УДК 577.112+371.24

С. З. Мальченко, Н. А. Чащин

ПРЕДСКАЗАНИЕ ВТОРИЧНОЙ СТРУКТУРЫ БЕЛКОВ

В работе рассмотрена проблема предсказания вторичной структуры белков. С помощью анализа ряда методов, относящихся к разным классам (статистические методы, методы сходства последовательностей, физико-химические методы и комбинированные методы), выявлены преимущества и ограничения существующих подходов к предсказанию вторичной структуры белков. Предлагается гипотеза о возможном направлении прогресса в этой области, основанная на выявлении и анализе информационных сайтов белков, разработке логических методов анализа вторичной структуры белков.

Проблема корректного предсказания вторичной структуры белков является одной из важнейших в молекулярной биологии и белковой инженерии. Решение ее позволит, с одной стороны, теоретически предсказывать и конструировать новые белки с заданной пространственной структурой, с другой — появится возможность, не прибегая к экспериментальной проверке, рассчитывать пространственную структуру природных белков и их модификаций. Это имеет определяющее значение как для белковой инженерии, так и в решении традиционных задач биологии, в частности, для изучения молекулярной эволюции различных семейств белков.

С точки зрения моделирования вторичную структуру можно представить в виде двух массивов данных — SEQ(i) и STR(i), где i — порядковый номер аминокислоты в последовательности. В массиве SEQ представлена информация о первичной структуре белка, в массиве STR (согласно данным рентгеноструктурного анализа белка) — о вторичной. Задача предсказания вторичной структуры заключается в поиске алгоритма (метода), с помощью которого можно добиться наибольшего соответствия между массивами SEQ и STR. Это даст возможность конструировать искусственные белки путем формирования аминокислотной последовательности, используя так называемый белковый код — множество правил, характерных для определения вторичной структуры (α-спираль, β-складка, нерегулярная структура, повторы). Несмотря на кажущуюся простоту постановки проблемы, за последние 10 лет точность методов предсказания вторичной структуры возросла на 10 % (с 55 до 65 %) [12]. Все известные на данный момент методы предсказания вторичной структуры можно разделить на четыре класса: 1) статистические методы; 2) методы, основанные на схожести (гомологии) последовательностей; 3) физико-химические методы; 4) комбинированные методы (экспертные системы).

В табл. 1 представлены данные о точности предсказания различных методов. Хотя спектр предсказанных алгоритмов довольно широк (от частотных методов до моделирования с помощью нейронной сети), заметна тенденция подхода к определенной границе точности (70 %), которую не удается перешагнуть [12]. Ряд авторов видит в этом объективные причины, которые будут рассмотрены ниже. Предварительно проанализируем подробнее несколько методов из вышеупомянутых классов.

Статистические методы базируются на применении стандартных математических процедур статистического анализа, таких как частоты встречаемости, вероятность, информационная теория и т. д.

Самым простым и удобным методом в этой группе по-прежнему остается метод Чоу — Фасмана [5, 6], основанный на вычислении частот вхождения аминокислотных остатков в разные конформации вторичной структуры. Анализ этих частот по последовательности белка с использованием ряда простых правил приводит к формированию дискретных участков вторичной структуры с той или иной конформацией. Хотя точность предсказаний на основе данного метода мала по срав-

нению с другими методами (см. табл. 1), он позволяет быстро и просто сделать предварительный анализ последовательности белка для нахождения его вторичной структуры. Достаточно простым и одним из самых точных методов предсказания является семейство GOR-методов [8, 9, 13]. Авторы этих работ заметили, что, хотя метод Чоу — Фасмана и отражает какую-то предпочтительность вхождения остатков в определенную вторичную структуру, но, может быть, из-за упрощенности используемого математического аппарата не дает нужного результата [12]. Поэтому, оставив в основе методов ту же идею — частоты вхождения в определенную вторичную структуру,— они использовали более сложный математический аппарат информационной теории.

Семейство GOR-методов. Методы основаны на применении информационной теории и суть их заключается в оценке вероятности нахождения каждой аминокислоты в каком-либо из конформационных состояний (α-спираль, β-складка, нерегулярная структура — H, E, C). С этой целью для каждой аминокислоты в каждой конформации высчитывается информационная функция I(S; R) по формулам: $I(S; R) = \ln P(S; R)/P(S)$, $P(S; R) = P(S; R)/P(R)$, где P(S) — вероятность конформации S в базе данных; P(S; R) — условная вероятность конформации S при условии, что остаток R присутствует в ней; ln — натуральный логарифм, и где P(S; R) — вероятность совместного события — остаток R в конформации S; P(R) — вероятность наблюдения остатка R.

На основе обучающей базы данных формируются таблицы со значениями I(S; R) для трех конформаций: α-спираль, β-складка и нерегулярная структура для всех остатков. Для каждого остатка высчитываются информационные функции во всех конформациях. С учетом наибольшего значения делается вывод о принадлежности остатка к той или иной конформации. Повторив эту процедуру для всех аминокислот белка, получают коэффициенты, отражающие вероятность различных участков белка находиться в какой-либо конформации. Существенным моментом в предсказании является оценка разницы между максимальным значением и вторым максимальным значением информационной функции. Под максимальным значением подразумевается наибольшее значение информационной функции I(S; R) для определенного остатка, а под вторым максимальным значением — следующее за максимальным значение информационной функции I(S; R). Если

Таблица 1

Точность предсказания, рассчитанная для трех конформаций в современных методах предсказания вторичной структуры белков (данные частично взяты из [12])

Класс методов	Точность предсказания, %	Ссылка
Статистические методы		
Чоу — Фасман	51	[5, 6]
GOR	55	[8]
GORIII	63,3(67)	[13]
Шестопалов	59	[2]
Методы сходства последовательностей		
Нишикава и Оои	60	[22]
Свит	59	[29]
Левин и Гарниер	63(67)	[19]
Физико-химические методы		
Филькенштейн — Птицын	63	[24, 25]
Бью и др.	59	[3]
Лим	56	[20]
Холли и Карплюс	63,2	[14]
Комбинированные методы		
Нейронная сеть	64,3	[26]
COMBINE	65,5	[3]
Нишикава и др.	67,4	[23]
Соловьев и др.	63	[1]

эта разница существенна, то вероятность нахождения данного остатка в данной конформации очень высока.

Самым простым из этого семейства является GOR-метод [8], последней же разработкой — GORIII-метод [13], в котором расчет информационной функции производится с учетом всех остатков из окружения ±8 аминокислот вокруг изучаемой аминокислоты.

Метод дублетного кода [2] основан на анализе распределения пар аминокислотных остатков во вторичной структуре белка.

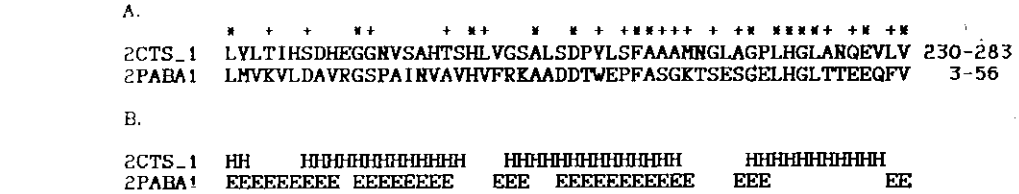

Рис. 1. Пример сравнения вторичных структур схожих последовательностей [18]. Схожие последовательности имеют совершенно различную вторичную структуру. В схеме используются следующие обозначения: 2CTS-1 — свиная цитратсинтетаза; 2PABA1 — человеческий преальбумин; H — α-спираль; E — β-складка; * — идентичные аминокислоты; + — консервативные замены; A — сравнение первичных последовательностей; B — сравнение соответствующих вторичных структур

Выдвигается гипотеза о том, что α-спирали, β-складки и повторы составлены из элементов — структуронов — длиной 5, 3 и 2 остатка соответственно.

Далее предполагается, что структуроны закодированы в последовательности белка дублетами, соответствующими крайним положениям структуронов. На основе анализа базы данных из 38 белков были построены таблицы дублетного кода для спиралей, складок и поворотов. Используя эти таблицы и ряд эмпирических правил отбора между конформациями, можно предсказать структуру неизвестного белка с точностью 59 %. Хотя практическая ценность данного метода мала, однако интересно, что с его помощью показано определенное предпочтение дублетов разными конформациями и вероятность самоорганизации вторичной структуры (наличие структуронов).

Методы сходства последовательностей. Все методы данной группы базируются на простой гипотезе — короткие пептиды с высокой степенью гомологии первичной последовательности должны иметь схожую вторичную структуру. Различие этих методов состоит только в использовании разных рамок сравниваемых участков и разных матриц схожести остатков. В методе [22] используется рамка из 11 остатков и матрица сходства физических и химических свойств, в методе [29] — рамка из 12 остатков и матрица Дайхофф; в методе [19] — рамка из 17 остатков и оптимизированная матрица схожести остатков. Однако всем этим методам присущи следующие недостатки.

1. Показано, в частности, что идентичные пентапептиды могут находиться в разных конформациях. Из всего возможного множества пентапептидов (20^5) было обнаружено 13 разновидностей, встречающихся более чем в одной конформации [18].

2. В Брухевенском банке данных, содержащем 143 белка (23 351 аминокислотный остаток) с рентгеноструктурным разрешением менее 0,25 нм, определены 146 схожих последовательностей длиной от 20 до 262 остатков, среди которых только 41 % последовательностей имели совпадающую вторичную структуру. Показано, что в общем выравненные последовательности не имели совпадения в пространственной структуре (рис. 1) [28].

3. Методы сходства последовательностей предполагают произведение процедуры выравнивания последовательностей. Разнообразие современных методов ставит вопрос о поиске оптимального метода выравнивания, от результатов которого зависит предсказание вторичной структуры.

Несмотря на описанные недостатки, эти методы дают высокие результаты в предсказании вторичной структуры гомологичных белков (75—90 % точности), однако для негомологичных белков точность предсказания по-прежнему не превышает 60—65 %.

Метод локальной гомологии [14]. Алгоритм предсказания вторичной структуры в этом методе состоит из двух частей. В первой части с использованием матрицы сходства вторичных структур (рис. 2) производится поиск в базе данных по вторичным структурам белка, гомологичного предсказываемому белку. Во второй части алгоритма используется следующая стратегия: начиная с первой позиции предсказываемого белка, выделяются участки длиной 17 остатков; участки такой же длины выделяются из белков, входящих в базу дан-

G	2																		
P	1	3																	
D	0	0	2																
E	0	-1	1	2															
A	0	-1	0	1	2														
N	0	0	1	0	0	3													
Q	0	0	0	1	0	1	2												
S	0	0	0	0	1	0	0	2											
T	0	0	0	0	0	0	0	0	2										
K	0	0	0	0	0	1	0	0	0	2									
R	0	0	0	0	0	0	0	0	0	1	2								
H	0	0	0	0	0	0	0	0	0	0	0	2							
V	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	2						
I	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	2					
M	-1	1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	0	0	2				
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2			
L	-1	-1	-1	-1	0	-1	-1	-1	0	-1	-1	-1	1	0	2	0	2		
F	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	1	0	-1	0	2	
Y	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	0	0	0	-1	0	1	2
W	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	0	0	0	-1	0	0	0	2
	G	P	D	E	A	N	Q	S	T	K	R	H	V	I	M	C	L	F	Y	W

Рис. 2. Оптимизированная матрица схожести вторичных структур белков [14]. Коэффициенты в матрице схожести определяют следующие свойства остатков для вторичной структуры: положительные — функциональное совпадение; ноль — отсутствие функциональной схожести; отрицательные — полное различие свойств

ных по вторичным структурам и гомологичных предсказываемому. Далее, сравнивают выделенные участки друг с другом, применяя рамку сравнения из семи остатков, при этом за каждый совпадающий остаток из рамки дается 1 балл участку из 17 остатков. Например, из базы данных выделен участок из 17 остатков, гомологичный предсказываемому и имеющий следующие первичную (1) и вторичную (2) структуры:

A L P G C T A L E G A V . . . . (1)

C C E E E E H H H H H H . . . . (2)

Предсказываемый участок имеет следующую первичную структуру — SLCGTVADLK... . Из него с первой позиции выделяется участок из семи остатков и сравнивается с участком из базы данных. Потом выделяется участок со второй позиции, третьей и т. д.

Шаг 1: A L P G C T A L E G A V . . .

S L C G T V A 3 балла

Шаг 2: A L P G C T A L E G A V . . . .

S L C G T V A 1 балл

Шаг 3: A L P G C T A L E G A V . . . .

S L C G T V A 1 балл и т. д.

В результате формируется таблица баллов для всех остатков в различных конформациях. Остатку присваивается та конформация, в ко-

торой он набрал наибольшее количество баллов:

Остаток	H	E	C	Предсказание
S	1	0	5 (3 + 1 + 1)	C
L	0	0	3	C
C	0	6 (3 + 3)	3 (2 + 1)	E
.	.	.	.	.
.	.	.	.	. и т. д.

Данный метод тестирован на базе данных из 67 белков с разрешением лучше 0,28 нм, включающей 12 058 аминокислот (27 % H, 22 % E, 51 % C). Точность предсказания данного метода составляет в общем

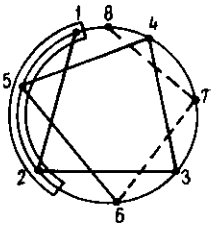

Рис. 3. Серия остатков, составляющих гидрофобную и гидрофильную стороны α-спирали. В модели остатки 1, 2, 5 аппроксимируются как гидрофобная сторона (выделенный сектор), остатки 3, 4, 7 — как гидрофильная сторона; остатки 6, 8 — как переходные к следующим зонам (сериям остатков) [3]

63 %; по отдельным конформациям H — 58 %; E — 54 %, C — 68 %. Для белков, гомологичных более чем на 30 %, точность составила 70 % и выше.

Физико-химические методы. В этой группе методов реализуются различные физико-химические модели вторичной структуры. Часть методов развилась из следующей модели конформации [27]: все остатки разделены на гидрофобные и гидрофильные, а вторичная структура представлена в виде белковой нити, обвивающейся вокруг гипотетического цилиндра. В этом случае одна из сторон цилиндра является гидрофобной, вторая — гидрофильной. При изучении базы данных определяются правила, позволяющие реализовать данную физическую модель (см., например, [20]). Один из последних методов данного класса более подробно рассмотрим ниже.

Метод поиска паттернов [3] основан на предсказании зон, похожих на регулярные участки вторичной структуры. Известно, что обычно α-спираль состоит из гидрофобной и гидрофильной сторон, одна из которых контактирует с белковым кором, а вторая — с раствором. Образует эти стороны серия гидрофобных и гидрофильных остатков (рис. 3). Все остатки в этом методе разделены на три группы по признаку гидрофобности: 1) K, R, E, D, Q, N, T, S, P — гидрофильные (двоичный индекс=0); 2) F, C, I, L, W, V, M, V, H — гидрофобные (двоичный индекс=1); 3) A, G — могут быть как гидрофильными, так и гидрофобными (двоичный индекс может быть 0 или 1).

Анализируемая зона содержит восемь остатков. Каждому остатку в этой зоне присваивается двоичный индекс 0 или 1, который потом переводится в десятичный индекс. После анализа десятичных индексов делается вывод о принадлежности зоны к α-спирали или β-складке. Например:

2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	двоичные индексы;
1	2	4	8	16	32	64	128	десятичные индексы;
V	L	E	Q	A	L	S	T	предсказываемая последовательность;
1	1	0	0	0	1	0	0 = 35	преобразование при A=0 (1-й паттерн);
1	1	0	0	1	1	0	0 = 51	преобразование при A=1 (2-й паттерн).

Второй паттерн похож на α-спираль, так как он образует гидрофобные и гидрофильные стороны, в то время как первый — нет, по-

скольку гидрофобная Ala окружена гидрофильными остатками. Далее десятичные индексы полученных паттернов сравниваются с таблицей десятичных индексов паттернов, похожих на α-спираль: 9, 12, 13, 17, 18, 19, 25, 27, 29, 31, 34, 36, 38, 44, 45, 46, 47, 50, 51, 54, 55, 59, 61, 62, 77, 201, 205, 217, 219, 237.

Десятичный индекс 51 есть в приводимой таблице, значит, изучаемая последовательность VLEQALST предсказывается как α-спираль.

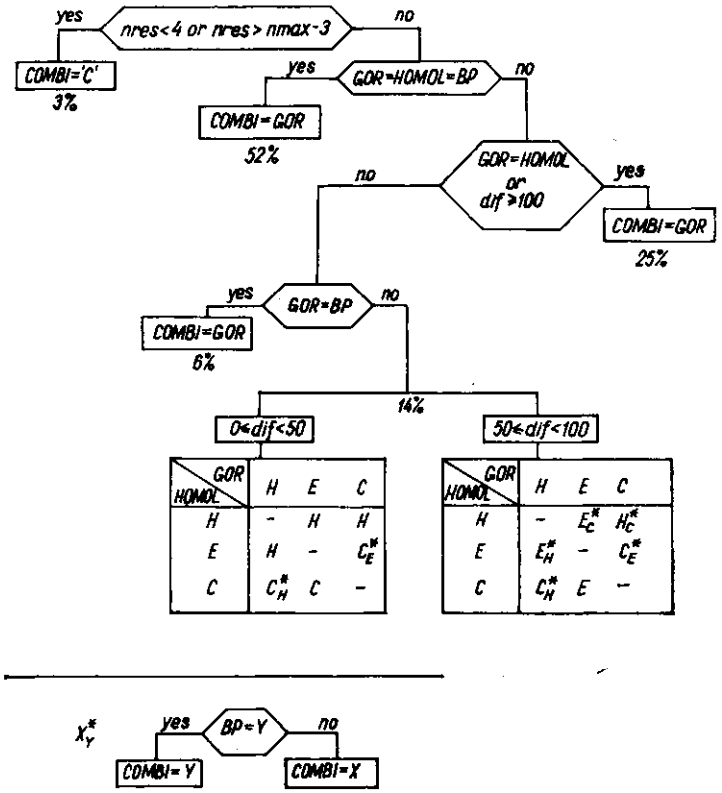

Рис. 4. Общая схема экспертной программы COMBINE [3, 9]. В схеме используются следующие обозначения: GOR — GORIII-метод [13]; HOMOL — метод локальной гомологии [14]; BP — метод поиска паттернов [3]; % — пропорция остатков в базе данных, вовлеченных в предсказание COMBINE; DIF — различие в информационном уровне предсказываемой конформации и вторым информационным уровнем; n_{max} — количество остатков в белковой цепи; n_{res} — позиция текущего предсказываемого остатка в белковой цепи

Кроме этого, в методе учитывается влияние Gly и Pro, которые имеют тенденцию разрушать α-спираль. Аналогичным образом рассчитываются β-складки. Точность предсказания метода в общем составляет 59 %, а по отдельным конформациям 54,5 % для H, 23,3 % для E и 76,5 % для C.

Комбинированные методы (экспертные системы). Рассмотрим наиболее сложные из методов предсказания вторичной структуры. Методы [3, 23] основаны на объединении нескольких более простых методов, при совмещении которых выбирается наиболее вероятный (согласующийся со всеми использованными методами) результат предсказания вторичной структуры. Метод [1] использует как можно более полный список отдельных свойств остатков, анализ которых приводит к выбору наиболее информативных свойств, используемых в дальнейшем для предсказания. Метод [26] реализует модель нейронной сети, способной к самообучению и запоминанию связей в обучающей выборке и поиску этих связей в неизвестном белке. Однако даже настолько сложные методы анализа вторичной структуры не поднимаются в точности предсказания выше 70 % (см. табл. 1).

Метод COMBINE [3]. В этом методе использованы ранее предложенные разработки — GORIII-метод [11], метод локальной гомологии [19] и метод поиска паттернов [3]. COMBINE объединил лучшие свойства предыдущих методов: хорошую эффективность предсказания коротких белков (менее 100 аминокислот) метода поиска паттернов и больших белков (более 200 аминокислот) GORIII-метода. Метод реализуется в три этапа.

1. Используя GORIII-метод вычисляются информационные уровни трех конформаций (α-спираль, β-складка и нерегулярная) для всех остатков в предсказываемом белке. Разницу между самым высоким информационным уровнем и следующим информационным уровнем (DIF) используют для определения дальнейшего пути предсказания. Если DIF$>$100, тогда предсказание в COMBINE сводится к предсказанию GORIII-метода. В другом случае происходит разветвление (рис. 4) и предсказание производится несколькими методами сразу.

2. Если в регулярной вторичной структуре X встречается остаток в другой конформации Y, например XYXXX, XXYXX или XXXYX, то остатку Y присваивается конформация X.

3. Этот этап реализуется при предсказании β-складчатых зон. В этом случае высчитывается уровень DIF для трех остатков в данной зоне и, если DIF$<$50 и вторая предсказанная конформация для граничного остатка β-складка, тогда COMBINE присваивает этому остатку конформацию β-складки. То же производится для второго и третьего остатков.

Метод COMBINE дает более высокую точность предсказания по сравнению с используемыми в нем методами — 65,5 %, в то время как GORIII — 63,3 %, метод локальной гомологии — 62,9 %, метод поиска паттернов — 59 %. Точность предсказания также зависит и от длины белка. Для белков длиной до 100 аминокислот — 73,1 %, 100—200 аминокислот — 66,2 % и более 200 аминокислот — 61,2 %.

Метод поиска наиболее значимых информационных характеристик [1]. Для работы в этом методе используется большой набор отдельных свойств аминокислот (полярность, объем боковой группы, поперечное сечение, заряд и т. д.), а также учитываются локальные, средние и дальние взаимодействия в белковой молекуле. Применяя дискриминационный анализ, производится поиск из всего набора используемых свойств для остатков в той или иной конформации наиболее информативных признаков. Далее белок анализируется полученными информативными признаками и предсказывается его вторичная структура. Точность предсказания при учете признаков отдельных аминокислот составляет 57 %, при учете специфических дипептидов — 63 %. Точность локализации протяженных α-спиралей и β-структур достигает 90 %.

Метод обобщенного предсказания [23] включает в себя несколько различных способов предсказания и собирает вместе лучшие варианты индивидуальных методов. Для обобщенного метода были выбраны пять различных методов ([9, 13, 21, 22, 26]), имеющих в среднем точность предсказания 60—65 % по трем конформациям: α-спираль, β-складка и нерегулярная структура. Общее повышение точности по сравнению с индивидуальными методами было небольшим, однако при предсказании структуры гомологичных белков она возрастала до 70 %. В работе было подтверждено, что точность предсказания возрастает с увеличением числа добавочных данных по гомологичным последовательностям, при этом лучшие предсказания были достигнуты для белков, у которых 2/3 последовательностей были гомологичны таковым базовых белков.

Метод нейронной сети [26]. В ряде работ [4, 15—17] разработаны теория и модель нейронной сети, состоящей из большого количества нейронов (компьютерных единиц). В работе [26] сделано приложение данной модели к предсказанию вторичной структуры. Суть заключается в определении общих свойств обучающего множества, фор-

мировании нейронных цепей при определенной ситуации и их узнавании на другом множестве данных.

В результате работы метода на выходе образуется скрытая способность определенного участка белка находиться в определенной конформации, которая затем преобразуется в кривую, отражающую наиболее вероятную вторичную структуру данного участка белка. Возможные погрешности отсекаются использованием минимального фактора погрешности, полученного из обучающего множества белков.

Итак, рассмотрев на нескольких примерах различные методы предсказания вторичной структуры белков, можно заметить, что все они при многообразии гипотез, подходов и алгоритмов к предсказанию вторичной структуры, имеют ограниченную точность — 70 % (см. табл. 1). Бóльшая точность предсказания (90—95 %), как правило, достигается только на обучающем множестве белков [14]. Объективные причины этого явления могут быть следующими.

1. Огромное количество возможных белковых последовательностей. Например, для пептида длиной 17 остатков количество возможных белковых последовательностей составляет $20^{17}=1{,}3\cdot10^{22}$. Даже принимая во внимание тот факт, что в процессе эволюции организмов не все возможные комбинации были отобраны, количество их все равно остается огромным [12].

2. Использование небольших баз данных белков с расшифрованной рентгеноструктурным анализом вторичной структурой. На сегодняшний день наибольшая база данных белков, вторичная структура которых определена с точностью менее 0,2 нм, удовлетворительной для методов предсказания вторичной структуры, включает 110 белков с общим количеством аминокислотных остатков около 19 000.

3. Работа с неспециализированными базами данных. При предсказании вторичной структуры все белки можно условно разделить на три группы с различной точностью предсказания: низкой (40—50 %), средней для каждого конкретного метода, и высокой (80—90 %). Показано, что для некоторых белков существующие методы предсказания вторичной структуры не подходят вообще [7]. Возможно, это происходит благодаря наличию различных механизмов сборки белков.

Можно предположить, что, разделив обучающую базу на несколько подбаз (по точности предсказания), вероятно получить увеличение точности в этих подбазах и соответственно в основной базе данных. Однако при таком подходе сразу возникает проблема выбора подбазы для данного предсказываемого белка.

4. До сих пор не существует непредубежденной базы данных для тестирования предсказательных методов [7]. Возможно, после создания такой базы будет достигнуто развитие объективных процедур предсказания и вскрыты некоторые общие трудности, характерные для всех современных методов предсказания вторичной структуры.

5. Учет только ближних взаимодействий аминокислот. Как показано в работе [19], рамка окружения изучаемого остатка длиной восемь аминокислот достаточна и расширение ее не приводит к существенному увеличению точности предсказания. Хотя в этой же работе обращено внимание на то, что в современных методах не учитываются дальние взаимодействия аминокислот в белке. На сегодняшний день методы предсказания с использованием схожести последовательностей частично решают данную проблему (по-видимому, за счет этого и происходит увеличение точности в комбинированных методах, использующих гомологию последовательностей). Однако она требует дальнейшего изучения.

6. Возможная деградация белкового кода с течением времени [12]. Это предположение частично подтверждает разброс в точности предсказания в обучающей базе данных (см. пункт 3).

7. Отсутствие единого взгляда (теории) на сборку пространственной структуры белков и правил, определяющих эту сборку.

Анализируя современные методы предсказания вторичной структуры, Гарниер и Левин [12] пришли к выводу о том, что в настоящий момент проблема предсказания вторичной структуры себя исчерпала, и поднять точность предсказания возможно только при изучении третичной структуры белков и их сборки. К аналогичным выводам пришли Соловьев и соавт. [1], суммируя полученные в настоящее время результаты в следующем принципе: окончательная локализация α-спиралей и β-структур определяется при формировании белковой глобулы и зависит от тонких стереохимических взаимодействий в третичной структуре белка.

Однако, как нам кажется, этот вывод несколько преждевременный. Детальный анализ существующих методов показывает, что нет ни одного из них, в котором применяют достаточно разработанный логический анализ баз данных, а также, несмотря на очевидное предположение о самоорганизации вторичной структуры, ни в одном из методов не используется это предположение достаточно глубоко. Обобщая данные, полученные разными методами, можно выделить такие закономерности.

1. Имеется определенная предпочтительность аминокислот (табл. 2), дуплетов [2] и участков белка [18, 28] находиться в определенной конформации.

2. Из результатов исследований следует, что встречаются участки белка, несущие определенную информационную нагрузку в детерминации вторичной структуры. В работе [1] показано, что информационные характеристики N- и С-концевых остатков дискретных участков вторичной структуры отличаются от таковых остатков, находящихся в середине подобных участков. В работе [7] сделан вывод о том, что внутренние остатки в дискретных участках вторичной структуры более детерминированны, чем остатки по их границам. В работе [28] речь идет уже об «индикаторах» определенной конформации.

3. В основе современных методов предсказания вторичной структуры лежит следующая схема: предсказание вторичной структуры индивидуальных остатков, анализ соседних остатков и на основе этих данных вывод о конформации целого сегмента белка.

Исследование полученных закономерностей приводит к вероятному направлению прогресса в методах предсказания вторичной структуры по схеме: отход от изучения вхождения индивидуальных остатков в определенную конформацию вторичной структуры и переход на поиск и изучение информационных участков-индикаторов. Возможно, существует целый класс таких структур по примеру промоторов, терминаторов и регулирующих сайтов в ДНК. Изменение стратегии предсказательных методов (см. пункт 3): переход от анализа индивидуальных остатков к анализу белковых блоков предполагает, как нам кажется, использование логических методов анализа. Определенный прогресс, вероятно, будет достигнут и в работе со специализированными базами данных, например α-, β-, αβ-белков, суперсемейств, стабильных и нестабильных белков и т. д., а также в изучении и моделировании процесса сборки белков [12].

Таблица 2

Данные о предпочтительности нахождения аминокислот в определенной конформации вторичной структуры [7]

Конформация	Аминокислоты		
	Предпочтительные	Индифферентные	Разрушающие
α-спираль	ALMHEQKC	VIFWDNR	YTGSP
β-складка	VIFYTW	ALMHGSR	EQKDNPC
Реверсивный поворот	GSDNP	GQKYTR	ALMHVIFTCR

Сегодня уже очевиден факт вырожденности белкового кода, когда правила вхождения аминокислот в ту или иную конформацию неоднозначно отвечают их реальной встречаемости в соответствующей вторичной структуре. Возможно, логический анализ информационных сайтов белков позволит скорректировать данные правила, определить более значимые и привести к существенному прогрессу в области предсказания вторичной структуры белков.

Summary. The present state of protein secondary structure prediction is discussed. The benefits and limitations of usually used approaches for the protein secondary prediction structure were determined on the base of analysis of various prediction methods (statistical, physico-chemical, combine methods and methods of sequences similarity). The hypothesis on possible progress in this area is proposed. This idea is to reveal and analyze the informative protein sites and to elaborate the logical methods for analysis of protein secondary structure.

СПИСОК ЛИТЕРАТУРЫ

1. *Соловьев В. В., Саламов А. А., Салиханова А. К.* Компьютерная система для исследования структурной организации глобулярных белков // Компьютерный анализ структуры, функции и эволюции генет. макромолекул: пробл. интеллектуализации.— Новосибирск, 1989.— С. 111—154.
2. *Шестопалов Б. В.* Алгоритм предсказания вторичной структуры белков по методу дублетного кода // III Всесоюз. совещ. «Теорет. исследования и банки данных по молекуляр. биологии и генетике» : Тез. докл.— Новосибирск, 1988.— С. 57—60.
3. *Secondary* structure prediction: combination of three different methods / V. Biou, J. F. Gibrat, J. M. Levin et al. // Prot. Engng.—1988.—**2**, N 3.— P. 185—191.
4. *A novel* approach to prediction of the 3-dimensional structures of protein backbones by neural network / H. Bohr, J. Bohr, S. Brunak et al. // FEBS. Lett.—1990.—**261**, N 1.— P. 43—46.
5. *Chou P. Y., Fasman G. D.* Prediction of protein conformation // Biochemistry.—1974.—**13**.— P. 222—245.
6. *Chou P. Y., Fasman G. D.* Prediction of the secondary structure of protein from their amino acid sequence // Adv. Enzymol.—1978.—**47**, N 1.— P. 45—48.
7. *Computational* molecular biology: Sources and methods for sequence analysis / Ed. Ar. M. Lesk.— Oxford : Univ. press, 1988.— 537 p.
8. *Garnier J., Osguthorp D. J., Robson B.* Analysis of the accuracy and implications of simple method for predicting the secondary structure of globular proteins // J. Mol. Biol.—1978.—**120**, N 1.— P. 97—120.
9. *Garnier J., Robson B.* The GOR method for predicting secondary structures in proteins. Prediction of protein structure and the principles of protein conformation / Ed. G. D. Fasman.— New York: Plenum publ. press corporation, 1989.— P. 417—465.
10. *Garnier J.* Protein structure prediction // Biochimie.—1990.—**72**, N 1.— P. 513—524.
11. *Secondary* structure prediction and protein desing / J. Garnier, J. M. Levin, J. F. Gibrat, V. Biou // Biochem. Soc. Symp.—1989.—**57**.— P. 11—24.
12. *Garnier J., Levin J. M.* The protein structure code: what is its present status? // CABIOS.—1991.—**7**, N 2.— P. 133—142.
13. *Gibrat J. F., Garnier J., Robson B.* Futher developments of protein secondary structure prediction using informational theory // J. Mol. Biol.—1987.—**198**, N 1.— P. 425—443.
14. *Holley H. W., Karplus M.* Protein secondary structure prediction with a neural network // Proc. Nat. Acad. Sci. USA.—1989.—**86**, N 1.— P. 152—156.
15. *Hopfield J. J.* Neural networks and physical system with emergent collective computational abilities // Ibid.—1982.—**79**.— P. 2554—2558.
16. *Hopfield J. J.* Neurons with graded response have collective computational properties like those of two-state neurons // Ibid.—1984.—**81**.— P. 3088—3092.
17. *Hopfield J. J., Tank D. W.* Computing with neural circuits : a model // Science.—1986.—**233**.— P. 625—633.
18. *Kabsch W., Sander C.* On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations // Proc. Nat. Acad. Sci. USA.—1984.—**81**, N 4.— P. 1075—1078.
19. *Levin J. M., Garnier J.* Improvements in a secondary structure prediction method on a search for local sequence homologies and its use as a model building tool // Biochim. et biophys. acta.—1988.—**955**, N 1.— P. 283—295.
20. *Lim V. I.* Algorithms for prediction of α-helices and β-structural regions in globular proteins // J. Mol. Biol.—1974.—**88**.— P. 873—894.
21. *Nagano K.* Triplet information in helix prediction applied to the analysis of supersecondary structure // Ibid.—1977.—**109**, N 2.— P. 251—274.

22. *Nishikawa K., Ooi T.* Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods // Biochim. et biophys. acta.— 1986.— **871**, N 1.— P. 45—54.
23. *Nishikawa K., Noguchi T., Konishi Y.* Highly reliable secondary structure prediction of proteins by a new joint method // Prot. Engng.—1990.—**3**, N 4.— P. 283—284.
24. *Ptitsyn O. B., Finkelstein A. V.* Theory of protein secondary structure and algorithm of its prediction // Biopolymers.—1983.—**22**, N 1.— P. 15—25.
25. *Ptitsyn O. B., Finkelstein A. V.* Prediction of protein secondary structure based on physical theory. Histones // Prot. Engng.—1989.—**2**.— P. 443—447.
26. *Qian N., Sejnowski T. J.* Predicting the secondary structure of globular proteins using neural network models // J. Mol. Biol.—1988.—**202**.— P. 865—884.
27. *Schiffer M., Edmundson A. B.* Use of helical wheels to represent the structures of protein and a identify segments with helical potential // Biophysics.— 1967.— **7**, N 1.— P. 121—135.
28. *Sternberg M. J. E., Islam S. A.* Local protein sequence similarity does not imply a structural relationship // Prot. Engng.—1990.—**4**, N 2.— P. 125—131.
29. *Sweet R. M.* Evolutionary similarity among peptide segments is a basis for prediction of protein folding // Biopolymers.—1986.—**25**.— P. 1565—1577.

Ин-т молекуляр. биологии и генетики
АН Украины, Киев

Получено 22.05.92

УДК 577.344

Л. В. Карабут, А. А. Серейская

ОСОБЕННОСТИ СТРОЕНИЯ КОНТАКТНОЙ ЗОНЫ ТРОМБИНА. ВОЗМОЖНЫЕ МЕХАНИЗМЫ ВЗАИМОДЕЙСТВИЯ ЕГО С НИЗКО- И ВЫСОКОМОЛЕКУЛЯРНЫМИ СУБСТРАТАМИ

Показаны особенности строения активного центра тромбина и его вторичного связывающего участка. Рассмотрены альтернативные механизмы взаимодействия тромбина с низко- и высокомолекулярными субстратами, а также возможная роль так называемого дополнительного центра в образовании продуктивного фермент-субстратного комплекса.

Тромбин (КФ 3.4.21.5) — ключевой фермент системы свертывания крови, является сериновой протеиназой трипсиноподобного действия. В ходе субстратного и ингибиторного анализа нативного и модифицированного тромбина были обнаружены характерные черты строения активного центра (АЦ) фермента и особенности взаимодействия его с различными лигандами [1—3]. Полученные результаты нашли свое подтверждение и дальнейшую детализацию при рентгеноструктурном анализе (РСА) комплекса тромбина с низкомолекулярным ингибитором [4]. В работах [5, 6] обсуждалась роль так называемого дополнительного центра (ДЦ) тромбина в эффективном гидролизе специфического высокомолекулярного субстрата — фибриногена. Конструирование и рассмотрение атомно-молекулярной и скелетно-проволочной модели фермента [7] позволило нам сформулировать некоторые гипотезы, касающиеся сопряжения указанного дополнительного центра с активным центром.

На основании этих данных можно представить себе структуру АЦ фермента следующим образом. Субцентры фермента, т. е. участки белка, взаимодействующие с конкретным аминокислотным остатком субстрата, формируются за счет нескольких аминокислотных остатков белковой молекулы, причем один и тот же остаток может участвовать в создании различных субцентров. Например, его боковой радикал может находиться в составе одного субцентра, а участок главной цепи — в составе другого.

Карман первичного связывания тромбина входит в субцентр S_1. На его дне находится остаток Асп189, боковой радикал которого своим карбоксилом ориентирован навстречу гуанидиновой группе остатка