БИООРГАНИЧЕСКАЯ ХИМИЯ

Том 20 * № 11 * 1994

УДК 577.112.5.087

© 1994 *И. И. Исмаилов, М. В. Шаров, З. У. Бекмухаметова*

ВЕРОЯТНОСТНЫЕ ОЦЕНКИ СХОДСТВА ПЕРВИЧНЫХ СТРУКТУР БЕЛКОВ

Институт физиологии и биофизики АН Республики Узбекистан, Ташкент

Ключевые слова: белки; структура первичная; критерии сходства белков.

Разработаны критерии, позволяющие оценивать, может ли сходство сегментов фиксированной длины, принадлежащих аминокислотным последовательностям различных белков, считаться признаком сходства данных белков. Составлена таблица минимальных значений числа идентичных аминокислот, для которого совпадение в сегменте заданной длины имеет вероятность менее 1%. Критерии подтверждаются при сравнении природных белков, их справедливость проверена на модельных аминокислотных последовательностях, созданных с помощью генератора случайных чисел.

В настоящее время имеется ряд работ [1—14], посвященных изучению аминокислотных (АК) последовательностей белков, в которых были предприняты попытки анализа их структуры и функции. В связи с тем что различные авторы приводят совпадающие фрагменты различной длины и с разным процентом совпадений, возникает вопрос о вероятностных критериях оценки значимости получаемых совпадений.

Приведем описание двух наиболее популярных методов, позволяющих оценить, можно ли наблюдаемые совпадения аминокислотных последовательностей двух белков считать чисто случайными или вероятность случайного совпадения весьма мала, а затем предложим собственную модификацию.

Суть первого из методов состоит в следующем [15]: вводится некоторая (основанная на тех или иных соображениях) матрица сходства аминокислот, выбирается определенная длина l участка аминокислотной последовательности и производится попарное сравнение всех участков длины l белка А со всеми участками длины l белка В. Для каждого сравнения подсчитывается сумма баллов сходства, получающаяся суммированием баллов сходства для АК-остатков, находящихся на соответствующих местах каждой из выбранных АК-последовательностей. Простейшим случаем матрицы сходства является диагональная матрица, у которой на диагонали находятся единицы (каждый АК-остаток сходен только сам с собой), а все остальные элементы нулевые. Очевидно, что в этом случае будет подсчитываться просто количество совпадающих АК-остатков на соответствующих позициях. Затем строится большое количество (до 200) пар модельных АК-последовательностей, набираемых случайным образом, но с сохранением аминокислотного состава природных белков. Для каждой пары проводится то же

сравнение. После этого для модельных «белков» определяется среднее число участков данной длины l, набравших при сравнении заданную сумму баллов сходства, а также дисперсия этой величины. Разность среднего и соответствующего значения для сравниваемых природных белков делится на величину дисперсии. Если полученный результат составляет три значения дисперсии и более, это считается признаком близости сравниваемых белков. Метод достаточно широко применяется, однако, очевидно, не лишен недостатков. Во-первых, при использовании данного метода для проведения одного сравнения природных белков необходимо тем же способом, т. е. с теми же затратами машинного времени, многократно сравнивать модельные последовательности. Соответственно, чем длиннее сравниваемые белки, тем больше непроизводительных затрат времени. Во-вторых, результаты сравнения могут сильно зависеть от свойств генератора случайных чисел, применяемого для набора модельных АК-последовательностей. Отметим также, что результаты каждого сравнения уникальны и с помощью данного метода невозможно получить некие более или менее универсальные критерии того, являются ли данные совпадения случайными, не проделав весь объем работ.

Другим методом сравнения является метод Маклаклана (см., например, [16]). Аналогично предыдущему описанному методу вводится матрица сходства, выбирается определенная длина l участка сравнения и сравниваются по тому же алгоритму попарно все участки длиной l белков А и В. Однако вероятность чисто случайно получить определенное число участков с некоторой заданной суммой баллов сходства в этом методе оценивается уже из соображений теории вероятностей по итерационному алгоритму с учетом аминокислотного состава сравниваемых белков. Вычисляется вероятность чисто случайного получения определенной суммы баллов сходства при одном сравнении пары участков длины l, затем подсчитывается число таких сравнений, производимых для двух белков, которое равно

$$N_{\text{общ}} = (N_{\text{А}} - l + 1)\,(N_{\text{В}} - l + 1),$$

где $N_{\text{А}}$ и $N_{\text{В}}$ — длины АК-последовательностей сравниваемых белков. Если произведение вероятности получения при однократном сравнении данного числа баллов на общее число сравнений не превышает 0,01, совпадение считается неслучайным. Этот метод признан наиболее приспособленным для локального сравнения белков. Однако и с его помощью получить какие-либо общие критерии для сравнения белков сложно, так как в выражениях для вероятности присутствуют процентные содержания аминокислот каждого из сравниваемых белков.

Вышеприведенные методы хорошо работают при сравнении белков с относительно короткой АК-последовательностью. Если же длина белков достаточно велика, то и сбор, и обработка результатов будут весьма затруднительны. В уже упоминавшихся работах [1—14] приводятся результаты сравнения АК-последовательностей транспортных белков, обладающих достаточно длинными АК-цепочками. Представляется желательным иметь более или менее простые критерии, позволяющие определить, какое минимальное число АК-остатков должно совпадать на участке заданной длины в двух белках любой длины, чтобы этот результат нельзя было отнести к разряду случайных. Целью настоящей работы является попытка получения таких критериев, которые могут использоваться при сравнении пары белков любой длины вне зависимости от их аминокислотного состава.

Для простоты будем считать, что нахождение любой аминокислоты (из 20 природных) в произвольном месте белка — событие равновероятное, вне зависимости от окружения. Наша цель — решить следующую задачу: если сравниваются АК-последовательности двух белков длиной $L1$ и $L2$, то какова вероятность обнаружения в каждом из этих белков $m1$ сегментов длиной l, в которых совпадают m аминокислот.

Решение данной задачи логично разбить на три части:

1) оценить вероятность того, что в двух АК-последовательностях длиной l, одна из которых считается заданной, а вторая произвольной, совпадут m аминокислот;

2) если мы имеем последовательности двух белков длиной $L1$ и $L2$, то какова вероятность обнаружить при их сравнении один сегмент длиной l, в котором совпадают m аминокислот;

3) какова вероятность обнаружить в этих двух АК-последовательностях длиной $L1$ и $L2$ более одного такого сегмента.

Вероятность совпадения двух аминокислот, одна из которых принадлежит к последовательности, считающейся заданной, а другая — к произвольной, равна 1/20. Можно назвать совпадение успехом, а несовпадение неудачей. Таким образом, вероятность успеха в нашем случае P (усп) = 1/20, а неудачи — P (неуд) = = 1 — P (усп) = 19/20. Сравнение двух аминокислот на соответствующих местах в двух различных аминокислотных последовательностях равной длины составляет единичный эксперимент. Тогда при сравнении двух последовательностей длиной l мы проводим всего l экспериментов, каждый с вероятностью успеха P (усп) и неудачи P (неуд). $P(m)$ — вероятность m успехов и $(l - m)$ неудач при сделанных предположениях определяется биномиальным распределением. Если ввести обозначения P (усп) = 1/20 = p; P (неуд) = q; $C(l, m) = l!(m!(l - m)!)$ — число сочетаний из l по m, то $P(m)$ запишется в стандартном виде (для последовательности длины l):

$$P(m) = C(l, m)\, p^m q^{l-m}. \qquad (1)$$

Таким образом, получено аналитическое выражение для вероятности совпадения m аминокислот в двух последовательностях длиной l, одна из которых считается заданной, а вторая произвольной.

Приступим к рассмотрению второй части задачи. Назовем экспериментом второго рода сравнение двух аминокислотных последовательностей длиной l. Назовем успехом второго рода совпадение в этих последовательностях m аминокислот, а неудачей второго рода — противоположное событие. Вероятность успеха второго рода в таком случае дается формулой (1). Введем обозначения: P (усп2) = $p2$; P (неуд2) = $q2$, где $p2 = P(m)$, $q2 = 1 - p2 = 1 - P(m)$.

Оценим теперь число экспериментов второго рода, которые мы можем произвести, располагая белками длиной $L1$ и $L2$. Очевидно, что из первого белка мы можем выбрать для сравнения сначала первые l аминокислот, затем участок длиной l начиная со второй аминокислоты, с третьей и т. д. до аминокислоты с номером, равным $L1 - l + 1$. Таким образом, всего $L1 - l + 1$ вариантов, причем каждый выбранный участок из первого белка может сравниваться со всеми $(L2 - -l + 1)$ возможными участками из второго белка. Поэтому общее число экспериментов второго рода, которые можно поставить, располагая белками длиной $L1$ и $L2$, равно

$$N_{\text{эксп}} = (L1 - l + 1)(L2 - l + 1).$$

Так как мы опять имеем ряд независимых (см. обсуждение данного вопроса ниже) экспериментов с известной вероятностью успеха и неудачи, то можно записать выражения для вероятности, дающие решение второго и третьего этапов задачи:

1) вероятность нахождения одного участка длиной l с m совпадающими аминокислотами в двух белках длиной $L1$ и $L2$

$$P'(1) = C(N_{\text{эксп}}, 1)\, p2^1 q2^{N_{\text{эксп}}-1} \qquad (2)$$

(здесь, как и ранее, $C(N_{\text{эксп}}, 1)$ — биномиальный коэффициент);

2) вероятность нахождения n участков длиной l с m совпадающими аминокислотами в двух белках длиной $L1$ и $L2$

$$P'(n) = C(N_{\text{эксп}}, n)\, p2^n q2^{N_{\text{эксп}}-n}. \qquad (3)$$

Очевидно, что данное рассмотрение легко обобщается на случай сравнения трех, четырех и более белков.

Итак, получены аналитические выражения для вероятности обнаружения заданного количества участков длиной l с m совпадающими АК-остатками при описанном методе сравнения и сделанных допущениях. С их помощью легко оценить пределы, в которых должно лежать число обнаруженных при сравнении случайным образом выбранных белков участков для заданных значений l, m, $L1$ и $L2$. Действительно, по формуле (3) можно рассчитать вероятности обнаружения нуля, одного, двух и т. д. участков длиной l с m совпадающими АК-остатками, и сумма этих вероятностей составляет единицу. Выбрав в качестве граничных значений величины n значения n_{min} и n_{max}, при которых сумма вероятностей

$$P'(n_{min}) + P'(n_{min} + 1) + P'(n_{min} + 2) + \ldots + P'(n_{max}) = 0{,}99,$$

можно получить разумную оценку для величины n, наблюдаемой при сравнении двух природных белков n_{real}.

Можно также подсчитать, каково должно быть минимальное значение m (при прочих известных параметрах), чтобы вероятность обнаружить хотя бы одну такую последовательность в данных условиях сравнения не превышала 1%, т. е. определить именно ту границу, начиная с которой совпадение не может считаться чисто случайным. (Очевидно, что это должно быть минимальное значение m, для которого из формулы (3) вероятность обнаружения нуля таких участков $P'(0) > 0{,}99$.) Преимущество данного подхода в том, что для определенного значения l соответствующее граничное значение m зависит только от числа проводимых сравнений (и зависимость достаточно плавная), и ни от чего больше, что позволяет один раз построить таблицу граничных значений и в дальнейшем пользоваться ею (табл. 1*а*).

С помощью формулы (3) можно оценивать также вероятность совпадений аминокислот в последовательности подряд (очевидно, что для этого достаточно положить l равным m). Обозначим через $P(m; j)$ вероятность обнаружить при сравнении двух белков заданное число j полностью совпадающих участков длиной в m АК. Расчет показывает, что совпадение даже 6 аминокислот подряд при сравнении двух белков, каждый из которых имеет длину порядка 1000 аминокислотных остатков,— событие весьма маловероятное ($P(6; 0) = 98{,}5\%$; $P(6; 1) = 1{,}5\%$). Соответственно совпадение большего числа аминокислот подряд не может быть отнесено в разряд случайных ($P(7; 0) = 99{,}92\%$; $P(7; 1) = 0{,}08\%$) и свидетельствует о существовании гомологии.

В рамках нашего рассмотрения концепция введения консервативных заместителей (см., например, [1]) означает то, что аминокислоты, объединенные в гомологичные группы, становятся неразличимыми. Таким образом, мы имеем не 20 независимых элементов, а всего 11 (эти 11 групп АК, считающихся различными, см. в работе [1]). Наличие на данном месте в двух белках не тождественных, но отнесенных к одной группе аминокислот трактуется как совпадение. Все приведенные выше формулы остаются верными, однако в формуле (1) в случае учета консервативных заместителей значение величин p и q должно быть изменено: $p = 1/11$; $q = 10/11$. Легко заметить, что введение понятия консервативных заместителей приводит к существенному увеличению вероятности нахождения совпадающих участков (табл. 1*б*) и это должно учитываться при проведении поиска совпадающих участков. В табл. 1 приведено минимальное число аминокислот, совпадение которых на участке длиной l не может быть случайным (вероятность такого совпадения составляет менее 1%).

Подчеркнем еще раз, что в табл. 1 приведен основной итог, ради которого и строилась вся предыдущая модель: мы получили то минимальное значение числа совпадающих АК-остатков на длине l, при котором их совпадение в сравниваемых белках определенной длины (т. е. фактически при проведении определенного числа экспериментов) не может быть объяснено причинами слу-

Таблица 1

Минимальное количество идентичных аминокислот (m), для которых случайное совпадение в окне сравнения длиной l (10—300 а. о.) имеет вероятность менее 1% (при заданном числе проведенных сравнений $N_{эксп}$)

А

$N_{эксп}$	10	20	30	50	75	100	150	200	250	300
5 000	7	9	10	13	16	18	23	28	32	36
50 000	7	9	11	14	17	20	25	30	34	38
250 000	8	10	12	15	18	21	26	31	35	40
1 000 000	8	11	12	16	19	22	28	32	36	41
2 000 000	8	11	13	16	19	22	28	32	37	41

Б

$N_{эксп}$	10	20	30	50	75	100	150	200	250	300
5 000	8	11	13	17	21	25	33	40	46	53
50 000	8	12	14	18	23	27	35	42	49	56
250 000	9	12	15	19	24	28	36	44	51	58
1 000 000	9	13	15	20	25	29	38	45	52	59
2 000 000	9	13	16	20	25	30	38	46	53	60

Примечание. Результаты таблицы А получены в предположении, что имеется 20 различных элементов для сравнения (20 природных АК); результаты таблицы Б получены для случая введения консервативных заместителей (см. [1]), когда имеется только 11 различных элементов для сравнения (11 групп АК; все АК, принадлежащие каждой данной группе, считаются неотличимыми друг от друга).

чайными и служит признаком близости этих белков. Назовем эти минимальные значения, приведенные в табл. 1, пороговыми значениями.

Необходимо отметить, что при выводе наших формул не делалось никаких предположений о природе сравниваемых белков и они, видимо, могут применяться для оценки результатов сравнения любых типов белков. Отметим, однако, что сравниваемые белки должны быть достаточно длинными (не менее 100 АК-остатков), так как все наши рассуждения базируются на вероятностной основе, а следовательно, на законе больших чисел (см. также обсуждение ограничений данного метода ниже).

С использованием описанного выше формализма и для проверки применимости полученных критериев был проведен компьютерный анализ сходства ряда природных транспортных белков (сравнение некоторых из них между собой другими методами см. в работе [1], сравнение ряда других — в сообщениях [2—14]). Результаты сравнения для семи пар природных транспортных белков приводятся в табл. 2—4 (список белков, использованных при сравнении, см. в табл. 5). В качестве длины l выбирались значения 30, 50, 100, 200, 300. Было получено удовлетворительное (до множителя 3—4 в худших случаях) согласие теоретических расчетных величин с результатами анализа совпадений реально существующих белков (табл. 2—4).

Для правильной интерпретации результатов сравнения необходимо иметь в виду следующее. Точного совпадения теоретических оценок интервалов для всех значений m ожидать не приходится, во-первых, из-за вероятностного характера предсказаний, во-вторых, из-за коррелированности результатов проводимых сравнений (обсуждение этого вопроса см. ниже). Поэтому точность предсказаний до

Таблица 2

Результаты попарного сравнения аминокислотных последовательностей природных и модельных белков

Длина сравниваемого участка (окна сравнения) составляет 30 а. о.
(Обозначения см. в примечании к табл. 2)

Наименование сравниваемых белков и номер в табл. 5	m	n_{real}	n_{min}	n_{max}	n_{mod}	N_{exp}	m пороговое
Арабинозо-протонный симпортер (1)	6	537	309	353	345	122 268	11
	7	139	46	73	87		
α-Цепь Na,K-АТР-азы (2)	8	24	3	16	21		(—)
	9	0	0	4	4		
Белок К-канала (3)	7	1738	514	561	483	1 099 674	12
	8	426	66	96	58		
	9	77	4	18	0		(+)
	10	21	0	4	0		
Белок Na-канала (4)	11	21	0	1	0		
	12	11	0	0	0		
	13	6	0	0	0		
	14	0	0	0	0		
Пермеаза лактозы (5)	7	697	250	290	236	553 536	12
	8	110	30	53	41		
	9	17	1	11	0		(—)
α-Цепь Na,K-АТР-азы (6)	10	3	0	3	0		
	11	0	0	1	0		
Na,Ca-Обменник (7)	8	493	116	151	156	1 808 602	13
	9	75	9	26	6		
Белок Na-канала (тип 3) (8)	10	25	0	5	0		(—)
	11	0	0	0	0		
α-Цепь Na,K-АТР-азы (9)	7	1319	434	480	493	935 354	12
	8	295	55	83	57		
	9	53	3	16	2		(+)
	10	8	0	4	0		
Na,Ca-Обменник (7)	11	4	0	1	0		
	12	5	0	0	0		
Транспортер галактозы (10)	7	594	183	258	225	449 080	12
	8	135	18	50	50		
Белок К-канала (11)	9	17	0	12	5		(—)
	10	0	0	4	4		
β-2-Цепь Na,K-АТР-азы (12)	6	979	641	763	749	258 912	12
	7	173	98	157	142		
α-Цепь Na,K-АТР-азы (6)	8	17	8	33	8		(—)
	9	0	0	8	0		

Примечание. В табл. 2—4 введены следующие обозначения: n_{real} — количество участков длиной l, в которых обнаружено m совпадающих АК-остатков в природных белках; n_{mod} — то же, полученное на модельных АК-последовательностях; n_{min} и n_{max} — границы, в которых с вероятностью 99% должно находиться теоретически число участков длиной l с заданным количеством совпадений m; N_{exp} — число попарных сравнений участков длиной l, проводимых при сравнении данной пары белков (см. текст); m пороговое — минимальное значение числа совпадающих АК-остатков на участке длиной l, которое уже нельзя считать случайным; (+) — при сравнении природных белков обнаружены совпадающие участки со значением m, равным пороговому или превышающим его; (—) — таких участков не обнаружено.

Таблица 3

Результаты попарного сравнения АК-последовательностей природных и модельных белков
Длина сравниваемого участка (окна сравнения) составляет 100 а. о.

Наименование сравниваемых белков и номер в табл. 5	m	n_{real}	n_{min}	n_{max}	n_{mod}	N_{exp}	m пороговое
Арабинозо-протонный симпортер (1)	13	195	54	101	115	76 838	20
	14	89	12	40	69		
β-Цепь Na,K-АТР-азы (2)	15	46	1	17	3		(—)
	16	0	0	8	0		
Белок К-канала (3)	15	696	68	119	113	936 224	22
	16	253	12	41	20		
	17	103	0	16	0		(+)
	18	30	0	7	0		
	19	33	0	3	0		
	20	19	0	2	0		
	21	24	0	1	0		
	22	55	0	1	0		
	23	30	0	0	0		
Белок Na-канала (4)	24	9	0	0	0		
	25	0	0	0	0		
Пермеаза лактозы (5)	15	247	27	64	41	449 936	21
	16	152	3	24	19		
	17	69	0	10	0		(—)
α-Цепь Na,K-АТР-азы (6)	18	24	0	5	0		
	19	20	0	2	0		
	20	0	0	1	0		
Na,Ca-Обменник (7)	16	310	27	64	67	1 613 092	22
	17	108	3	23	14		
Белок Na-канала (тип 3) (8)	18	8	0	9	4		(—)
	19	10	0	4	0		
	20	0	0	2	0		
α-Цепь Na,K-АТР-азы (9)	15	733	56	104	96	804 804	22
	16	189	10	37	28		
	17	54	0	14	0		(—)
	18	32	0	6	0		
Na, Ca-Обменник (7)	19	6	0	3	0		
	20	0	0	2	0		
Транспортер галактозы (10)	14	457	89	147	111	358 150	21
	15	154	20	53	14		
	16	104	2	20	0		(—)
Белок К-канала (11)	17	56	0	9	0		
	18	6	0	4	0		
	19	0	0	2	0		
β-2-Цепь Na,K-АТР-азы (12)	12	873	442	548	558	176 102	20
	13	298	143	211	139		
	14	100	38	79	28		(—)
α-Цепь Na,K-АТР-азы (6)	15	12	6	30	11		
	16	0	0	13	0		

небольшого множителя может считаться удовлетворительной. Однако все значения количества совпадающих АК-остатков m ниже порогового должны нас интересовать лишь как свидетельства общей правильности теории.

Наиболее важны и интересны результаты, получающиеся для значений m, равных пороговому и превышающих его, поскольку наличие даже одной пары совпадающих участков с числом совпадающих АК-остатков, равным или превышающим пороговое,— событие исключительно мало вероятное и практически однозначно свидетельствует о близости сравниваемых белков (в табл. 2—4 приведено число фактически проводимых при сравнении каждой пары белков экспериментов; пороговые значения m для каждого данного количества экспериментов были взяты из табл. 1). Очевидно, что наличие корреляций в проводимых сравнениях может лишь завысить число имеющихся совпадений для данного значения m, но если наличие даже одного такого совпадения — событие практически невероятное, то влияние корреляций в данном случае не должно нас интересовать. Отметим, что согласование теоретических значений и реальных результатов анализа улучшается при увеличении m, а именно число совпадающих участков с большими значениями m и является наиболее информативной величиной. Особенно убедительный результат — тот факт, что совпадения, отличные от случайных для заданной длины l (см. табл. 1), не встречаются у различных белков и проявляются у родственных (см. работы [1—14]).

Характерно, что совпадение весьма небольшого количества АК (в процентном отношении) на больших длинах (длина сравниваемого участка, или окно сравнения, 100, 200, 300) может служить признаком близости белков. Так, например, совпадение уже 41 аминокислоты на участке длиной 300 (событие, вероятность которого меньше 1%) имеет место лишь в случае сравнения белков, гомологичность которых можно считать практически определенной (см. первую пару сравниваемых белков в табл. 4 и работу [1]).

Совпадать могут либо один или несколько коротких фрагментов, либо гомология может быть «следовой», что выражается в совпадении некоторых АК на протяженном участке. В первом случае результатом анализа совпадений будет аномально высокое число совпадающих аминокислот при сравнении с использованием коротких окон сравнения, во втором случае — при использовании длинных окон. При этом, правда, остается открытым вопрос о влиянии вставок и делеций, однако получение совпадений с очень малой вероятностью при использовании достаточно длинных окон должно по крайней мере служить поводом для дополнительного анализа.

Например, пермеаза лактозы при сравнении с α-субъединицей Na,K-АТР-азы дает статистически невероятное число совпадений АК в окне 300, чего не наблюдается в малых окнах, т. е. гомология этих двух белков явно следовая и может быть следствием сходства мембранной организации этих переносчиков, но не конкретного механизма переноса. В то же время Na,Ca-обменник демонстрирует аномально высокое количество совпадений с α-субъединицей Na,K-АТР-азы при сравнении с использованием окна 30, а в больших окнах эти два транспортера совпадают мало. Этот факт может быть результатом наличия сравнительно короткого гомологичного участка, который может быть вовлечен в непосредственный механизм переноса иона натрия и/или его узнавание.

Натриевый канал электрического угря и калиевый канал дрозофилы дают статистически невероятные количества совпадений во всех окнах, что свидетельствует о единстве как структурной, так и функциональной организации этих белков, а также единстве механизма переноса одновалентного катиона [1].

С помощью генератора случайных чисел были смоделированы «аминокислотные последовательности» той же длины, что и природные белки, использованные для сравнения с целью обеспечения такого же количества возможных экспериментов. Генератор выдавал последовательность целых чисел, равномерно распределенных в интервале 1—20. Аминокислотам были присвоены номера из этого интервала, и они вносились в последовательность. В этих экспериментах было получено

Таблица 4

Результаты попарного сравнения АК-последовательностей, природных и модельных белков
Длина сравниваемого участка (окна сравнения) составляет 300 а. о.

Наименование сравниваемых белков и номер в табл. 5	m	n_{real}	n_{min}	n_{max}	n_{mod}	N_{exp}	m пороговое
Белок К-канала (3)	30	1613	58	107	62	523 224	40
	31	827	21	55	9		
	32	496	6	29	3		(+)
	33	310	1	16	0		
	34	266	0	7	0		
	35	257	0	4	0		
	36	168	0	2	0		
	37	72	0	2	0		
	38	19	0	1	0		
	39	17	0	1	0		
Белок Na-канала (4)	40	15	0	0	0		
	41	22	0	0	0		
	42	5	0	0	0		
	43	0	0	0	0		
Пермеаза лактозы (5)	27	1653	229	311	326	207 936	40
	28	1055	108	170	30		(+)
	29	651	47	91	0		
	30	257	17	49	0		
	31	160	5	27	0		
	32	119	0	15	0		
	33	57	0	9	0		
	34	28	0	6	0		
	35	22	0	4	0		
	36	42	0	2	0		
	37	21	0	2	0		
α-Цепь Na,K-АТР-азы (6)	38	10	0	1	0		
	39	10	0	1	0		
	40	13	0	1	0		
	41	11	0	0	0		
	42	34	0	0	0		
	43	14	0	0	0		
	44	0	0	0	0		
Na,Ca-Обменник (7)	32	802	19	53	11	1 108 492	41
	33	249	5	27	6		(—)
	34	227	0	15	15		
	35	66	0	8	0		
	36	30	0	5	0		
	37	39	0	2	0		
Белок Na-канала (тип 3) (8)	38	41	0	1	0		
	39	12	0	1	0		
	40	3	0	1	0		
	41	0	0	1	0		

Таблица 4 (окончание)

Наименование сравниваемых белков и номер в табл. 5	m	n_{real}	n_{min}	n_{max}	n_{mod}	N_{exp}	m пороговое
α-Цепь Na,K-АТР-азы (9)	31	757	19	52	113	485 804	40
	32	356	5	28	53		
	33	206	0	15	11		(—)
	34	118	0	9	0		
	35	30	0	5	0		
Na,Ca-Обменник (7)	36	9	0	3	0		
	37	0	0	2	0		
Транспортер галактозы (10)	28	966	75	126	76	152 350	41
	29	363	31	70	4		
	30	281	11	39	0		(—)
Белок К-канала (11)	31	179	2	22	0		
	32	46	0	12	0		
	33	0	0	8	0		

почти полное согласие теоретических расчетных величин с результатами анализа совпадений случайных последовательностей (в таблицах приведены средние величины по 10 модельным парам). (Более подробно об интерпретации результатов см. выше — обсуждение итогов сравнения природных белков.) Как и следовало ожидать, совпадения, которые превышали бы рассчитанные пороговые значения, не наблюдались ни разу. С целью проверки влияния аминокислотного состава конкретных сравниваемых белков на результаты сравнения были смоделированы также последовательности с сохранением этого состава. Характер результатов сравнения практически не изменился.

Необходимо отметить, что встречаемость аминокислот в белках неодинакова для всех аминокислот. Поэтому, строго говоря, вероятность встретить на данном месте последовательности заданную аминокислоту не равна 1/20. Для часто встречающихся аминокислот эта величина больше 1/20, для редко встречающихся — меньше. Например, если данная аминокислота очень редкая, вероятность встретить ее в белке в данной позиции близка к нулю, и, следовательно, вероятность встретить любую другую близка к 1/19. Однако так легко можно было бы учесть поправки только в простейшем случае. В реальной ситуации введение некоторых поправочных коэффициентов для вероятности нахождения каждой из аминокислот в данной позиции лишило бы рассмотрение общности и существенно затруднило бы получение аналитических результатов.

При выводе формул (1)—(3) неявно предполагалось, что $N_{эксп}$ раз сравниваются различные сегменты длиной l. Однако при нашем способе сравнения сравниваемые сегменты существенным образом коррелированы. Очевидно, что если в двух сегментах длиной l, расположенных на заданных местах в двух белках, мы обнаружили m совпадающих АК-остатков, то при смещении на одну позицию в том и другом белке и новом сравнении двух сегментов длиной l мы обнаружим $m-1$, m или $m+1$ совпадений. Таким образом, по результатам предыдущего опыта можно в некоторой мере предсказать результаты последующего.

Коррелированность сравниваемых участков последовательностей приводит и к наблюдаемым расхождениям между теоретически предсказываемым и реально получаемым числом участков с данным количеством совпадающих АК-остатков.

Тем не менее полученные результаты позволяют считать данные ограничения не слишком существенными (поскольку очень хорошо подтверждаются пороговые оценки, приведенные в табл. 1 (общий вид), 2—4 (конкретно для каждого сравнения), и наблюдается удовлетворительное (с точностью до множителя порядка

Таблица 5

Список белков, использованных для сравнения

Номер белка в табл. 2-4	Наименование сравниваемых белков	Наименование организма	Длина белка, а. о.	Номер ссылки
1	Арабинозо-протонный симпортер	*Escherichia coli*	472	6 7
2	β-Цепь Na,K-АТР-азы	*Gallus gallus*	305	13
3	Белок К-канала	*Drosophila melanogaster*	643	3
4	Белок Na-канала	*Electrophorus electricus*	1820	10
5	Пермеаза лактозы	*Kluyveromyces lactis*	587	2
6	α-Цепь Na,K-АТР-азы	*Gallus gallus*	1021	5
7	Na,Ca-Обменник	*Canis*	970	9
8	БелокNa-канала (тип 3)	*Rattus norvegicus*	1951	10
9	α-Цепь Na,K-АТР-азы	*Homo sapiens*	1023	5
10	Транспортер галактозы	*Saccharomyces cerevisiae*	574	12
11	Белок К-канала	*Rattus norvegicus*	853	3
12	β-2-Цепь Na,K-АТР-азы	*Homo sapiens*	290	8

рядка двух) согласование предсказываемых интервалов для количества совпадений и получаемых значений совпадений для значений m меньше порогового). Отметим также, что от влияния коррелированности сравниваемых участков несвободны также и два приведенных в начале статьи метода. Принципиально важно то, что коррелированность АК-последовательностей в нашем случае, как и в методе Маклаклана, не влияет на пороговые оценки, т. е. критерий неслучайности совпадений остается верным.

Таким образом, разработанные критерии позволяют оценивать, может ли обнаруженное сходство участков фиксированной длины, принадлежащих различным белковым последовательностям, считаться признаком родства данных белков. Эти критерии подтверждаются при сравнении природных белков, их справедливость проверена на модельных АК-последовательностях, созданных с помощью генератора случайных чисел. Принципиально важно то, что данные критерии в достаточной мере универсальны и однократно вычисленная таблица граничных значений может быть использована для оценки результатов сравнения любых белков.

СПИСОК ЛИТЕРАТУРЫ

1. *Catterall W. A.*//Science. 1988. V. 242. № T-41. P. 50—61.
2. *Chang Y.-D., Dickson R. C.*//J. Biol. Chem. 1988. V. 263. № 32. P. 16696—16703.
3. *Frech G. C., VanDongen A. M. J., Schuster G., Brown A. M., Joho R. H.*//Nature. 1989. V. 340. № 6235. P. 642—645.
4. *Karlin S., Bucher Ph.*//Proc. Nat. Acad. Sci. USA. 1992. V. 89. № 12. P. 12165—12169.
5. *Kawakami K., Ohta T., Nojima H., Nagano K.*//J. Biochem. 1986. V. 100. P. 389—397.
6. *Maiden M. C. J., Davis E. O., Baldwin S. A., Moore D. C. M., Henderson P. J. F.*//Nature. 1987. V. 325. № 6105. P. 641—643.
7. *Maiden M. C. J., Jones-Mortimer M. C., Henderson P. J. F.*//J. Biol. Chem. 1988. V. 263. № 17. P. 8003—8010.
8. *Martin-Vasallo P., Dackowski W. W., Emanuel J. R., Levenson R.*//J. Biol. Chem. 1989. V. 264. № 8. P. 4613—4618.
9. *Nicoll D. A., Longoni S., Philipson K. D.*//Science. 1990. V. 250. № T-43. P. 562—565.
10. *Noda M., Shimizu S., Tanabe T., Takai T., Kayano T., Ikeda T., Takahashi H., Nakayama H., Kanaoka Y., Minamino N., Kangawa K., Matsuo H., Raftery M. A., Hirose T., Inayama S., Hayashida H., Miyata T., Numa S.*//Nature. 1984. V. 312. № 5990. P. 121—127.

11. *Rackovsky S.*//Proc. Nat. Acad. Sci. USA. 1993. V. 90. № 3. P. 644—648.
12. *Szkutnicka K., Tshopp J. F., Andrews L., Cirillo V. P.*//J. Bacteriol. 1989. V. 171. P. 4486—4493.
13. *Takeyasu K., Tamkun M. M., Siegel N. R., Fambrough D. M.*//J. Biol. Chem. 1987. V. 262. № 22. P. 10733—10740.
14. *Takeyasu K., Tamkun M. M., Renaud K. J., Fambrough D. M.*//J. Biol. Chem. 1988. V. 263. № 9. P. 4347—4354.
15. *Гришин Е. В., Солдатова Л. Н., Солдатов Н. М., Костецкий П. В., Овчинников Ю. А.*//Биоорган. химия. 1979. Т. 5. № 9. С. 1285—1293.
16. *Поздняков В. И., Панков Ю. А.*//Молекуляр. биология. 1980. Т. 14. Вып. 1. С. 136—146.

Поступила в редакцию
25.III.1993
После доработки
25.III.1994

I. I. Ismailov, M. V. Sharov, Z. U. Beckmukhametova

PROBABILISTIC EVALUATION OF RESEMBLANCE OF PRIMARY STRUCTURES OF PROTEINS

Institute of Physiology and Biophysics,
Uzbek Academy of Sciences, Tashkent

Key words: proteins, primery structure.

Criteria have been developed for resemblance of segments of certain length of different proteins to be considered as an evidence of those proteins relationship. A table is formed of minimal numbers of identical amino acids for which the probability of the casual coincidence in a segment of certain length is less then 1%. The criteria were corroborated by comparison of natural proteins and model amino acid sequences concocted by means of the random numbers generator.