quad (4)$$

Далее будем вписывать гауссовы функции с шириной $2y > 0$ и высотой h в график функции $F_I(x)$, т. е. рассмотрим функцию

$$H_I(x,y) = \max h : \min[F_I(z) - he^{-\frac{(z-x)^2}{2y^2}}] \ge 0, \quad (5)$$

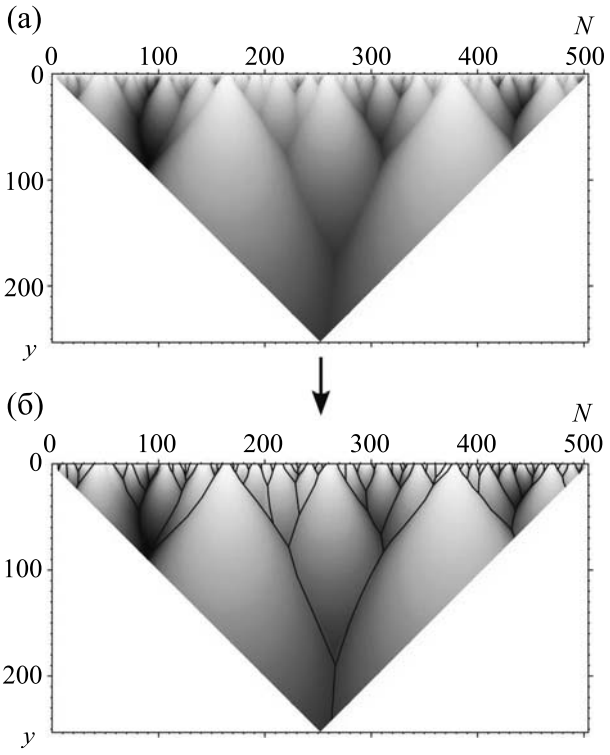

Рис. 1. (а) – Информационной структура последовательности белка 1GWE.PDB (цепь A), которая получена с помощью метода АНИС. Обозначения осей: N – номер аминокислотного остатка в первичной структуре, y – параметр из уравнения (5). (б) – Древовидный граф, построенный по локальным максимумам функции $H_I(x, y)$, в которой выделены ЭЛИС.

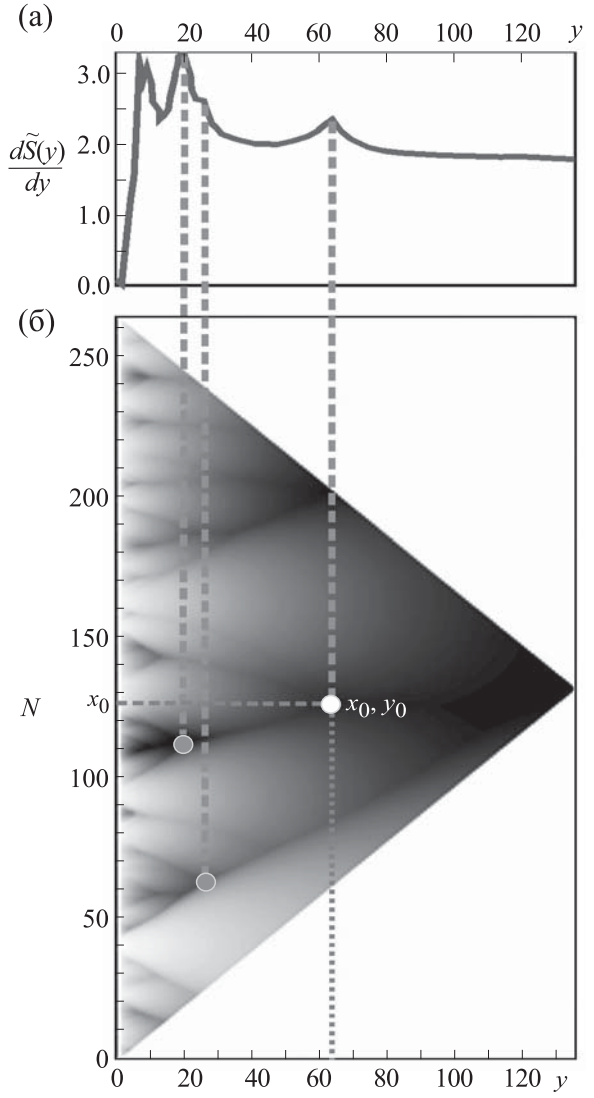

Рис. 2. Разностная производная $\tilde{S}'_I(y)$ по y. (а) – График первой производной от разности между реальной и модельными значениями энтропии. (б) – Иерархическая структура белка, на которой отмечены три точки слияния ветвей. Точка слияния ветвей x_0y_0 упомянута в тексте статьи и отмечена белым кружком. Две другие точки слияния ветвей отмечены серыми кружками. Пунктирными линиями показана взаимосвязь между точками слияния в иерархической структуре белка с пиками на графике первой производной.

измеряющую, насколько высокую гауссову функцию с центром в x и шириной $2y$ можно вписать в график функции $F_I(x)$.

Носитель такой функции будет определен на равнобедренном треугольнике на координатной плоскости (x, y) с основанием в виде отрезка $[0, N]$ на оси абсцисс и высотой $N/2$. Пример функции $H_I(x, y)$ на своем носителе показан на рис. 1а. На локальных максимумах этой функции (рис. 1б) можно построить древовидный граф. Ветви этого графа естественно сопоставить промежуточным элементам иерархической организации белка (ЭЛИС, ЭЛементы Информационной Структуры [20]). Полученный древовидный граф описывает иерархическую организацию последовательности белка (рис. 1).

Любой иерархический элемент можно охарактеризовать его положением в последовательности белка и количеством слившихся ветвей, его формирующих (рангом). Рассмотрим один из иерархических элементов графика $H_I(x, y)$ с ветвлением в точке (x_0, y_0) (т.е. из этой точки вниз отходят несколько «ветвей») (рис. 2б). Точке (x_0, y_0) отвечает отрезок белка I, содержащий аминокислотные остатки с номерами, попадающими в отрезок оси абсцисс с центром в точке x_0 и шириной $2y_0$.

ИЕРАРХИЯ СТРУКТУР В БЕЛКЕ

Рассмотрим набор из $M = 24647$ белков. Для каждого из белков из этого набора вычисляем функцию $H_I(x, y)$ (уравнение (5)). Поскольку рассматриваемые белки имеют разную длину, разме-

ры носителей таких функций для разных белков будут различаться.

Рассмотрим набор значений $y = 1, ..., L$, используемых для построения функции $H_I(x, y)$. Величину L выбирали равной от 50 до 150.

Для белка I и фиксированного значения y рассмотрим функцию $H_I(x, y)$ как функцию распределения вероятностей, т.е. нормируем функцию так, чтобы

$$\int H_I(x, y)dx = 1. \quad (6)$$

Для такого нормированного распределения вероятностей вычислим его энтропию Шеннона:

$$S_I(y) = -\int H_I(x, y)\log H_I(x, y)dx. \quad (7)$$

Аппроксимируем равномерное распределение на отрезке [0, 1] равномерным распределением на разбиении отрезка на n одинаковых частей (вероятность каждой такой части будет равняться $1/n$) и получим для энтропии следующую зависимость от разбиения:

$$-\sum_{i=1}^{n}\frac{1}{n}\log\frac{1}{n} = -\log\frac{1}{n} = \log n. \quad (8)$$

Вычислим регуляризованную энтропию следующим образом. Функция $H_I(x, y)$ имеет носителем отрезок длиной $N - 2y$ (где N есть длина белка I). Разделим такой отрезок на n одинаковых подотрезков, тогда энтропия Шеннона для равномерного распределения на таком разбиении будет равна $\log n$ (как показано выше). Вычислим энтропию $S_I(y)$ для нормированного распределения $H_I(x, y)$ *на разбиении* носителя $H_I(x, y)$ на n одинаковых подотрезков и вычтем для регуляризации энтропию для равномерного распределения:

$$\tilde{S}_I(y) = S_I(y) - \log n. \quad (9)$$

Естественно предположить, что такая функция будет испытывать скачки в точках ветвления дерева, когда происходит разделение ветвей на ЭЛИС меньшего ранга. Вычислим для $\tilde{S}'_I(y)$ разностную производную по y, максимумы такой функции должны отвечать точкам ветвления. На рис. 2б приведен пример иерархической информационной структуры белка и рассчитанной вышеописанным способом разностной производной (рис. 2а).

Далее рассмотрим для каждого $y = 1, ..., L$ вектор V_y, содержащий разностные производные $\tilde{S}'_I(y)$, матричные элементы вектора нумеруются белками I (в численном эксперименте мы рассматриваем, как было сказано выше, 24647 белков). Рассмотрим корреляцию таких векторов для разных y:

$$r_{yy'} = \frac{\langle V_y, V_{y'}\rangle}{\sqrt{\langle V_y, V_{y'}\rangle\langle V_{y'}, V_{y'}\rangle}},\ \langle A, B\rangle = \sum_I A_I B_I. \quad (10)$$

Такая матрица корреляций ($r_{yy'}$) содержит данные о корреляциях между ЭЛИС различных размеров (y и y') в иерархической структуре белков. Эта матрица отражает совпадения размеров фрагментов белковых последовательностей, на которых метод АНИС выявил точки ветвления иерархических структур. Матрица корреляций ($r_{yy'}$) была построена для 24647 иерархических структур белковых последовательностей, размер которых лежал в интервале от 50 до 400 аминокислотных остатков из базы NRDB90 [29]. Значения элементов матрицы ($r_{yy'}$) лежат в интервале от 0.01 до 1.0. На рис. 3 представлены изображения элементов матрицы ($r_{yy'}$) (см. формулу (10)).

Если у двух исследуемых белков в иерархической структуре точки ветвления находятся на близких значениях полуширины функции сглаживания, то для матрицы $r_{yy'}$ в придиагональной области возникают утолщения (рис. 3). Такие утолщения возникают при разных дискретных значениях полуширины функции сглаживания и в пределе формируют серию дискретных квадратных областей, разделенных между собой. Можно утверждать, что таким способом удалось выявить новые уровни организации белковых молекул, меньшие, чем структурные домены белков.

На рис. 3а–д видно, как при уменьшении величины значения критерия фильтрации формируются придиагональные области и связи между придиагональными областями. На диагонали наблюдается самый высокий уровень взаимной корреляции первых производных (на рис. 3а–д они отмечены черным цветом). При снижении критерия фильтрации от 0.9 до 0.01 видно, что квадратные области на диагонали формируются и принимают максимальные размеры. Эти области отмечены на рисунке квадратами с границами серого цвета. При дальнейшем снижении величины критерия фильтрации увеличения размеров и количества придиагональных областей не происходит. Если учесть, что были исследованы 24647 негомологичных последовательностей с размером от 50 до 400 аминокислотных остатков, то можно говорить, что полученная характеристика является общей новой уникальной характеристикой белковых последовательностей.

База NRDB90 [29] включает очень разнородные последовательности, т.е. последовательности белков всех известных структурных типов и функций. В данной работе предпринята попытка получить структурные характеристики для всей совокупности этих разнородных данных. Вследствие этого для получения устойчивых результа-

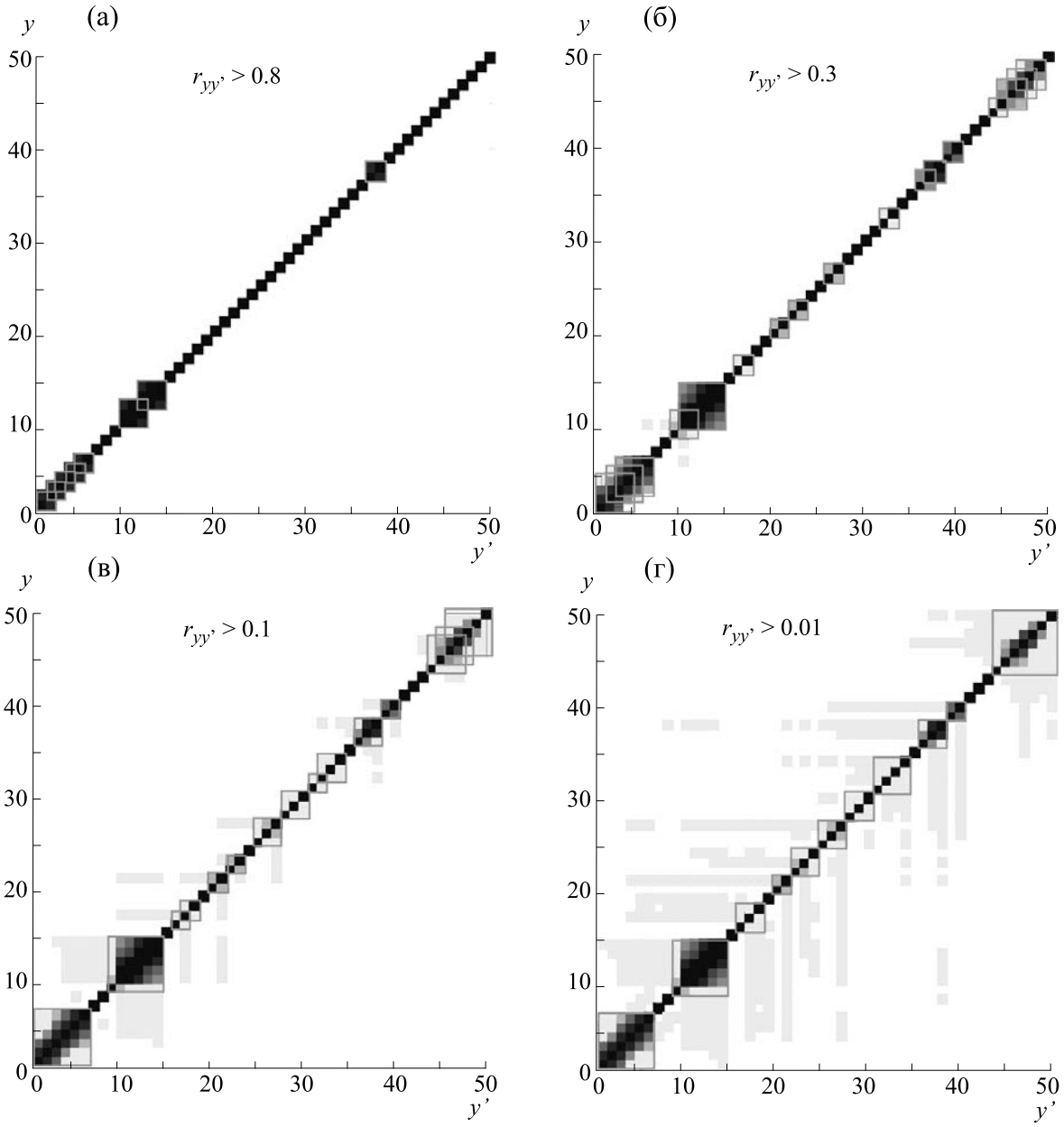

Рис. 3. Элементы $r_{yy'}$ матрицы взаимной корреляции векторов V_y, содержащих разностные производные $\tilde{S'}_1(y)$ (формула (10)) при различных значениях критериев фильтрации. Данные получены для 24647 негомологичных белковых последовательностей размером от 50 до 400 аминокислотных остатков из базы NRDB90 [29].

тов потребовалось использовать столь маленькие значения критерия фильтрации.

В таблице приведены размеры придиагональных областей, размеры соответствующих им элементов структурной организации белков.

Таким образом, удалось выявить одиннадцать уровней организации белковых последовательностей. Структуры с подобными характерными размерами ранее рассматривались как элементы супер-вторичной структуры. Применение описанного метода к белкам с известной пространственной структурой должно позволить выявить новые топологии структурно-устойчивых элементов.

ЗАКЛЮЧЕНИЕ

В работе дано описание метода исследования информационной структуры белковых последовательностей (метод АНИС), который позволяет выявить их иерархическую организацию. Приме-

Уровни организации элементов последовательности белков

Имя уровня структурной организации	Диапазон значений полуширины сглаживающей функции	Диапазон длины соответствующей последовательности белка, а.к.о.
X1	1–6	7–12
X2	9–14	15–20
X3	16–18	22–24
X4	20–21	26–27
X5	22–24	28–30
X6	25–27	31–33
X7	28–30	34–36
X8	31–34	37–40
X9	36–38	42–44
X10	39–40	45–46
X11	44–50	50–56

нение метода АНИС для исследования 24647 негомологичных белковых последовательностей размером от 50 до 400 аминокислотных остатков из базы NRDB90 [29] позволило выявить одиннадцать уровней организации белковых последовательностей.

В дальнейшем планируется изучение фрагментов пространственной структуры белков, соответствующих элементам разных уровней иерархии, размеры которых соответствуют выявленным в этой работе интервалам. Есть предположение, что такие фрагменты будут включать как уже известные элементы супер-вторичной структуры, так и элементы пространственной организации белков, ранее не выделяемые как элементы супер-вторичной структуры. Мы предполагаем, что выявляемые фрагменты обладают повышенной структурной устойчивостью, способны к самосборке и могут являться ядрами сворачивания, начиная формирование нативной пространственной организации белков.

Полученные в этой статье результаты могут быть использованы для выявления элементов супер-вторичной структуры, из которых формируются структурные домены, исследования структурной организации доменов, для изучения молекулярной эволюции полипептидных цепей, для дизайна белков рекомбинантных белков и проектирования белков с новыми типами пространственной организации.

ФИНАНСИРОВАНИЕ РАБОТЫ

Работа выполнена при частичной поддержке Российского фонда фундаментальных исследований (грант 20-04-01085а).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

СОБЛЮДЕНИЕ ЭТИЧЕСКИХ СТАНДАРТОВ

Настоящая работа не содержит описания исследований с использованием людей и животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. G. M. Crippen, J. Mol. Biol. **126**, 315 (1978). DOI: 10.1016/0022-2836(78)90043-8
2. G. D. Rose, J. Mol. Biol. **134**, 447 (1979).
3. L. Holm and C. Sander, Proteins **19**, 256 (1994). DOI: 10.1002/prot.340190309
4. I. N. Berezovsky, V. G. Tumanyan, and N. G. Esipova, FEBS Lett. **418**, 43 (1997). DOI: 10.1016/S0014-5793(97)01346-X
5. A. M. Lesk and G. D. Rose, Proc. Natl. Acad. Sci. USA **78**, 4304 (1981).
6. D. B. Wetlaufer, Proc. Natl. Acad. Sci. USA **70**, 697 (1973).
7. M. J. Zaki, V. Nadimpally, D. Bardhan, and C. Bystroff, Bioinformatics **20** (Suppl. 1), i386 (2004). DOI: 10.1093/bioinformatics/bth935
8. P. Bork and E. V. Koonin, Curr. Opin. Struct. Biol. **6**, 366 (1996). DOI: 10.1016/s0959-440x(96)80057-1
9. T. P. Exarchos, C. Papaloukas, C. Lampros, and D. I. Fotiadis, J. Biomed. Inform. **41**, 165 (2008). DOI: 10.1016/j.jbi.2007.05.004
10. N. Hulo, A. Bairoch, V. Bulliard, et al., Nucl. Acids Res. **34**, D227 (2006). DOI: 10.1093/nar/gkj063
11. F. S. Domingues and T. Lengauer, Appl. Bioinformatics **2**, 3 (2003).
12. J. D. Watson, R. A. Laskowski, and J. M. Thornton, Curr. Opin. Struct. Biol. **15**, 275 (2005). DOI: 10.1016/j.sbi.2005.04.003
13. A. Valencia, Curr. Opin. Struct. Biol. **15**, 267 (2005). DOI: 10.1016/j.sbi.2005.05.010
14. A. N. Nekrasov, J. Biomol. Struct. Dyn. **20**, 87 (2002). DOI: 10.1080/07391102.2002.10506825
15. C. E. Shannon, Bell System Techn. J. **27**, 379 (1948). DOI: 10.1002/j.1538-7305.1948.tb01338.x
16. A. G. de Brevern, Biophys. J. **114**, 231a (2018). DOI: 10.1016/j.bpj.2017.11.1286
17. A. N. Nekrasov, L. G. Alekseeva, R. A. Pogosyan, et al., Biochimie **160**, 88 (2019). DOI: 10.1016/j.biochi.2019.02.006

18. W. Jurkowski, M. Brylinski, L. Konieczny, et al., Proteins **55**, 115 (2004). DOI: 10.1002/prot.20002
19. W. Jurkowski, T. Kułaga, and I. Roterman, J. Biomol. Struct. Dyn. **29**, 79 (2011). DOI: 10.1080/07391102.2011.10507376
20. A. N. Nekrasov, A. A. Anashkina, and A. A. Zinchenko, in *Proc. 2nd Int. Conf. "Theoretical Approaches to Bioinformation Systems"(TABIS 2013)* (Institute of Physics, Belgrade, 2014), pp. 1–22.
21. A. A. Anashkina and A. N. Nekrasov, Russ. J. Numeric. Analysis Math. Model. **29**, 265 (2014).
22. A. N. Nekrasov and A. A. Zinchenko, J. Biomol. Structure & Dynamics **25**, 553 (2008). DOI: 10.1080/07391102.2008.10507202
23. A. N. Nekrasov, V. V. Radchenko, T. M. Shuvaeva, et al., J. Biomol. Structure & Dynamics **24**, 455 (2007). DOI: 10.1080/07391102.2007.10507133
24. Y. Briers, K. Miroshnikov, O. Chertkov, et al., Biochem. Biophys. Res. Commun. **374**, 747 (2008). DOI: 10.1016/j.bbrc.2008.07.102
25. A. N. Nekrasov, L. E. Petrovskaya, V. A. Toporova, et al., Biochemistry (Moscow) **74**, 399 (2009).
26. A. G. Mikhailova, A. N. Nekrasov, A. A. Zinchenko, et al., Biochemistry (Moscow) **80**, 1331 (2015). DOI: 10.1134/S0006297915100156
27. R. V. Chertkova, N. A. Brazhe, T. V. Bryantseva, et al., PLoS One **12** (2017). DOI: 10.1371/journal.pone.0178280
28. L. N. Shingarova, L. E. Petrovskaya, A. N. Nekrasov, et al., Russ. J. Bioorg. Chem. **36**, 301 (2010). DOI: 10.1134/S1068162010030040
29. L. Holm and C. Sander, Bioinformatics **14**, 423 (1998). DOI: 10.1093/bioinformatics/14.5.423

Levels of Hierarchical Organization of Protein Sequences. Analysis of Entropy Characteristics

A.N. Nekrasov*, Y.P. Kozmin*, S.V. Kozyrev, N.G. Esipova***, R.H. Ziganshin*, and A.A. Anashkina*****

**Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, Moscow, 117997 Russia*

***Steklov Mathematical Institute, Russian Academy of Sciences, ul. Gubkina 8, Moscow, 119991 Russia*

****Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, ul. Vavilova 32, Moscow, 119991 Russia*

This research investigates 24647 non-homologous protein sequences. The occurrence profile of peptapeptides was constructed for every sequence and hierarchically organized elements of various sizes were revealed by a special mathematical method in each profile. The correlations between these hierarchical elements were analyzed and it was shown that in the tested set of protein sequences there are 11 levels of protein organization with elements ranging in length from 7 to 56 amino acid residues. It was suggested that the identified levels of organization correspond to elements of a super-secondary structure with different topology.

Keywords: protein sequences, hierarchical organization, entropy