

УДК: 519

## Распознавание и анализ устойчивости структурных мотивов типа $\alpha$ - $\alpha$ -уголок в глобулярных белках

©2013 Руднев В.Р.<sup>\*1</sup>, Панкратов А.Н.<sup>1,2</sup>, Куликова Л.И.<sup>1</sup>, Дедус Ф.Ф.<sup>2,1</sup>, Тихонов Д.А.<sup>1</sup>, Ефимов А.В.<sup>3</sup>

<sup>1</sup> *Институт математических проблем биологии РАН, 142290, Пущино, ул. Институтская, д.4*

<sup>2</sup> *Факультет ВМК МГУ имени М.В. Ломоносова, 119991 ГСП-1, Москва, Ленинские горы*

<sup>3</sup> *Институт белка РАН, 142290, Пущино, ул. Институтская, д.4*

**Аннотация.** Рассматривается задача распознавания структурных мотивов белков на примере  $\alpha$ - $\alpha$ -уголков. Предложен комбинированный подход к решению задач анализа пространственной структуры белков на основе аналитического описания основной цепи белковой глобулы и спектрального метода распознавания повторов. Нами найдено 110  $\alpha$ - $\alpha$ -уголков в базе данных PDB, соответствующих заданному эталону. Методом молекулярной динамики показано, что  $\alpha$ - $\alpha$ -уголки как автономные структуры устойчивы в водной среде.

**Ключевые слова:** распознавание образов, ортогональные полиномы, аналитическое описание, супервторичные структуры белков.

### ВВЕДЕНИЕ

Определение пространственной структуры белка является актуальной и трудоемкой задачей. Наиболее развитым и применяемым на практике является метод определения структуры на основе данных рентгеновского рассеяния. Более поздним и довольно успешно применяемым является метод определения структуры по спектрам ядерного магнитного резонанса. Этот метод особенно плодотворен в тех случаях, когда потерпели неудачу все попытки вырастить большие кристаллы белка. Еще одним преимуществом метода ЯМР является тот факт, что на конформацию белка не оказывают влияния молекулы кристаллического окружения. Поскольку исследование структуры этим методом проводится в растворе, результаты сильно зависят от растворителя. Более реалистичную картину структуры белка дает исследование водных растворов белков.

Объем информации о белках очень велик и постоянно растет: в базе данных UniProt доступны последовательности более чем трех миллионов белков, что значительно превышает число известных структур. В базе данных PDB доступны структуры более 60000 белков [1]. Поэтому разработка теоретических методов предсказания трехмерной структуры белка представляется крайне актуальной и необходимой. Наиболее надежное предсказание третичной структуры белка дает моделирование на основе гомологии (или сравнительное моделирование), то есть только в том случае, когда известна пространственная структура какого-либо гомологичного белка. Этот метод

---

\*volodyarv@mail.ru.

разработан различными авторами [2–5], и идея метода состоит в использовании накопленного объема знаний об уже известных структурах белков.

Но ни один из разработанных теоретических методов предсказания не является совершенным и абсолютно надежным. Это связано с тем, что прогнозирование вторичной структуры белка ( $\alpha$ - и  $\beta$ -структур) по первичной структуре является некорректно поставленной задачей, поскольку вторичная структура зависит от третичной структуры. В силу актуальности и большого интереса исследователей к данной проблеме разрабатываются различные подходы к решению задачи прогнозирования вторичной структуры белка по первичной структуре [6, 7].

Регулярные структуры ( $\alpha$  и  $\beta$ ) являются частью хорошо организованной пространственной структуры белка, образованной водородными связями, которые осуществляются между элементами цепи полимера в случае нативного состояния. В денатурированном (расплавленном) состоянии белок имеет другую структуру, в которой, по-видимому, все или почти все водородные связи между элементами цепи белка замещены связями с растворителем.

Значение вторичной структуры белка заключается в том, что, во-первых, возможность ее образования определяется локальными характеристиками белковой цепи и, во-вторых, она имеет относительно простую геометрическую форму. Однако она является частью более сложной пространственной структуры, которая имеет нелокальный характер, поскольку отдаленные участки цепи полимера могут контактировать друг с другом в пространственной укладке.

Ранее были разработаны простые правила укладки белковой цепи [8], которые привели к построению структурных деревьев, описывающие многообразие белковых структур. Как в случае с рентгеноструктурным анализом и ЯМР, дальнейшее развитие метода связано с разработкой информационных технологий. В частности, необходимо автоматически распознавать структурные мотивы в известных структурах белков, что может способствовать построению соответствия между первичной и третичной структурой белка. Известны структурные мотивы, состоящие из двух и более элементов вторичной структуры и имеющих уникальные укладки полипептидной цепи в пространстве ( $\alpha$ - $\alpha$ -уголки,  $\alpha$ - $\alpha$ -шпильки,  $L$ -образные и  $V$ -образные структуры и др.). Одним из часто встречаемых структурных мотивов в гомологичных и негомологичных белках являются  $\alpha$ - $\alpha$ -уголки [9]. Эта супервторичная структура образована двумя соседними по полипептидной цепи  $\alpha$ -спиралями, связанными между собой перетяжками и упакованных ортогонально (крестообразно). В белках  $\alpha$ - $\alpha$ -уголки встречаются в форме левой суперспирали. Их последовательности имеют определенное расположение в цепи гидрофобных, гидрофильных и глициновых остатков. Выдвинута гипотеза, что структурные мотивы с уникальными укладками могут быть зародышами при сворачивании белка, а остальные участки цепи пристраиваются к ним в соответствии с простыми принципами. Однако, независимо от того, по какому механизму происходит сворачивание белка, структурные мотивы могут быть использованы в качестве стартовых структур для поиска возможных укладок полипептидной цепи [8].

## ПОСТАНОВКА ЗАДАЧ

В работе рассмотрены три основные задачи:

– Подготовка обучающей выборки супервторичных структур белковых молекул. На этом этапе решается задача распознавания супервторичных структур в базе данных PDB по заданному шаблону эталона методом спектрально-аналитического сравнения пространственных структур белковых молекул. Алгоритм распознавания  $\alpha$ - $\alpha$ -уголков в белковых молекулах сводится к спектральному алгоритму поиска повторов [10].

- Поиск характерных признаков супервторичных структур в аминокислотных последовательностях на обучающей выборке.
- Верификация супервторичных структур методом молекулярной динамики. На этом этапе используется свойство автономной устойчивости супервторичных структур для проверки результатов распознавания.

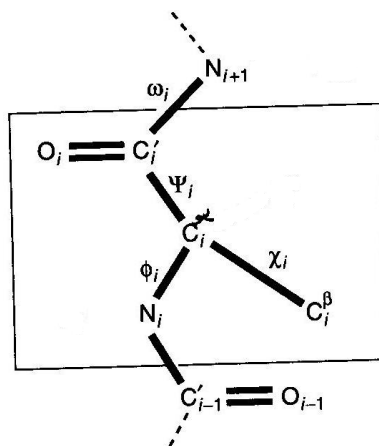
## ОБСУЖДЕНИЕ

### Подготовка обучающей выборки

В рамках поставленной задачи необходимо сформировать выборку  $\alpha$ - $\alpha$ -уголков из базы данных PDB. Выборка представляет собой перечень белков с указанием координат атомов, образующих эти структуры.

### Метод решения

Пространственная структура вторичных и супервторичных структур в белках определяется координатами  $C^\alpha$ -атомов основной цепи белковой глобулы. Отсюда следует, что при решении данной задачи координатами атомов боковых цепей белковых молекул можно пренебречь.



**Рис. 1.** Фрагмент основной цепи полипептида. Обозначение атомов и торсионных углов белка. Основная цепь белка состоит из повторяющихся последовательностей трех атомов, относящихся к одному остатку – амидного остатка  $N$ , атома  $C^\alpha$ , и карбонильного углерода  $C'$ ; обычно эти атомы обозначают как  $N_i$ ,  $C_i^\alpha$ ,  $C_i'$ , соответственно, где  $i$  – номер остатка, начиная с  $N$ -конца цепи.  $\phi$ ,  $\psi$ ,  $\omega$  – торсионные углы основной цепи,  $\chi$  – углы боковых цепей.

Таким образом, дальше под пространственной структурой белковых молекул понимается пространственная структура  $C^\alpha$ -атомов основной цепи.

Пространственную структуру белка в целом, а также пространственную структуру вторичных и супервторичных структур, можно представить в виде параметрического уравнения кривой в трёхмерном пространстве:

$$\begin{cases} x(t) = \sum_{i=0}^N A_i \varphi_i(t) \\ y(t) = \sum_{i=0}^N B_i \varphi_i(t), \\ z(t) = \sum_{i=0}^N C_i \varphi_i(t) \end{cases} \quad (1)$$

где  $\{\varphi_i\}$  - система ортогональных полиномов,  $A_i, B_i, C_i$  - коэффициенты разложения функций по ортогональным базисам. Кривая (1) получена с помощью методов аналитического описания основной цепи, образуемой координатами  $C^\alpha$ -атомов. В качестве таких методов использовались сплайны и ортогональные многочлены [11, 12].

Наиболее подходящими для исследуемых функций являются полиномы Лежандра и Чебышева. Коэффициенты разложения, которые рассматриваются как спектральные признаки сигнала, рассчитываются по общей формуле

$$Z_i = \int_a^b f(t)\varphi_i(t)\rho(t)dt \quad (2)$$

с заданным весом  $\rho(t)$ .

На рис. 2 представлена пространственная структура классического  $\alpha$ - $\alpha$ -уголка, а на рис. 3 её аналитическое описание.

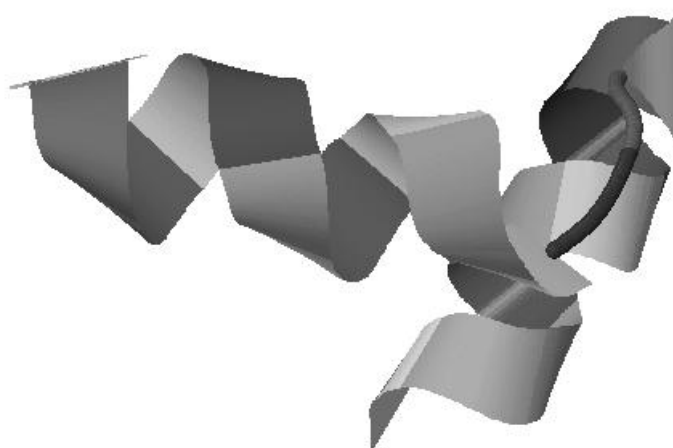


Рис. 2. Ленточная модель  $\alpha$ - $\alpha$  уголка с короткой перетяжкой.

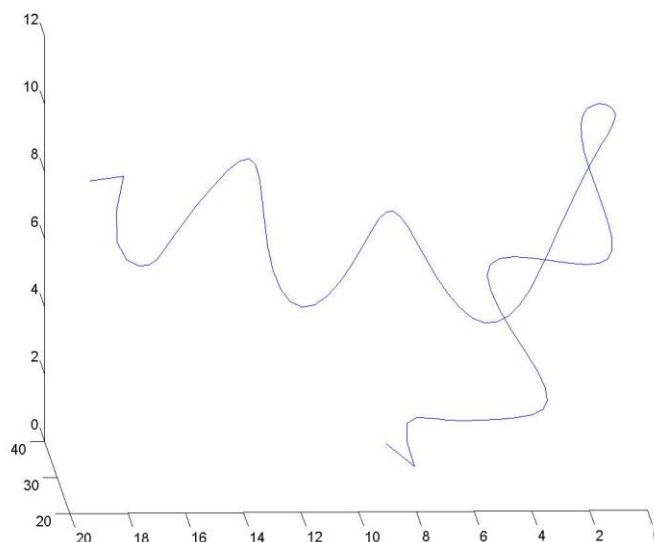


Рис. 3. Аналитическое описание  $\alpha$ - $\alpha$  уголка с короткой перетяжкой по полиномам Лежандра.

Переход от параметрического уравнения кривой в трёхмерном пространстве к натуральному уравнению кривой в трёхмерном пространстве даёт параметризованное описание кривизны и кручения в зависимости от натурального параметра кривой - длины дуги (2).

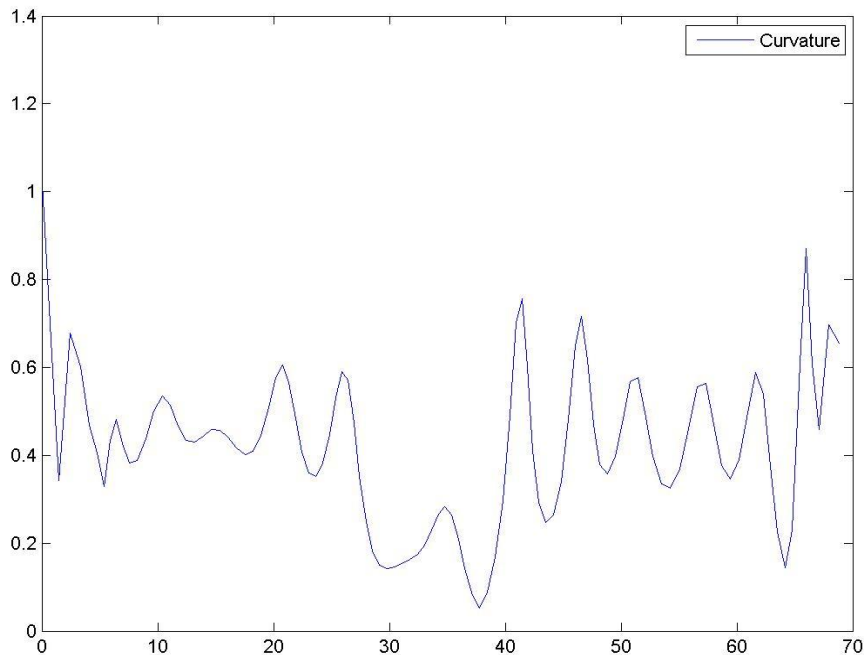
$$\begin{cases} x = x(t), \\ y = y(t), \\ z = z(t) \end{cases} \Rightarrow \begin{cases} C(s) = \sqrt{\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2}, \\ T(s) = \frac{\begin{vmatrix} \dot{x} & \dot{y} & \dot{z} \\ \ddot{x} & \ddot{y} & \ddot{z} \\ \ddot{x} & \ddot{y} & \ddot{z} \end{vmatrix}}{\ddot{x}^2 + \ddot{y}^2 + \ddot{z}^2} \end{cases}, \quad (3)$$

где  $C$  – функция кривизны,  $T$  – функция кручения,  $s$  – натуральный параметр кривой – длина дуги:

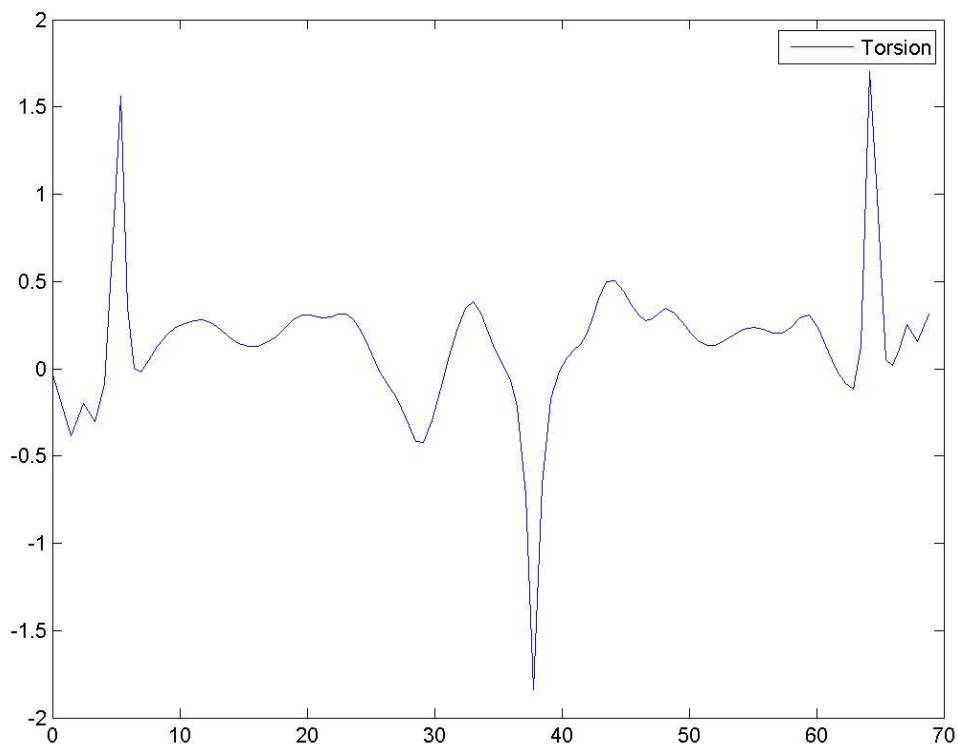
$$s(t) = \int_{t_1}^{t_2} \sqrt{\dot{x}^2 + \dot{y}^2 + \dot{z}^2} dt .$$

Такое описание инвариантно по отношению к выбору декартовой системы координат. Функции кривизны и кручения являются характеристическими профилями пространственной структуры белковой молекулы. При этом регулярные участки,  $\alpha$ -спирали, представляются областями постоянных значений кривизны и кручения. На рис. 4 и 5 представлены профили кривизны и кручения, полученные из аналитического описания структуры  $\alpha$ - $\alpha$  уголка, представленного на рис. 1.

По заданным профилям можно однозначно восстановить пространственную структуру образца рис. 1 с точностью до выбора системы отсчета декартовых координат. Переход от натурального уравнения кривой в трёхмерном пространстве к параметрическому уравнению кривой в трёхмерном пространстве осуществляется по формуле Френе [13].



**Рис. 4.** Профиль кривизны, полученный из аналитического описания  $\alpha$ - $\alpha$  уголка, представленного на рис. 1.



**Рис. 5.** Профиль кручения, полученный из аналитического описания  $\alpha$ - $\alpha$  уголка, представленного на рис. 1.

Алгоритм поиска  $\alpha$ - $\alpha$ -уголков в белковых молекулах основан на обобщенном спектрально-аналитическом методе, он разработан и ранее был успешно применен для решения задачи поиска повторов в геномных последовательностях [14].

Описанный выше подход реализован в виде программно-алгоритмического комплекса «Protein Reviser». Поиск  $\alpha$ - $\alpha$ -уголков осуществлялся по эталону, представленному на рис. 1, в базе данных PDB. С использованием разработанного метода были найдены 110  $\alpha$ - $\alpha$ -уголков в базе данных PDB, соответствующих заданному эталону 1D1L  $C^\alpha$ : 15-37.

### Автономная устойчивость $\alpha$ - $\alpha$ -уголков

Была выдвинута гипотеза об автономной устойчивости  $\alpha$ - $\alpha$  уголков в водной среде, которая легла в основу исследования, описанного в данном разделе.

Под автономной устойчивостью в данном случае понимается устойчивость пространственной структуры исследуемого структурного мотива отдельно от белковой молекулы, в которой данная структура была обнаружена.

Основная идея заключается в том, что если исследуемый структурный мотив является автономно устойчивым, то справедливы следующие утверждения:

- все признаки этой структуры локализованы в соответствующем участке первичной структуры белковой молекулы;
- данное свойство можно использовать как дополнительный признак (свойство) этой структуры.

Для проверки гипотезы об автономной устойчивости были отобраны 57  $\alpha$ - $\alpha$ -уголков. Для данной выборки был произведен численный эксперимент методом молекулярной динамики. Полученные траектории были проанализированы экспертами на предмет устойчивости. Исследование показало, что  $\alpha$ - $\alpha$ -уголок является автономно

устойчивой структурой, и это свойство может рассматриваться как дополнительный верифицирующий признак в исследованиях. Моделирование производилось средствами Amber 11. Параметры, при которых проводилось моделирование, представлены ниже:

- потенциальное поле FF03,
- температура 300 К,
- длина равновесной траектории 1 нс,
- слой воды 9 Å.

В дальнейшем планируется определить граничные условия устойчивости данных структур, провести исследование на более обширной выборке.

### Поиск характерных признаков $\alpha$ - $\alpha$ -уголков в первичной структуре

Данное исследование опирается на свойстве автономной устойчивости структурного мотива данного вида.

Из свойства автономной устойчивости следует, что все признаки  $\alpha$ - $\alpha$ -уголков в первичной структуре локализованы в самой структуре  $\alpha$ - $\alpha$ -уголков. Таким образом, достаточно исследовать первичные структуры только  $\alpha$ - $\alpha$ -уголков отдельно от белковых молекул, в которых были найдены эти структуры.

В результате исследования были обнаружены интересные закономерности чередования определённых групп аминокислотных остатков, в частности, было подтверждено статистически наличие глицина в перетяжке и чередование гидрофобных аминокислотных остатков в  $\alpha$ -спиралях в определённых позициях, относительно глицина. На рис. 6 представлена диаграмма, иллюстрирующая распределение гидрофобных аминокислотных остатков относительно глицина в  $\alpha$ - $\alpha$ -уголках.

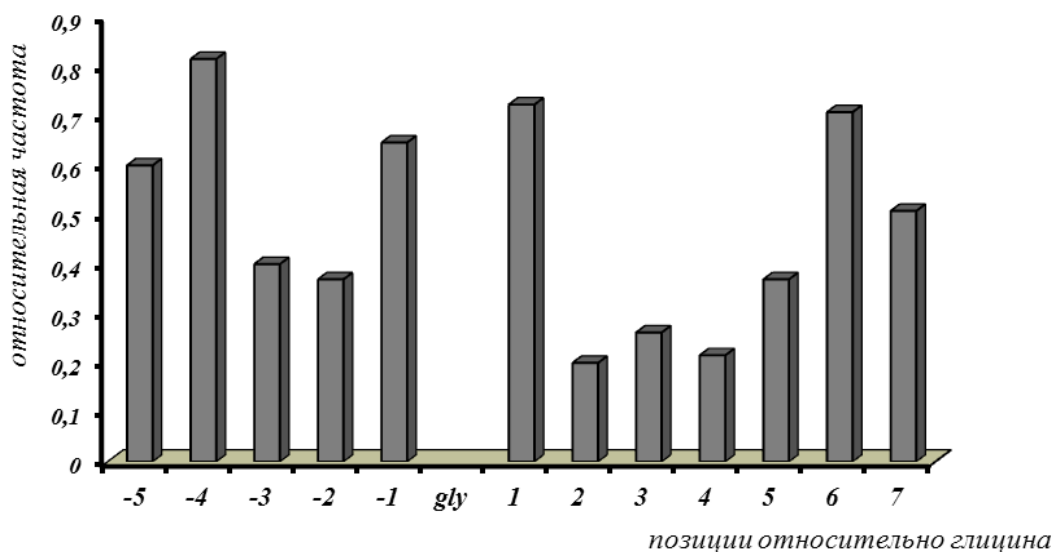


Рис. 6. Распределение гидрофобных аминокислотных остатков относительно глицина в  $\alpha$ - $\alpha$ -уголках.

### Верификация $\alpha$ - $\alpha$ -уголков

Предсказание вторичных и супервторичных структур всегда носит вероятностный характер, так как до сих пор не обнаружено 100%-ой зависимости пространственной белковой структуры от соответствующей ей аминокислотной последовательности. Обнаруженное свойство автономной устойчивости позволит верифицировать предсказанные  $\alpha$ - $\alpha$ -уголки. Для этого предлагается следующий процесс:

1. В первичной структуре белковой молекулы распознать структурный мотив (метод распознавания находится в процессе разработки и не представлен в этой статье).
  2. По углам  $\phi$  и  $\psi$  можно легко построить пространственную модель  $\alpha$ - $\alpha$ -уголка, так как  $\alpha$ - $\alpha$ -уголок имеет определённую пространственную структуру.
  3. Опираясь на свойство автономной устойчивости, осуществить молекулярно-динамический расчёт и получить итоговую пространственную модель распознанного  $\alpha$ - $\alpha$ -уголка.
  4. Методом, предложенным в первом разделе, сравнить получившуюся модель с эталоном. При совпадении можно считать, что предсказание прошло успешно.
- Также нужно заметить, что при относительно небольшой длине супервторичных структур молекулярно-динамический расчёт осуществляется довольно быстро.

## ЗАКЛЮЧЕНИЕ

В ходе исследований были получены следующие основные результаты:

На первом этапе подготовки обучающей выборки найдено 110  $\alpha$ - $\alpha$ -уголков в базе данных PDB, соответствующих заданному эталону 1D1L  $C^{\alpha}$ : 15-37. Для распознавания был предложен подход, основанный на континуальном описании цепи белковой молекулы и спектральном методе распознавания повторов. Аналогичный подход для сравнения белковых структур был предложен в работе [15].

На втором этапе были выявлены характерные признаки  $\alpha$ - $\alpha$ -уголков в аминокислотных последовательностях, что позволяет осуществлять распознавание  $\alpha$ - $\alpha$ -уголков в первичных структурах белковых молекул, опираясь также на эти признаки.

На третьем этапе проведён анализ автономной устойчивости  $\alpha$ - $\alpha$ -уголков методом молекулярной динамики в воде. Эксперимент ставился на выборке из 57  $\alpha$ - $\alpha$ -уголков. Моделирование производилось средствами Amber 11. Исследование показало, что  $\alpha$ - $\alpha$ -уголок является автономно устойчивой структурой, и это свойство может рассматриваться как дополнительный верифицирующий признак в исследованиях.

Работа выполнена при поддержке проектов РФФИ № 13-01-00340, 11-07-00519, 11-07-00716.

## СПИСОК ЛИТЕРАТУРЫ

1. Holtje H.-D., Sippl W., Rognan D., Folkers G. *Molecular Modeling. Basic Principles and Applications. Third, Revised and Expanded Edition*. WILEY-VCH Verlag GmbH&Co.KGaA, 2008. 318 p.
2. Jonson M.S., Srinivasan N., Sowdhamini R. and Blundell T.L. Knowledge-based protein modelling. *Critical Reviews in Biochemistry and Molecular Biology*. 1994. V. 29. P. 193–316.
3. Sali A., Overington M.S. and Blundell T.L. From comparisons of protein sequences and structures to protein modelling and design. *Trends in Biochemical Science*. 1990. V. 15. P. 235–240.
4. Jones T.A., Thirup S. Using known substructures in protein model building and crystallography. *EMBO Journal*. 1986. V. 5. P. 819–822.
5. Dudek M.J., Scheraga H.A. Protein structure prediction uses a combination of sequence gemology and global energy minimization. *Journal of Computational Chemistry*. 1990. V. 11. P. 121–151.
6. Рудаков К.В., Торшин И.Ю. Вопросы разрешимости задачи распознавания вторичной структуры белка. *Информатика и ее применения*. 2010. Т. 4. Вып. 2. С. 25–35.
7. Карасев В.А., Лучинин В.В. Способ прогнозирования вторичной структуры белка. *Федеральная служба по интеллектуальной собственности, патентами и*



- товарным знакам. URL: [http://vector-machine.narod.ru/Description\\_of\\_patent.pdf](http://vector-machine.narod.ru/Description_of_patent.pdf) (дата обращения: 09.04.2010.)
8. Efimov A.V. Standard Structures in Proteins. *Prog. Biophys. Molec. Biol.* 1993. V. 60. P. 201–239.
  9. Ефимов А.В. Новая супервторичная структура белков:  $\alpha$ - $\alpha$  уголок. *Молекулярная биология.* 1984. Т. 18. С. 1524–1537.
  10. Pankratov A.N., Gorchakov M.A., Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova D.A., Olshevets M.M., Pyatkov M.I., Rudnev V.R., Tetuev R.K., Filippov V.V. Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences. *Pattern Recognition and Image Analysis.* 2009. V. 19. № 4. P. 687–692.
  11. Дедус Ф.Ф., Куликова Л.И., Панкратов А.Н., Тетуев Р.К. *Классические ортогональные базисы в задачах аналитического описания и обработки информационных сигналов: учебное пособие.* М.: Издательский отдел ВМиК МГУ, 2004. 147 с.
  12. Дедус Ф.Ф., Махортых С.А., Устинин М.Н., Дедус А.Ф. *Обобщенный спектрально – аналитический метод обработки информационных массивов. Задачи анализа изображений и распознавания образов.* М.: Машиностроение, 1999. 357 с.
  13. Бронштейн И.Н., Семендяев К.А. *Справочник по математике для инженеров и учащихся втузов.* М.: Наука, 1980. 976 с.
  14. Дедус Ф.Ф., Куликова Л.И., Махортых С.А., Назипова Н.Н., Панкратов А.Н., Тетуев Р.К. Аналитические методы распознавания повторяющихся структур в геномах. *Доклады Академии Наук.* 2006. Т. 411. № 5. С. 599–602.
  15. Ranganathan S., Izotov D., Kraka E., Cremer D. Description and recognition of regular and distorted secondary structures in proteins using the automated protein structure analysis method. *Proteins: Structure, Function, and Bioinformatics.* 2009. V. 76. P. 418–438.

Материал поступил в редакцию 21.05.2013, опубликован 17.07.2013.