

ПРЕДСКАЗАНИЕ ТОРСИОННЫХ УГЛОВ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ БЕЛКОВ НА ОСНОВЕ БАЙЕСОВСКОЙ ПРОЦЕДУРЫ РАСПОЗНАВАНИЯ НА ЦЕПЯХ МАРКОВА

Ключевые слова: торсионный угол, байесовская процедура, цепь Маркова, вторичная структура белка, переходные вероятности.

Введение. В настоящее время общими усилиями ученых всего мира расшифрованы геномы человека, шимпанзе, курицы, рыбы *Tetraodon* и некоторых других животных, нескольких видов растений, а также свыше тысячи бактерий. Зная нуклеотидную последовательность гена, можно однозначно определить последовательность аминокислот белка, так как каждая из 20 аминокислот кодируется определенным триплетом нуклеотидов (кодоном). После трансляции последовательности аминокислот из молекулы РНК белок сразу начинает сворачиваться в пространственную конфигурацию. Именно пространственная конфигурация белка определяет его функциональность, поскольку белки в живых организмах взаимодействуют как трехмерные объекты в пространстве. Поэтому в исследованиях белков и их функций придерживаются доктрины «последовательность — структура — функциональность» [1]. Это означает, что функциональность белка определяется его пространственной структурой, а пространственная конфигурация задается его аминокислотной последовательностью.

Существует четыре уровня организации структуры белка:

- первичная структура — линейная последовательность аминокислотных остатков в молекуле белка;
- вторичная структура — формирование на линейной последовательности локальных регулярных структур: α -спиралей и β -слоев;
- третичная структура — расположение элементов вторичной структуры (α -спиралей и β -слоев) в пространстве один относительно другого;
- четвертичная структура — формирование белкового комплекса из отдельных белков.

Структура белка на каждом уровне организации оказывает решающее влияние на формирование структуры на последующем уровне, т.е. первичная структура определяет вторичную, вторичная определяет третичную и т.д. (рис. 1).

Первичная структура белка, т.е. его аминокислотная последовательность, определяется экспериментальным путем относительно легко. Нахождение структуры высших порядков связано с большими сложностями, поскольку требует применения дорогих методов рентгеноструктурного анализа и магнитно-ядерного резонанса. Высокая стоимость экспериментального определения структуры белка обуславливает развитие математических методов ее предсказания. Проблема нахождения пространственной структуры белка по его аминокислотной последовательности является одной из важнейших неразрешенных проблем вычислительной биологии и биоинформатики.

В работах [2, 3] байесовская процедура распознавания на цепях Маркова применялась для определения вторичной структуры белка. В настоящей статье аналогичная процедура применяется для предсказания третичной структуры белка.

Постановка задачи. На вход задачи распознавания подается первичная структура — последовательность символов из алфавита A , $|A| = 20$. Каждый элемент последовательности соответствует аминокислотному остатку. Необходимо получить последовательность точек в пространстве R^3 , которые соответствуют координатам аминокислотных остатков. Длина аминокислотной последовательности (белка) находится в пределах от нескольких десятков до нескольких тысяч аминокислотных остатков.

© И.В. Сергиенко, Б.А. Белецкий, А.М. Гупал, 2010

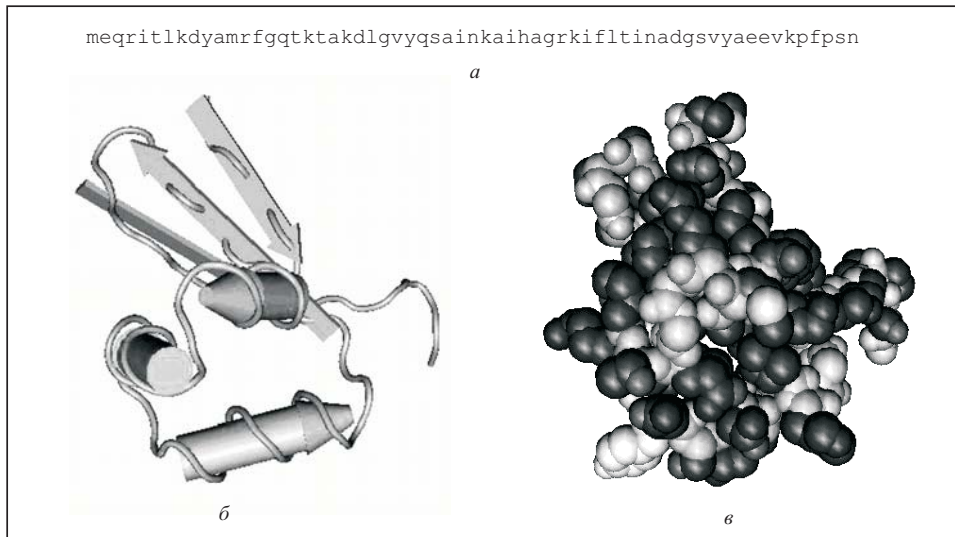


Рис. 1. Структура белка CRO: первичная (а); вторичная (б); третичная (в)

Естественным упрощением исходной задачи является построение третичной структуры белка на решетках. Расстояния между соседними остатками в белковой цепи одинаковы, что позволяет поместить каждый остаток в вершину решетки Z^3 и интерпретировать третичную структуру как путь на решетке. Пусть Ω — множество всех возможных путей, соответствующих белковой цепи с фиксированной первичной структурой. Для определения третичной структуры белка применяется так называемый термодинамический подход, заключающийся в использовании функции свободной энергии $E: \Omega \rightarrow R$. Если предположить, что белок в равновесном состоянии принимает конфигурацию с наименьшей свободной энергией, то задача построения третичной структуры сводится к нахождению пути

$$\omega^* = \arg \min_{\omega \in \Omega} E(\omega).$$

Функция энергии $E(\omega)$ выбирается, исходя из физических соображений. Например, в модели *HP* [4] аминокислотные остатки разделяются по своим химическим свойствам на два класса: гидрофобные (*H*) и гидрофильные (*P*). Гидрофобные остатки избегают контактов с молекулами воды и стремятся занять место в центре белковой глобулы, формируя ее ядро. Гидрофильные остатки, наоборот, контактируют с молекулами воды, находясь преимущественно на поверхности белковой глобулы. Исходя из этого, функция свободной энергии определяется в виде

$$E(\omega) = -n_{HH}(\omega),$$

где $n_{HH}(\omega)$ — количество контактов (*HH*) между гидрофобными остатками в конфигурации ω .

В настоящей статье делается попытка отыскать конфигурации ω^* на основании обучающих выборок без привлечения функции энергии $E(\omega)$. При этом путь ω задается как последовательность торсионных углов.

Торсионным (двугранным) углом $\tau(r_1, r_2, r_3, r_4)$, $r_1, r_2, r_3, r_4 \in R^3$, $\tau \in [-\pi, \pi)$, называется угол между нормальными к плоскостям, образованным тройками точек (r_1, r_2, r_3) и (r_2, r_3, r_4) (рис. 2).

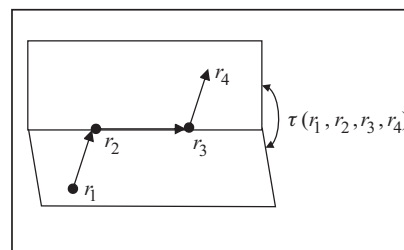
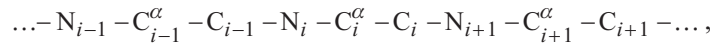


Рис. 2. Торсионный угол

Белковую молекулу можно представить в виде цепочки атомов



i -му аминокислотному остатку соответствуют атомы $N_i - C_i^{\alpha} - C_i$. Следует отметить, что в контексте распознавания вторичной структуры рассматриваются торсионные углы φ , ψ , ω , образованные четверками атомов $(C_{i-1}, N_i, C_i^{\alpha}, C_i)$, $(N_i, C_i^{\alpha}, C_i, N_{i+1})$ и $(C_i^{\alpha}, C_i, N_{i+1}, C_{i+1}^{\alpha})$ соответственно (рис. 3). Углы φ и ψ связа-

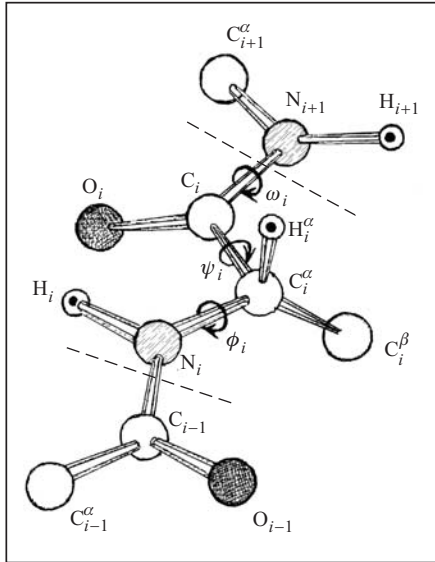


Рис. 3. Торсионные углы $\varphi_i, \psi_i, \omega_i$, образованные четверками атомов

ны с формированием вторичной структуры. Карты Рамачандрана [5] определяют множество допустимых значений пары торсионных углов $(\varphi, \psi) \in [-\pi, \pi] \times [-\pi, \pi]$ при заданном типе аминокислотного остатка и вторичной структуре, которой этот остаток принадлежит.

В данной статье рассматриваются торсионные углы, образованные атомами $(C_i^{\alpha}, C_{i+1}^{\alpha}, C_{i+2}^{\alpha}, C_{i+3}^{\alpha})$. Каждый атом C_i^{α} определяет положение в пространстве i -го аминокислотного остатка.

Разобьем множество значений торсионного угла $\tau \in [-\pi, \pi]$ на классы. Пусть X — множество классов,

$$X = \left\{ \left[-\pi, -\frac{\pi}{2} \right), \left[-\frac{\pi}{2}, 0 \right), \left[0, \frac{\pi}{2} \right), \left[\frac{\pi}{2}, \pi \right) \right\}.$$

Под нахождением торсионного угла будем понимать нахождение класса, которому принадлежит значение этого угла.

Таким образом, по первичной структуре $y = (y_1, \dots, y_n)$, $y_i \in A$, требуется найти последовательность торсионных углов $x = (x_1, \dots, x_{n-3})$, $x_i \in X$, на основе множества первичных структур с известными последовательностями торсионных углов.

Процесс нахождения всей последовательности $x = (x_1, \dots, x_{n-3})$ разбивается на решение $n-3$ подзадач, в каждой из которых независимо определяется один торсионный угол. При этом используется априорная информация о структуре совместного распределения $P(x_i, y)$, $i = 1, n-3$. Предполагается, что торсионный угол x_i , образованный четырьмя соседними остатками $a_i = (y_i, \dots, y_{i+3})$, зависит от этих остатков и остатков из некоторой окрестности радиуса k (рис. 4):

$$P(x_i | y_1, \dots, y_n) = P(x_i | y_{i-k}, \dots, y_{i+k+3}). \quad (1)$$

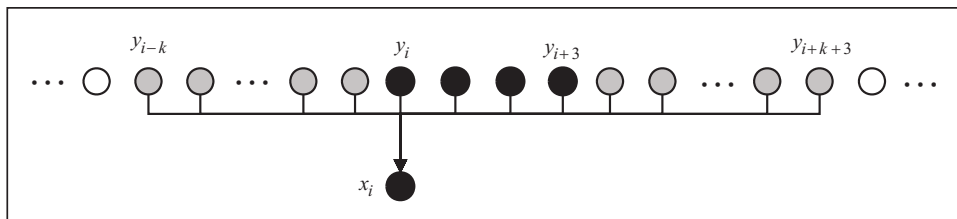


Рис. 4. Структура совместного распределения $P(x_i | y_1, \dots, y_n)$

Предполагается, что при фиксированных значениях x_i и четверки a_i для последовательностей оснований из окрестности радиуса k выполняется марковское свойство вида

$$P(y_{i-k}, \dots, y_{i+k+3} | x_i, a_i) = P_{i-k}(y_{i-k} | x_i, a_i) \prod_{j=i-k+1}^{i+k+3} P_j(y_j | y_{j-1}, x_i, a_i), \quad (2)$$

где $P_{i-k}(y_{i-k} | x_i, a_i)$ — вероятность значения y_{i-k} при фиксированных значениях x_i, a_i ; $P_j(y_j | y_{j-1}, x_i, a_i)$ — нестационарные переходные вероятности при фиксированных значениях x_i, a_i .

Для определения торсионного угла x_i необходимо найти значение

$$\arg \max_{x_i \in X} P(x_i, a_i | y_{i-k}, \dots, y_{i+k+3}).$$

Согласно формуле Байеса

$$P(x_i, a_i | y_{i-k}, \dots, y_{i+k+3}) = \frac{P(y_{i-k}, \dots, y_{i+k+3} | x_i, a_i) P(x_i, a_i)}{P(y_{i-k}, \dots, y_{i+k+3})}$$

и в силу (2) имеем

$$P(x_i, a_i | y_{i-k}, \dots, y_{i+k+3}) \sim P(x_i, a_i) P_{i-k}(y_{i-k} | x_i, a_i) \prod_{j=i-k+1}^{i+k+3} P_j(y_j | y_{j-1}, x_i, a_i).$$

В численных расчетах вероятности $P_j(y_j | x_i, a_i)$ заменяются оценками

$$\hat{P}_j(y_j | x_i, a_i) = \frac{k(y_j, x_i, a_i)}{\sum_{z \in Y} k(z, x_i, a_i)},$$

где $k(y_j, x_i, a_i)$ — число окружений четверки a_i соседними аминокислотами со значением y_j на позиции j при значении торсионного угла x_i , образованного четверкой остатков a_i .

Нестационарные переходные вероятности $P_j(y_j | y_{j-1}, x_i, a_i)$ заменяются оценками

$$\hat{P}_j(y_j | y_{j-1}, x_i, a_i) = \frac{k(y_{j-1}, y_j, x_i, a_i)}{\sum_{z \in Y} k(z, y_j, x_i, a_i)}, \quad (3)$$

где $k(y_{j-1}, y_j, x_i, a_i)$ — число окружений со значением y_j на позиции j , со значениями y_{j-1} на позиции $j-1$ при величине торсионного угла x_i , образованного четверкой остатков a_i .

Асимптотические свойства оценок вида (3) были изучены в работе [6]. Показано, что оценки асимптотически нормальны со средним нулем и дисперсией порядка $\frac{1}{\sum_{z \in Y} k(z, y_j, x_i, a_i)}$. Оценки $\hat{P}_j(y_j | y_{j-1}, x_i, a_i)$ асимптотически независимы для разных значений j .

Торсионному углу x_i присваивается значение

$$\arg \max_{x_i \in X} P(x_i, a_i | y_{i-k}, \dots, y_{i+k+3}).$$

В результате решения $n-3$ подзадач распознавания получаем искомую последовательность x .

Обучение и оценка точности. В качестве обучающей выборки используется множество порядка 20 тысяч негомولوجичных белков, сформированное с помощью методов выравнивания [7] (в свободном доступе на <http://www.rcsb.org/pdb/statistics/clusterStatistics.do>). Для проверки точности метода использовался сервер TESE [8] (в свободном доступе на <http://protein.bio.unipd.it/tese/>), который на основе методов выравнивания формирует репрезентативные тестовые наборы негомولوجичных белков. Для тестов использовались белки из трех структурных групп по класси-

фикации SATCN [9], состоящие в основном из α -спиралей (mainly alpha), из β -слоев (mainly beta), а также содержащие как α -спирали, так и β -слои (alpha beta). Для каждой структурной группы выбиралось по 10 белков с диапазоном длин 50–200, 200–500 и более 500 остатков. Точность определялась как частота правильного предсказания класса торсионного угла. Для сравнения различных реализаций метода между собой использовалась выборка из 10 белков структурного класса «alpha beta» с диапазоном длин 200–500 остатков. Точность распознавания торсионных углов с радиусом окружения $k=14$ для цепи Маркова первого порядка составила 0,75.

Перед тестированием из обучающей выборки удалялись белки, совпадающие с тестовыми или содержащие их в виде подпоследовательностей. На практике такая процедура не выполняется. Наоборот, один из наиболее успешных подходов к распознаванию структуры белка основывается именно на поиске гомологичных белков с известной структурой [10]. Без удаления повторов точность распознавания торсионных углов для радиуса окружения $k=14$ и цепи первого порядка составила 0,94.

Предсказание пар состояний. Точность предсказания можно повысить, если учесть зависимость между соседними торсионными углами x_i и x_{i+1} . Для этого процесс разбивается на подзадачи по предсказанию пар торсионных углов (рис. 5). Пара углов x_i, x_{i+1} образуется пятью соседними аминокислотными остатками $a_i = (y_i, \dots, y_{i+4})$.

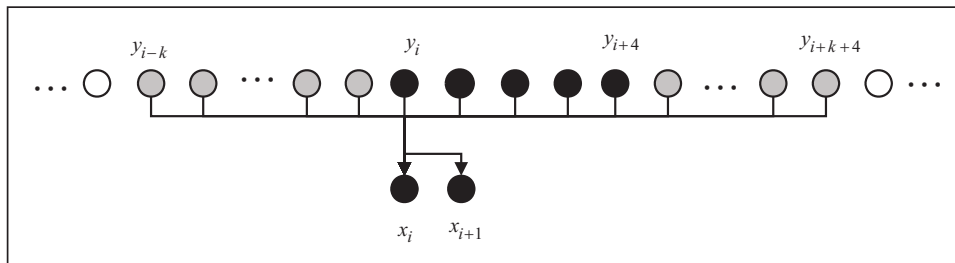


Рис. 5. Структура совместного распределения $P(x_i, x_{i+1} | y_1, \dots, y_n)$

Предполагается, что зависимость от окружения и нестационарное марковское свойство имеют место для пар x_i, x_{i+1} и формулируются аналогично (1), (2):

$$P(x_i, x_{i+1} | y_1, \dots, y_n) = P(x_i, x_{i+1} | y_{i-k}, \dots, y_{i+k+4}), \quad (4)$$

$$\begin{aligned} & P(y_{i-k}, \dots, y_{i+k+4} | x_i, x_{i+1}, a_i) \sim \\ & \sim P_{i-k}(y_{i-k} | x_i, x_{i+1}, a_i) \prod_{j=i-k+1}^{i+k+4} P_j(y_j | y_{j-1}, x_i, x_{i+1}, a_i). \end{aligned} \quad (5)$$

Применяя формулу Байеса и используя свойство (4), получаем

$$\begin{aligned} & P(x_i, x_{i+1}, a_i | y_{i-k}, \dots, y_{i+k+4}) \sim \\ & \sim P(x_i, x_{i+1}, a_i) P_{i-k}(y_{i-k} | x_i, x_{i+1}, a_i) \prod_{j=i-k+1}^{i+k+4} P_j(y_j | y_{j-1}, x_i, a_i). \end{aligned}$$

В численных расчетах вероятности $P_j(y_j | x_i, x_{i+1}, a_i)$ заменяются оценками

$$\hat{P}_j(y_j | x_i, x_{i+1}, a_i) = \frac{k(y_j, x_i, x_{i+1}, a_i)}{\sum_{z \in Y} k(z, x_i, x_{i+1}, a_i)},$$

где $k(y_j, x_i, x_{i+1}, a_i)$ — количество окружений со значением y_j на позиции j при величине пары торсионных углов x_i, x_{i+1} , образованных пятеркой остатков a_i .

Переходные вероятности $P_j(y_j | y_{j-1}, x_i, x_{i+1}, a_i)$ заменяются оценками

$$\hat{P}_j(y_j | y_{j-1}, x_i, x_{i+1}, a_i) = \frac{k(y_{j-1}, y_j, x_i, x_{i+1}, a_i)}{\sum_{z \in Y} k(z, y_j, x_i, x_{i+1}, a_i)},$$

где $k(y_{j-1}, y_j, x_i, x_{i+1}, a_i)$ — количество окружений со значением y_j на позиции j , и со значениями y_{j-1} на позиции $j-1$ при величине торсионных углов x_i, x_{i+1} , образованных пятеркой остатков a_i . Асимптотические свойства аналогичны (3).

В результате распознавания получаем последовательность распределений $\hat{P}_i(x_i, x_{i+1}, a_i | y_{i-k}, \dots, y_{i+k+4}), i = \overline{1, n-4}$. Поскольку первичная структура фиксирована, для краткости введем обозначение

$$P_i(x_i, x_{i+1}) := \hat{P}_i(x_i, x_{i+1}, a_i | y_{i-k}, \dots, y_{i+k+4}), i = \overline{1, n-4}.$$

Для каждого i переменная x_i (кроме $i=1$ и $i=n-4$) присутствует в распределении $P_i(x_i, x_{i+1})$ в качестве первого аргумента и вторым аргументом в распределении $P_{i-1}(x_{i-1}, x_i)$. Поэтому для построения последовательности x необходима процедура сведения результатов независимого распознавания пар торсионных углов x_i, x_{i+1} .

Опишем процедуру, показавшую наилучшую точность распознавания. Обозначим

$$P_i^-(x_i) := \sum_{x_{i-1} \in X} P_{i-1}(x_{i-1}, x_i), \quad P_i^+(x_i) := \sum_{x_{i+1} \in X} P_i(x_i, x_{i+1}).$$

Углу x_i присваивается значение

$$\arg \max_{x_i \in X} \ln P_i^+(x_i) + \ln P_i^-(x_i). \quad (6)$$

В табл. 1 приводятся результаты точности распознавания углов для трех структурных групп по всем диапазонам длин, а в табл. 2 — точности распознавания метода на тех же тестовых выборках и при таких же условиях, но без удаления повторов.

Таблица 1

Группа белков	Результаты точности распознавания углов по диапазонам длин		
	50–200 остатков	200–500 остатков	Более 500 остатков
Alpha beta	0.81	0.74	0.81
Mainly alpha	0.58	0.82	0.72
Mainly beta	0.74	0.76	0.76

Таблица 2

Группа белков	Результаты точности распознавания углов без удаления повторов по диапазонам длин		
	50–200 остатков	200–500 остатков	Более 500 остатков
Alpha beta	0.94	0.94	0.89
Mainly alpha	0.93	0.95	0.98
Mainly beta	0.94	0.91	0.94

Процедура сведения результатов оказывает большое влияние на конечную точность распознавания. В табл. 3 приводятся показатели точности независимого распознавания пар торсионных углов x_i, x_{i+1} до сведения результатов. Нетрудно убедиться, что процедура сведения результатов обеспечивает до 10 % результирующей точности метода.

Таблица 3

Группа белков	Результаты точности независимого распознавания пар углов по диапазонам длин		
	50–200 остатков	200–500 остатков	Более 500 остатков
Alpha beta	0.73	0.65	0.75
Mainly alpha	0.49	0.76	0.67
Mainly beta	0.61	0.66	0.66

Существует множество способов построения последовательности углов $x = (x_1, \dots, x_{n-4})$ с использованием последовательности распределений вида $P_i(x_i, x_{i+1}), i = \overline{1, n-4}$. Величины $-\ln P_i(x_i, x_{i+1})$ можно интерпретировать как расстояния и свести задачу построения x к нахождению кратчайшего пути между двумя вершинами на графе $G = (V, E, \rho)$ специального вида [11].

Пусть множество вершин V графа G состоит из вершин вида v_i^κ , $\kappa \in X$, $i = 1, \dots, n-3$, начальной вершины α и конечной вершины β :

$$V = \{v_i^\kappa \mid \kappa \in X, i = 1, \dots, n-3\} \cup \{\alpha, \beta\}.$$

Множество ребер графа G имеет вид

$$E = \{e_i(\kappa_1, \kappa_2) \mid i = 1, n-4, \kappa_1, \kappa_2 \in X\} \cup \{e_\alpha(\kappa) \mid \kappa \in X\} \cup \{e_\beta(\kappa) \mid \kappa \in X\},$$

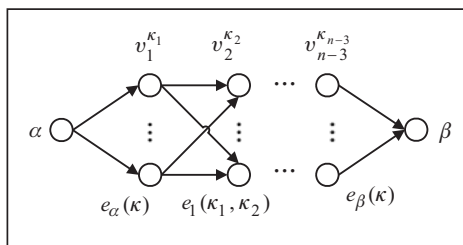


Рис. 6. Структура графа G

где $e_i(\kappa_1, \kappa_2)$ — дуга, соединяющая вершину $v_i^{\kappa_1}$ с вершиной $v_{i+1}^{\kappa_2}$; $e_\alpha(\kappa)$ — дуга, соединяющая вершину α с вершиной v_1^κ ; $e_\beta(\kappa)$ — дуга, соединяющая вершину v_{n-3}^κ с вершиной β (рис 6).

Зададим длины дуг $\rho(e)$, $e \in E$:

- 1) дуги, соединяющие начальную вершину α с вершинами v_1^κ , имеют нулевые длины $\rho(e_\alpha(\kappa)) = 0 \quad \forall \kappa \in X$;
- 2) дуги, соединяющие вершины v_{n-3}^κ , $\kappa \in X$, с конечной вершиной β , имеют нулевые длины $\rho(e_\beta(\kappa)) = 0 \quad \forall \kappa \in X$;
- 3) длины остальных дуг определяются как $\rho(e_i(\kappa_1, \kappa_2)) = -\ln P_i(\kappa_1, \kappa_2)$, $\kappa_1, \kappa_2 \in X$, $i = 1, \dots, n-4$.

Для определения последовательности торсионных углов x необходимо найти кратчайший путь от вершины α к вершине β . Эта задача эффективно решается методами динамического программирования. Получив кратчайший путь и отбросив вершины α и β , находим последовательность углов x . Точность такой модификации метода не позволила существенно улучшить результаты по сравнению с процедурой (6).

Построение третичной структуры. Получив последовательность торсионных углов x , можно построить путь на решетке Z^3 , соответствующий третичной структуре белка. Координата первого остатка r_1 выбирается произвольно. Далее делается два шага в произвольном направлении, чтобы зафиксировать точки r_2 и r_3 (см. рис. 2). Точку r_4 получаем вращением единичного вектора (r_3, r_4) вокруг оси (r_2, r_3) параллельно (r_1, r_2) на угол x_1 . Угол откладывается от плоскости (r_1, r_2, r_3) .

Заключение. Предложена процедура распознавания торсионных углов, образованных C^α атомами соседних аминокислотных остатков. На основе последовательности торсионных углов строится пространственная структура на решетке Z^3 . В будущем планируется сглаживать полученную структуру с помощью методов молекулярной динамики, используя такие приложения, как NAMD (<http://www.ks.uiuc.edu/Research/namd/>) или CHARMM (<http://www.charmm.org/>).

СПИСОК ЛИТЕРАТУРЫ

1. Practical lessons from protein structure prediction / K. Ginalski, N.V. Grishin, A. Godzik et al. // *Nucleic Acids Res.* — 2005. — **33**. — P. 1874–1891.
2. Сергиенко И.В., Белецкий Б.А., Васильев С.В., Гупал А.М. Предсказание вторичной структуры белков на основе байесовских процедур распознавания на цепях Маркова // *Кибернетика и системный анализ.* — 2007. — № 2. — С. 59–64.
3. Гупал А.М. Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.

4. Dill K. A. Theory for the folding and stability of globular proteins // *Biochemistry*. — 1985. — **24**, N 6. — P. 1501–1509.
5. Ramachandran G.N., Ramakrishnan C., Sasisekharan V. Stereochemistry of polypeptide chain configurations // *J. of Molecular Biology*. — 1963. — **7**. — P. 95–99.
6. Anderson T.W., Goodman L.A. Statistical inference about Markov chains // *The Annals of Mathematical Statistics*. — 1957. — **28**. — P. 89–110.
7. Basic local alignment search tool / S.F. Altschul, W. Gish, W. Miller et al. // *J. of Molecular Biology*. — 1990. — **215**, N 3. — P. 403–410.
8. Sirocco F., Tosatto S.C.E. TESE: Generating specific protein structure test set ensembles // *Bioinformatics*. — 2008. — **24**, N 22. — P. 2632–2633.
9. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution / L.H. Greene, T.E. Lewis, S. Addou et al. // *Nucleic Acids Research*. — 2007. — **35**. — P. 291–297.
10. McGuffin L.J. Protein fold recognition and threading // *Computational Structural Biology*. — World Scientific, 2008. — P. 37–60.
11. Шлезингер М. И., Главач В. Десять лекций по статистическому и структурному распознаванию. — Киев: Наук. думка, 2004. — 548 с.

Поступила 23.03.2010