УДК 547.96: 577.112.5/.2.01/.22

# МЕТОД ИССЛЕДОВАНИЯ ИЕРАРХИЧЕСКИХ ЭЛЕМЕНТОВ В ПРИРОДНЫХ АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

**[1]Некрасов А.Н., [1]Зинченко А.А., [1]Карлинский Д.М., [2]Козырев С.В.**

*[1]ФГБУН «Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова Российской академии наук», Москва;*

*[2]ФГБУН «Математический институт им. В.А. Стеклова Российской академии наук», Москва, e-mail: alexei_nekrasov@mail.ru*

Совокупность экспериментальных данных о пространственной структуре белковых молекул показывает, что белки имеют иерархическую структурную организацию. Тем не менее до настоящего момента не существовало метода, позволяющего выявлять иерархическую организацию в первичных структурах белков. Настоящее исследование позволяет восполнить этот пробел. Анализ значений позиционной информационной энтропии, полученных при анализе природных полипептидных последовательностей из базы NRDB, позволил обосновать новую парадигму структурной организации белков, в соответствии с которой элементарной единицей структурной организации является группа из пяти рядом расположенных остатков. В соответствии с этой парадигмой белковая последовательность рассматривается как совокупность коротких, перекрывающихся фрагментов («информационных единиц»). На основании данного представления, был разработан метод (метод АНализа Информационной Структуры – АНИС [4]), позволяющий выявлять иерархическую организацию информации в первичных структурах белков. Методом «бутстреп» была продемонстрирована устойчивость результатов, полученных с помощью метода АНИС.

**Ключевые слова: анализ аминокислотной последовательности, информационная энтропия, теория информации, иерархия**

# A METHOD FOR DISCOVERY OF HIERARCHY ELEMENTS IN NATURAL PROTEIN SEQUENCES

**[1]Nekrasov A.N., [1]Zinchenko A.A., [1]Karlinskiy D.M., [2]Kozyrev S.V.**

*[1]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow;*

*[2]Steklov Institute of Mathematics of the Russian Academy of Sciences, Moscow, e-mail: alexei_nekrasov@mail.ru*

The experimental data on the spatial structure of the protein molecules shows that proteins have a hierarchical structural organization. However, until now there was no method to identify a hierarchical organization of the primary structures of proteins. This study allows filling this gap. An analysis of the positional information entropy values obtained during an analysis of the natural polypeptide sequence database NRDB allowed to justify a new paradigm of structural organization of proteins, according to which the basic unit of structural organization is a group of five adjacent residues. According to this paradigm, a protein sequence is considered as a set of short, overlapping fragments («information units»). Based on this presentation, we developed a method (the method of ANalysis of Information Structure – «ANIS» [4]) that allows one to identify the hierarchical organization of the information in protein primary structures. The «bootstrap» method was used to demonstrate stability of the results obtained using the of ANIS method.

**Keywords: protein sequence analysis, information entropy, information theory, hierarchy**

Совокупность экспериментальных данных указывает на то, что белковые молекулы должны быть организованы определенным образом, позволяющим проявлять им весьма специфические структурные и функциональные свойства. Более полувека назад Анфинсен предположил, что структурные и функциональные свойства белков определяются их аминокислотной последовательностью: «информация... о нативных вторичных и третичных структурах белков, содержится в их аминокислотной последовательности» [1]. В соответствии с этим утверждением аминокислотная последовательность должна полностью определять пространственную организацию белка.

Однако в настоящее время методы анализа белковых последовательностей позволяют выявлять только трансмембранные участки, кластеры гидрофобных остатков (предположительно – те, которые участвуют в формировании гидрофобных ядер белковых глобул), а также более или менее надежно идентифицировать элементы вторичной структуры белков. Целый ряд других особенностей организации трехмерной структуры белков при анализе их первичных структур выявить не удается. В ряде работ предпринимались попытки оценить степень упорядоченности аминокислотных остатков в первичных структурах белков, которые привели к очень низким ее значе-

ниям [8, 9]. Более того, авторы работы [5] пришли к выводу, что «аминокислотные последовательности белков представляют собой слабо отредактированные случайные гетерополимеры». Эти результаты противоречат наблюдаемым свойствам белков, таким как способность к самоорганизации, селективное связывание с регуляторами и субстратами (в случае ферментов), высокая эффективность каталитических реакций и т.д. Задача требует своего решения. Очевидно, что большинство характерных свойств белков связаны с физико-химическими свойствами аминокислотных остатков, и, следовательно, необходимо найти такую статистическую модель, которая могла бы позволить описать физические и химические свойства аминокислотных остатков с помощью их статистических параметров, которые можно получить из полипептидных цепей. В работе [6] было показано, что физико-химические свойства аминокислотных остатков можно адекватно описывать, если рассматривать не только отдельные аминокислотные остатки, но и их непосредственное окружение в первичной структуре. Для того чтобы определить размер окрестности, которую необходимо учитывать, была исследована позиционная информационная энтропия в наборах негомологичных белковых последовательностей.

**Базы данных**

В качестве объектов исследований использовались последовательности из баз данных NRDB [3] трех различных релизов: NRDB 30, NRDB 60 и NRDB 90.

- Исходная база данных NRDB 30 содержит 192,518 белковых последовательностей.
- Исходная база данных NRDB 60 содержит 252,926 белковых последовательностей.
- Исходная база данных NRDB 90 содержит 534,936 белковых последовательностей.

**Теоретическое обоснование метода**

Чтобы получить искомую оценку области, адекватную особенностям природных полипептидных цепей (размер «информационной единицы») для аминокислотных последовательностей из различных релизов базы данных NRDB, были рассчитаны вероятности ($P^k$) появления различных пар аминокислотных остатков с фиксированным числом позиций ($k$) между ними (рис. 1). В каждой такой паре остатков один остаток (в положении $n$) будем называть «корневым», а второй остаток в паре (в положении $n + k$) – «переменным». В процессе расчета «корневой» остаток последовательно смещается от позиции 1 в аминокислотной последовательности белка до остатка номер $L$-$k$, где $L$ представляет собой общее количество остатков в белке.

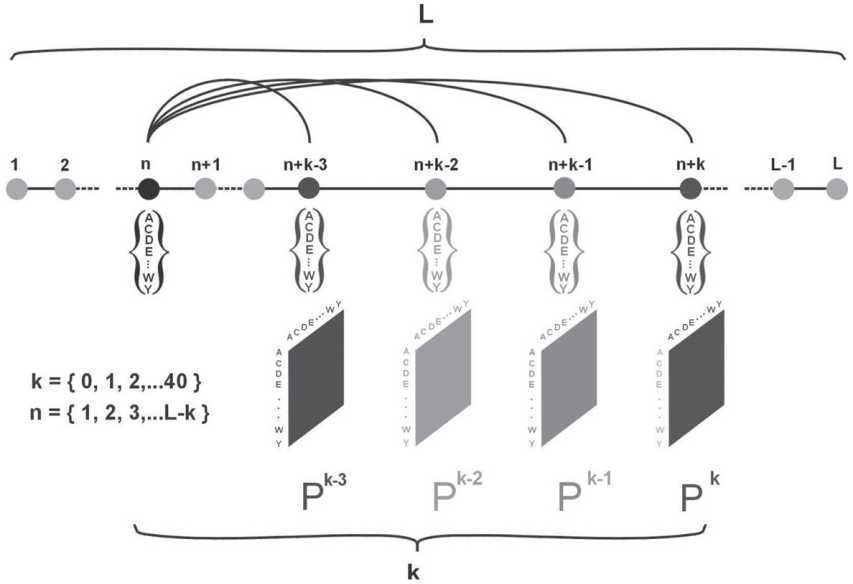


*Рис. 1. Схема анализа базы данных, описывающей белковые последовательности, выполненного для получения матриц частот встречаемости пар аминокислотных остатков. Положение «корневого» остатка в паре изменяется от 1 до L-k (где L представляет собой общее количество остатков в белке. «Переменный» остаток рассматривается в позициях от n + 1 до n + k при положении «корневого» остатка в позиции n последовательности белка*

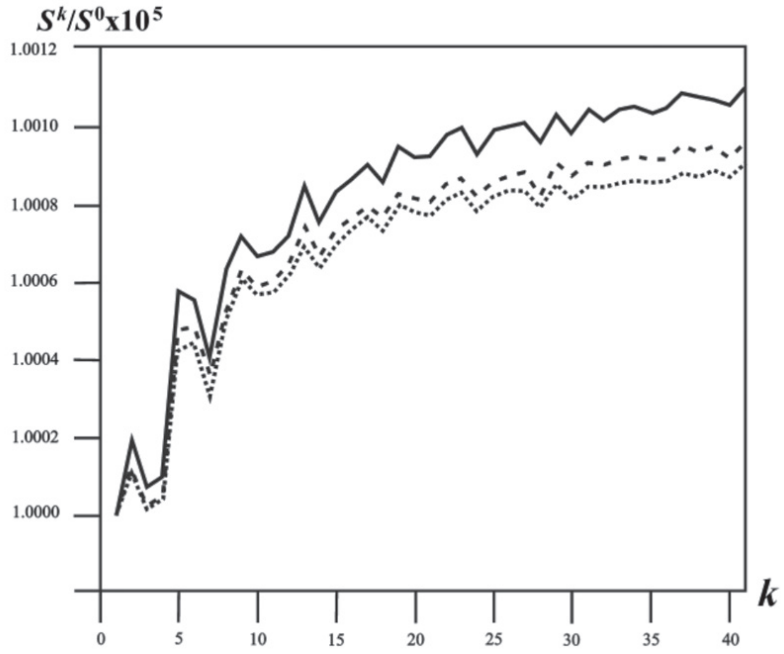


*Рис. 2. Зависимость нормированной информационной энтропии $S^k/S^l$ от расстояния между аминокислотными остатками (k) в паре. Точечная линия обозначает зависимость, полученную для NRDB релиза 90, штриховая – для релиза 60 и сплошная – для релиза 30*

«Переменный» остаток рассматривается в позициях от $n + 1$ до $n + k$ при положении «корневого» остатка в позиции $n$ последовательности белка. При расчете вероятностных матриц $P^k$ были использованы все аминокислотные последовательности белков, входящие в рассматриваемый релиз базы NRDB. В результате были сформированы 50 матриц размером 20x20 ($k = 1 \ldots 50$). Размер матриц определяется числом канонических аминокислотных остатков, встречающихся в белковых последовательностях. Полученные матрицы $P^k$ характеризуют только расстояние между аминокислотными остатками в паре и не зависят от каких-либо других параметров исследуемых белковых последовательностей.

Для того, чтобы охарактеризовать расстояние $\boldsymbol{k}$ между остатками в паре, была рассчитана информационная энтропия Шеннона $S^k$ для матриц $P^k$ по формуле [7]

$$S^k = -\sum_{i-1}^{20}\sum_{j=1}^{20} P_{ij}^k \log_2 P_{ij}^k \ . \qquad (1)$$

В результате были получены 50 значений информационной энтропии, по одному значению для каждой матрицы. Матрицы $P^k$ и значения информационной энтропии $S^k$ были получены для всех трех рассмотренных релизов базы данных NRDB. Исходя из предположения, что аминокислотные остатки, находящиеся в соседних положениях ($k = 1$), наиболее скоррелированы между собой в паре по типу, и, следовательно, матрица $P^1$ имеет минимальное значение энтропии $S^1$, все остальные значения $S^k$ были пронормированы на эту величину. График нормированных значений информационной энтропии $S^k/S^1$ для различных релизов NRDB приведен на рис. 2. Видно, что полученные кривые имеют S-образную форму и сходны для различных релизов NRDB. Для трех представленных зависимостей наблюдается почти полное совпадение локальных максимумов и минимумов. Это позволяет предположить, что полученные данные представляют собой специфические характеристики природных полипептидных цепей как класса молекул и не зависят от размера и состава анализируемых наборов данных.

Очевидно, что при росте $k$ величина $S^k/S^1$ увеличивается и наблюдается падение амплитуды колебаний. При $k > 30$ функция $S^k/S^1$ почти достигает плато.

Наиболее интересной особенностью полученной зависимости является то, что на расстоянии между аминокислотными остатками $k = 5$ начинается устойчивый рост значений нормированной информационной

энтропии $S^k/S^1$. Минимальные значения информационной энтропии наблюдаются при значениях $k$ от 1 до 4. Основываясь на этих данных, мы предположили, что пентапептидные фрагменты ($k = 4$) можно рассматривать как единицы структурной огранизации белка. Далее такие структурные единицы белков мы будем называть «информационными единицами» («information units», IU).

**Метод анализа информации о структуре белка**

Основываясь на предложенном приближении, белковую последовательность можно рассматривать не как последовательность аминокислотных остатков, а как систему последовательно расположенных и перекрывающихся (со сдвигом на одну позицию) информационных единиц. Для использования предложенного приближения при анализе белковых последовательностях был разработан специальный алгоритм.

Пусть первичная структура белка есть последовательность аминокислот $A_i$, $i = 1,...,L$, где аминокислоты могут быть 20 типов.

В соответствии с предложенной парадигмой структурной организацией белков, каждой последовательности аминокислот, имеющей длину $M = 5$, сопоставим величину, определенную следующим образом. Возьмем базу данных NRDB, состоящую из первичных структур белков. Последовательности $S = S_1...S_M$ из $M$ аминокислот сопоставим частоту $f(S)$ встречаемости в качестве всевозможных подпоследовательностей стоящих рядом аминокислот во всех белках из рассмотренной базы данных.

Выберем теперь набор последовательностей $S'$ длины $M$, отличающихся от $S$ заменой одной аминокислоты (таких последовательностей существует $20^M$ штук). Последовательности $S'$ сопоставим соответствующую частоту $f(S')$. Проведем суммирование по всем возможным последовательностям $S'$, получаемым заменой одной аминокислоты и отвечающим $S$. В результате получим функцию

$$F(S) = \sum_{S'} f(S'). \qquad (2)$$

Теперь для данного рассматриваемого белка $P = \{A_i\}$ длины $L$ мы будем рассматривать всевозможные подпоследовательности $S \subset P$ длины $M$ стоящих рядом аминокислот, таких подпоследовательностей в белке длины $L$ будет $L - M + 1 = L - 4$ штук. Введем нумерацию этих последовательностей $S = S_i$ длины 5 по их центрам $i$, таким образом, $i = 3,...,L-2$, и рассмотрим функцию

$$F(i) = F(S_i), \qquad (3)$$

то есть суммарную частоту встречаемости в базе данных NRDB всевозможных подпоследовательностей $S'$ длины 5, отвечающих подпоследовательности $S$ с центром в аминокислоте с номером $i$ в данном белке.

Ранее для белка из $L$ аминокислот была построена функция $F(i)$, $i = 3,...,L-2$ частоты встречаемости в базе данных подпоследовательности длины 5. Сопоставим этой функции гистограмму, то есть функцию $F(x)$ на отрезке $(2, L-2]$, принимающую для $x \in (i-1, i]$ значение $F(x) = F(i)$.

Построим теперь по гистограмме $F(x)$ функцию нелинейного сглаживания $G(a,x)$ по следующему правилу.

Рассмотрим сглаживающую функцию $\varphi(x)$ – непрерывную функцию с носителем на отрезке $[-1/2, 1/2]$, $\varphi(-1/2) = \varphi(1/2) = 0$, $\varphi(0) = 1$, $\varphi$ принимает положительные значения в интервале $(-1/2, 1/2)$, монотонно растёт на $[-1/2, 0]$, монотонно убывает на $[0, 1/2]$, график функции симметричен относительно отображения относительно прямой $x = 0$. Мы также считаем, что функция $\varphi$ гладкая, причём производная не обращается в нуль на отрезках $(-1/2, 0)$ и $(0, 1/2)$.

В качестве сглаживающей функции можно выбрать соответствующим образом сдвинутую и перерастянутую гауссовскую функцию $e^{-x^2}$, график которой обрезан на половине высоты.

Будем также рассматривать сдвиги и растяжения сглаживающей функции

$$\varphi^{(a,b)}(x) = \varphi\left(\frac{x-b}{a}\right), \qquad (4)$$

где $a \geq 1$. Функция $\varphi^{(a,b)}$ имеет носитель в отрезке $[-1/2a + ba, 1/2a + ba]$.

Определим теперь функцию нелинейного сглаживания $G(x, a)$ для функции $F$ по следующей формуле

$$G(b,a) = \sup c,\ c : c\varphi^{(ab)}(x) \leq F(x),\ \forall x. \qquad (5)$$

Таким образом, $G(b, a)$ есть максимальная высота sup$c$ сглаживающей функции шириной $a$ с центром носителя в точке $b$, которую можно вписать в гистограмму $F$. Параметр $a$ назовём масштабом сглаживания.

Носитель функции $G(x, a)$ имеет следующий вид. Функция $G(x, a)$ может быть отлична от нуля при $a \in [1, L-4]$, $x \in [2 + a/2, L - 2 - a/2]$. Таким образом, функция нелинейного сглаживания име-

ет носитель, являющийся подмножеством треугольника на плоскости с координатами $(x, a)$ с вершинами $(L/2, L-4)$, $(2+1/2, 1)$, $(L-2-1/2, 1)$.

Совокупность полученных значений сглаживающей функции $G(x, a)$ назовем информационной структурой исследуемого белка.

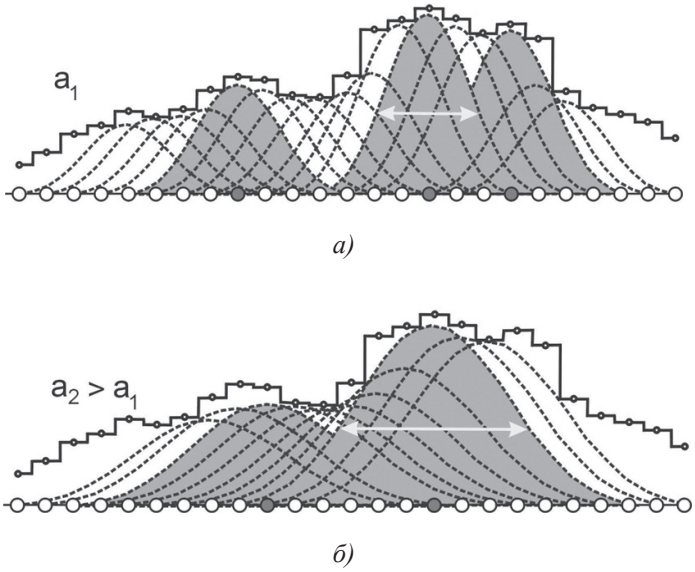


*Рис. 3. Нелинейное сглаживание «популяции» функции эквивалентных: IU. а) с масштабом сглаживания $a_1$; б) с масштабом сглаживания $a_2 > a_1$. Масштабы сглаживания $a_1$ и $a_2$ показаны горизонтальными стрелками. Пунктирные и сплошные линии показывают все возможные функции сглаживания для рассматриваемого фрагмента белковой последовательности. Локальные максимумы сглаживающих функций G(i, a) выделены серым фоном, а заполненные кружки показывают центральные аминокислотные остатки для таких сглаживающих функций*

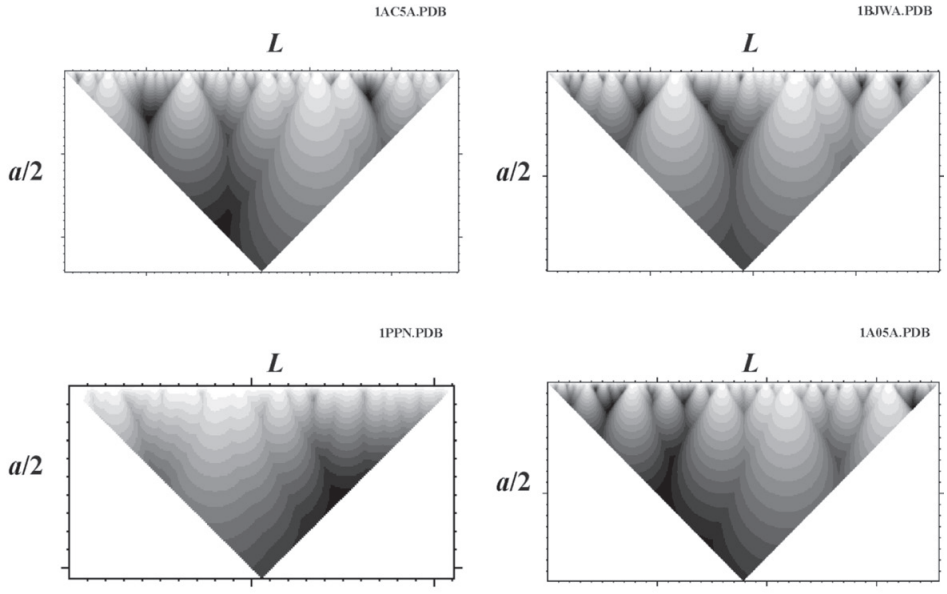


*Рис. 4. Примеры расчетов информационных структур для различных белков. 1AC5A.PDB – KEX1(DELTA)P (цепь A) из Saccharomyces Cerevisiae, 1BJWA – аспартатаминотрансфераза (цепь A) из Thermus Thermophilus, 1PPN – моноклинный папаин из Carica Papaya, 1A05A – 3-изопропилмалат дегидрогеназа (цепь A) из Thiobacillus Ferrooxidans, a – масштаб сглаживания, L – число остатков в белковой последовательности*

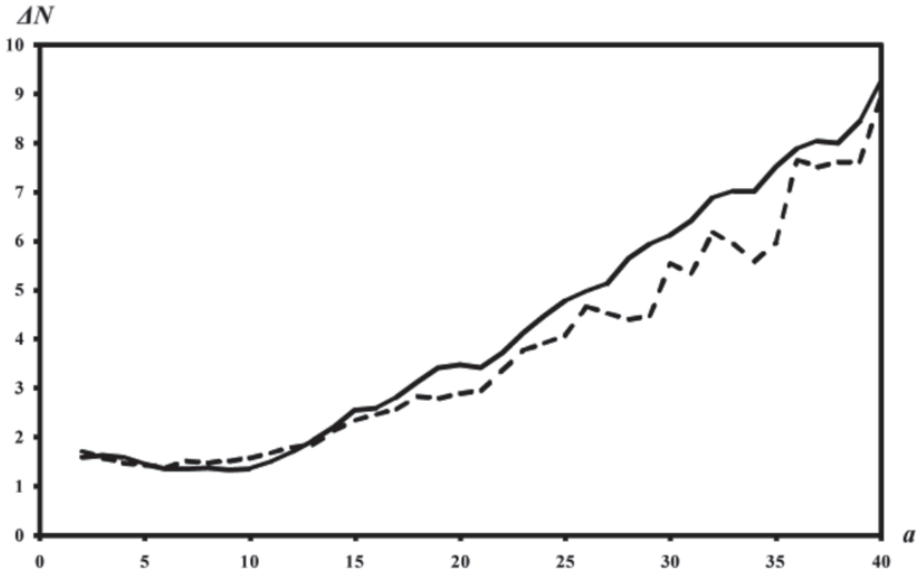


*Рис. 5. График соотношения между ΔN и a. ΔN – среднее отклонение сглаживающей функции G(i, a), выраженное в количестве позиций, для локальных максимумов (сплошная линия) и локальных минимумов (пунктирная линия), a – масштаб сглаживания, полученный для 100 случайно выбранных белковых последовательностей*

**Точность результатов, полученных методом АНИС**

Мы использовали метод «бутстреп» [2, 10], чтобы создать 100 тестовых наборов белковых последовательностей на основе ранее описанной базы данных NRDB релиз 90.

Эти тестовые наборы были использованы для проверки устойчивости результатов, получаемых АНИС методом [4]. На основе этих тестовых наборов была получена статистика встречаемости для IU. Эта статистика была использована для расчета информационных структур 100 случайно выбранных белковых последовательностей длиной более 300 аминокислотных остатков. На рис. 5 представлен график среднего отклонения в положениях локальных максимумов и локальных минимумов сглаживающей функции $G(x, a)$ при различных значениях масштаба сглаживания $a$.

На рис. 5 показано, что для масштаба сглаживания $a$, находящегося в интервале значений от 2 до 12, среднее отклонение в позициях локальных максимумов и минимумов – примерно ± 2 позиции, и среднее отклонение растет по мере увеличения масштаба сглаживания. Максимальное среднее отклонение при $a = 40$ для обоих максимумов и минимумов – почти ± 9. Ошибка идентификации элементов информационной структуры изменяется от 1 до 3 позиций для значений масштаба сглаживания от 2 до 22, т.е. для фрагментов белковой последовательности длиной от 5 до 45 аминокислотных остатков.

**Выводы**

Анализ данных позиционной информационной энтропии позволил обосновать новую парадигму структурной организации белков, в которой элементарной структурной единицей является группа из пяти рядом стоящих аминокислотных остатков. Использование этой парадигмы позволило разработать новый метод анализа аминокислотных последовательностей белков, позволяющий выявлять иерархическую организацию в их первичной структуре.

*Авторы благодарят коллектив Лаборатории химии протеолитических ферментов ИБХ РАН за полезное обсуждение работы и помощь в организации рабочего процесса.*

*Эта работа была поддержана Российской Академией Наук [грант по программе фундаментальных исследований в стратегических направлениях развития науки Президиума РАН «Фундаментальные проблемы математического моделирования» (код программы: II.4П), проект «Математическая модель пространственной орга-*

*низации природных полипептидных цепей на основе информационного контента первичной структуры»].*

### Список литературы

1. Anfinsen C.B., Haber E., Sela M., White F.H.Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain // Proc Natl Acad. Sci. USA. – 1961. – Vol. 47, № 9. – P. 1309–1314.

2. Efron B., Tibshirani R.J. An Introduction to the Bootstrap. NY: Chapman & Hall, 1993. – 456 P.

3. Holm L., Sander C. Removing near-neighbour redundancy from large protein sequence collections // Bioinformatics. – 1998. – Vol. 14, № 5. – P. 423–429.

4. Nekrasov A.N. Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of proteins // J. Biomol. Struct. Dyn. – 2004. – Vol. 21, № 5. – P. 615–624.

5. Ptitsyn O.B., Volkenstein M.V. Protein structures and neutral theory of evolution // J. Biomol. Struct. Dyn. – 1986. – Vol. 4, № 1. – P. 137–156.

6. Rogov S.I., Nekrasov A.N. A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences // Protein Eng. – 2001. – Vol. 14, № 7. – P. 459–463.

7. Shannon C.E. A Mathematical Theory of Communication // The Bell System Technical Journal. – 1948. – Vol. 27. – P. 379–423, P. 623–656.

8. Szoniec G., Ogorzalek M.J. Entropy of never born protein sequences // SpringerPlus. – 2013. – Vol. 2, № 1. – P. 200.

9. Weiss O., Jiménez-Montaño M.A., Herzel H. Information content of protein sequences // J. Theor. Biol. – 2000. – Vol. 206, № 3. – P. 379–386.

10. Wu C.F.J. Jackknife, bootstrap and other resampling methods in regression analysis // Annals of Statistics. – 1986. – Vol. 14, № 4. – P. 1261–1295.