БИООРГАНИЧЕСКАЯ ХИМИЯ
том 16 * №12* 1990

УДК 577.112.088.5

П. В. Костецкий, Р. Р. Владимирова

МЕТОД ОБНАРУЖЕНИЯ КОНСЕРВАТИВНЫХ И ВАРИАБЕЛЬНЫХ УЧАСТКОВ В АМИНОКИСЛОТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ ГОМОЛОГИЧНЫХ БЕЛКОВ

Институт биоорганической химии им. М. М. Шемякина АН СССР, Москва

Исходное семейство гомологичных последовательностей разбивают на две группы из m и n выравненных аминокислотных последовательностей, которые внутри каждой группы принадлежат наиболее близким организмам. Результатом сравнения является профиль межгрупповой изменчивости (ПМИ), каждая точка которого соответствует изменчивости сегмента из 10 аминокислотных позиций (столбцов). Мерой изменчивости служит число замен во всех столбцах при $m \times n$ возможных парных сравнениях последовательностей, деленное на максимальное число замен ($m \times n \times 10$). Площадь между кривой ПМИ и прямой, проведенной на уровне среднего значения ПМИ, характеризует неравномерность аминокислотных замен вдоль гомологичных белковых последовательностей. Наблюдаемую площадь (S) сравнивают со средней расчетной площадью (S_p) для серии из 1000 искусственных гомологичных семейств белков, получаемых перестановкой столбцов аминокислотных остатков исходного семейства. Если разница ($S - S_p$) составляет не менее двух стандартных отклонений (σ) от величины S_p, выполняют идентификацию экстремальных пиков и впадин. Для этого проводят две горизонтальные прямые, отсекающие «излишек» площади, равный $S - (S_p + 2\sigma)$. Идентифицируемые таким образом на кривой ПМИ пики и впадины отвечают наиболее вариабельным и консервативным участкам гомологичных белковых последовательностей.

Для шести изученных семейств гомологичных белков (фосфолипазы А2, аспартатаминотрансферазы, α-субъединицы Na^+, K^+-АТР-аз, родопсины, L- и М-субъединицы фотореакционного центра фотобактерий) нашли, что ПМИ характеризуются высоким значением общей неравномерности распределения аминокислотных замен $(S - S_p)/\sigma$ и наличием достоверных консервативных и вариабельных участков в белковых последовательностях.

При сравнении аминокислотных последовательностей гомологичных белков часто обнаруживаются участки повышенного сходства (консервативные) и участки высокой изменчивости (вариабельные) [1—5]. Очевидно, что их наличие — следствие неравномерного расположения аминокислотных замен вдоль белковых последовательностей. Естественно поэтому, что консервативные и вариабельные участки резко различаются степенью локального сходства [1, 3] сравниваемых аминокислотных последовательностей.

Для количественной оценки локального сходства аминокислотных последовательностей гомологичных белков предложено несколько способов, основанных главным образом на подсчете числа различных аминокислотных остатков (а. к. о.) в каждой отдельной или нескольких расположенных подряд позициях [2—5]. При этом для облегчения поиска зон с заметным отличием числа аминокислотных замен от среднего числа замен по всей длине гомологичных белков с успехом применяют графическое сканирование выравненных последовательностей сегментом фиксированной длины [1, 3, 5, 6]. До настоящего времени, однако, существуют трудности с оценкой достоверности отличия консервативных и вариабельных участков от участков со случайным распределением аминокислотных замен вдоль гомологичных белковых последовательностей.

В 1982 г. был предложен метод оценки достоверности неравномерного распределения аминокислотных замен в четырех гомологичных фосфолипазах [4]. Вероятность неслучайности консервативных зон у этих фосфо-

```
          10        20        30        40        50        60  65
 1  NLYQFKNMIQCTVPNRSWWDFADYGCYCGRGGSGTPVDDLDRCCQVHDNCYDEAEKISRCWPYFK
 2  ..............S.....................................N......G......
 3  .........K....S...L...N.....................I.....N..G...G......
 4  .........H......P..H..N.........K...........I..K.........G....I.
 5  .........H.........H..N.....................I.....G......G....I.
 6  .........H....S.P..H...........K..A..............G....L-G....LT
 7  .........H....S.P..H...........K..................EK.G.M-G.....T

 8  ..I..G...SAMTGK-.SLAY.S......W..K.Q.K..T....F...C..GK.D.C.PKM---I
 9  D.T..G...NKMGQ--.VF.YIY......W..K.K.I.AT....F...C..GKMGTYDTK.---T
10  D.T..G...NKMGQ--.VF.YIY......W..Q.K.R.AT....F...C..GKMGTYDTK.---T

    70*00*000*******5**8*7*000000*0080*2*07*0000*500*0089797*8**4***9

          70        80        90       100       110       120       130
 1  TYSYECSQGTLTCKNGNNACAAAVCDCDRLAAICFAGAPYNNN-NYNIDLKARCQ-------
 2  ..........G...CA.....................D.-D...N......E------
 3  ..............GDD.N...S.........Y.........D-....N......-------
 4  ..T..SC.......D.G-K...S......V..N...R.T..DK-.....FN....-------
 5  ..T.DSC.....SCGAA.N...S......V..N...R...IDK-.....FN....-------
 6  L.K.....K...SG...K.E....N..LV..N.......IDA-...VN..E...-------
 7  L.K.K....K...SG..SK.G....N..LV..N.....R.IDA-....NF.K...-------

 8  L...KFHN.NIV.-GDK...KKK..E...V......ASKHSY.K.LWRYPSSK.TTGTAEKC
 9  S.N..IQN.GID.--DEDPQKKEL.E...V......NNRNTY.S..FGHSSSK.TGTEQC--
10  S.N..FQD.DII.-GDKDPQKKEL.E...V......NSRNTY.SK.FGYSSSK.TETEQCC-

    *0805***0***2*69*8*8***70*0035008000******8*54*******0*
```

Рис. 1. Сравнение аминокислотных последовательностей фосфолипаз рода *Naja* и рода *Bitis*: *N. n. kaouthia* (*1*), *N. n. atra* (*2*), *N. melanoleuca* I (*3*), *N. melanoleuca* III (*4*), *N. m. mossambica* (*5*), *N. m. pallida* (*6*), *N. n. oxiana* (*7*), *B. caudalis* (*8*), *B. gabonica* (*9*), *B. nasicornis* (*10*). Для достижения максимума сходства в гомологичные последовательности введены делеции, обозначенные дефисом. Одинаковые с последовательностью 1 остатки в последовательностях 2—10 обозначены точками. В нижней строке представлена доля аминокислотных замен (d) в каждой позиции при 21 возможном межгрупповом парном сравнении фосфолипаз родов *Naja* и *Bitis*. Использована символьно-цифровая шкала: «0» — $d = 0\%$; «1» — $0 < d <= 10\%$; «2» — $10 < d <= 20\%$...; «*» — $90 < d <= 100\%$

липаз определяли с помощью численного эксперимента на искусственных гомологичных семействах, которые получали произвольной перестановкой столбцов букв в выравненных последовательностях исходных фосфолипаз, расположенных друг под другом. Для этого подсчитывали частоту случайного возникновения консервативных участков, в которых число аминокислотных замен было бы не больше, чем в семействе природных фосфолипаз. Недостатком этого метода является необходимость оценки достоверности каждого консервативного участка в отдельности.

В настоящей работе значительно улучшен предложенный нами ранее [4, 5] метод обнаружения и статистической оценки консервативных и вариабельных участков в семействах гомологичных белков. Были изучены 6 семейств: родопсины [7, 8], аспартатаминотрансферазы [9], α-субъединицы Na^+, K^+-АТР-аз [10—13], L- и M-субъединицы фотореакционного центра фотосинтезирующих бактерий [6, 14—19] и фосфолипазы А2 [4, 9]. Последние были представлены двумя группами последовательностей: семью фосфолипазами из яда кобр *Naja* и тремя фосфолипазами из яда гадюк *Bitis* (рис. 1).

Очевидно, что возможно сравнение белковых последовательностей, принадлежащих или разным группам (межгрупповое сравнение), или одной и той же группе (внутригрупповое сравнение). Результат межгруппового сравнения гомологичных фосфолипаз А2 яда змей родов *Naja* и *Bitis* представлен на рис. 2 в виде профилей изменчивости. Каждая точка на графике соответствует изменчивости сегмента из 10 аминокислотных позиций (столбцов), в котором подсчитывалась сумма замен во всех столбцах при сравнении 7×3 пары а. к. о. белковых последовательностей. Чтобы величина изменчивости сегмента находилась в пределах 0—1, полученное значение делили на максимально возможное число замен 21×10. На рис. 2 присутствует также сплошная горизонтальная

прямая, проведенная на уровне среднего значения изменчивости по всей длине гомологичных белков. Легко видеть, что вариабельность или консервативность какого-либо участка сравниваемых гомологичных белков связана именно с величиной соответствующего пика или впадины относительно прямой среднего значения профиля изменчивости.

Профиль межгрупповой изменчивости фосфолипаз (рис. 2) характеризуется наличием отчетливых пиков и впадин, соответствующих наиболее вариабельным и консервативным участкам. При этом разброс вокруг среднего значения изменчивости весьма велик. Площадь, соответствующая этому разбросу и характеризующая неравномерность аминокислотных замен вдоль последовательностей гомологичных белков, заключена между кривой изменчивости и прямой, отвечающей среднему значению изменчивости. В рассматриваемом случае наблюдаемая площадь S значительно превосходит среднюю расчетную площадь S_p для серии из 1000 искусственных гомологичных семейств фосфолипаз, получаемых перестановкой столбцов а.к.о. исходного семейства [4]. Наблюдаемое различие $(S - S_p)$ — общая неравномерность аминокислотных замен — составляет 4,7 стандартных отклонений (σ) от среднего значения расчетной площади S_p и заметно превышает величину 2σ, которая отвечает 95%-ной достоверности отличия S от случайного значения S_p. Это превышение логично связать с экстремальными участками графика, которые можно выделить с помощью горизонтальных прямых, отсекающих «излишек» площади, равной $S - (S_p + 2\sigma)$. Идентифицируемые таким образом вариабельные и консервативные участки гомологичных последовательностей можно с вероятностью не менее чем 95% считать ответственными за наблюдаемую общую неравномерность распределения аминокислотных замен при сравнении последовательностей фосфолипаз А2 змей *Naja* и *Bitis* (рис. 1).

Важно, что получаемые предложенным способом искусственные гомологичные семейства идентичны исходным семействам по аминокислотному составу каждой последовательности и по числу замен при сравнении любой пары белковых последовательностей. В то же время искусственные семейства гомологичных семейств отличаются полезным свойством, согласно которому наличие в них консервативных и вариабельных участков возможно только в виде «шума». В соответствующих профилях изменчивости встречаются одиночные редкие пики и впадины, не оказывающие заметного влияния на величину S_p. Поэтому условие $S - S_p > 2\sigma$ представляется достаточно естественным для обнаружения достоверных пиков и впадин.

Более точная локализация координат вариабельных и консервативных участков возможна путем дополнительного анализа семейства гомологичных последовательностей (рис. 1). Один или несколько столбцов выравненных последовательностей на краях участка должны быть малоизменяемыми в случае консервативности или, напротив, быть очень изменчивыми в случае вариабельности участка. Такой анализ облегчает дополнительная строка (рис. 1), каждый символ которой соответствует процентной доле аминокислотных замен в отдельных столбцах при всех возможных межгрупповых парных сравнениях.

Важно отметить, что количественный состав сравниваемых групп имеет ограниченное влияние на результат сравнения. Так, профиль межгрупповой изменчивости обладает заметной устойчивостью, т. е. сохраняет число и местоположение достоверных пиков и впадин, даже при сравнении трех фосфолипаз *Bitis* с одной фосфолипазой *N. melanoleuca* III (рис. 2*б*) и при сравнении семи фосфолипаз *Naja* с единственной фосфолипазой *B. gabonica* (рис. 2*в*). Устойчивость межгруппового профиля вполне коррелирует со стабильностью числа аминокислотных замен (54—64%) при сравнении пар гомологичных последовательностей.

Такой устойчивости не имеет профиль внутригрупповой изменчивости семи последовательностей фосфолипаз змей *Naja* (рис. 3*а*), получаемый в результате также 21 ($7 \times 6/2$) парного сравнения а. к. о. в каждом столбце последовательностей (рис. 1), усредненных по 10 столбцам и

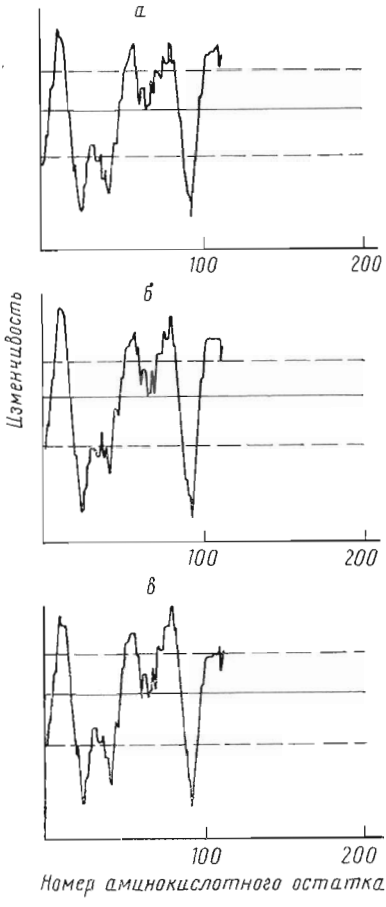

Рис. 2

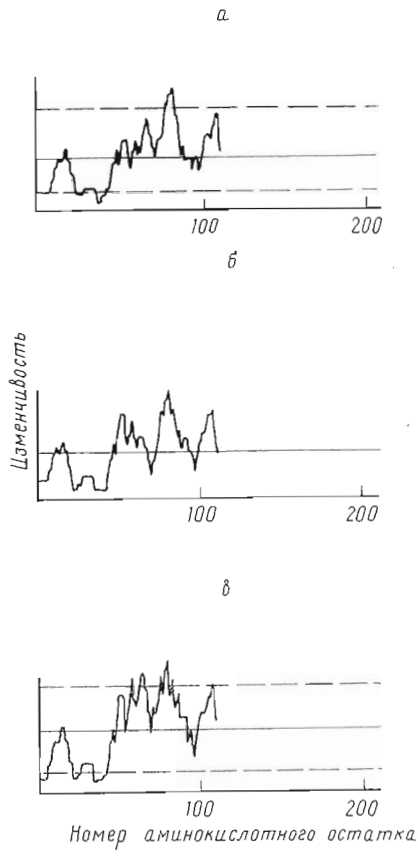

Рис. 3

Рис. 2. Профили межгрупповой изменчивости аминокислотных последовательностей фосфолипаз А2 яда змей рода *Naja* и рода *Bitis*. Сплошная горизонтальная линия отвечает среднему значению изменчивости по всей длине сравниваемых последовательностей. Верхняя и нижняя штриховые линии идентифицируют пики и впадины, отвечающие достоверным консервативным и вариабельным участкам гомологичных белковых последовательностей. *а* — семь фосфолипаз А2 *Naja* против трех фосфолипаз А2 *Bitis*; *б* — фосфолипаза А2 *N. melanoleuca* III против трех фосфолипаз А2 *Bitis*; *в* — семь фосфолипаз А2 *Naja* против фосфолипазы *B. gabonica*

Рис. 3. Профили изменчивости фосфолипаз А2 рода *Naja*. Обозначения те же, что и на рис. 2. *а* — внутригрупповой профиль изменчивости семи фосфолипаз (рис. 1); *б* — внутригрупповой профиль изменчивости пяти фосфолипаз А2 (без *N. melanoleuca* I и III); *в* — межгрупповой профиль изменчивости двух фосфолипаз *N. m. mossambica* и *N. m. pallida* против пяти фосфолипаз А2 остальных видов *Naja*

нормализованных к максимально возможному числу замен. Значение общей неравномерности оказывается заметно ниже и составляет 2,3σ. Одновременно уменьшается число пиков, выходящих за 95%-ный уровень случайного возникновения.

Профиль внутригрупповой изменчивости последовательностей фосфолипаз А2 видов *Naja* достаточно чувствителен к количественному и видовому составу группы. Например, исключение из этой группы фосфолипаз *N. melanoleuca* понижает значение общей неравномерности до 1,5σ и делает профиль изменчивости еще менее рельефным (рис. 3*б*). Меньшей устойчивости профиля внутригрупповой изменчивости соответствует значительный разброс числа аминокислотных замен при попарном сравнении фосфолипаз (8—29%). Более того, легко видеть, что средняя величина внутригрупповой изменчивости так же сильно зависит от состава группы.

При графической идентификации консервативных и вариабельных участков в семействах аминокислотных последовательностей гомологичных белков можно с успехом применять сегмент сканирования длиной

как меньше, так и больше 10 а.к.о. Следует, однако, иметь в виду, что при длине сегмента менее 5 или более 20 а.к.о. профиль изменчивости приближается к случайному, и это затрудняет идентификацию участков повышенной вариабельности и консервативности. Кроме того, при большой длине сегмента сканирования некоторые участки с резким различием числа аминокислотных замен могут остаться неидентифицированными. Напротив, при малой длине сегмента сканирования пики и впадины распадаются на большое число компонент, что не всегда удобно. При использовании сегмента сканирования длиной около 10 а.к.о. происходит определенное сглаживание профиля изменчивости и одновременно удается идентифицировать все наиболее консервативные и вариабельные участки аминокислотных последовательностей гомологичных белков.

Представленный выше метод «лишней площади» идентификации консервативных и вариабельных участков связан с оценкой площади, заключенной между кривой профиля изменчивости и прямой его среднего значения. При этом удается обнаружить сразу все пики и впадины, соответствующие вариабельным и консервативным участкам. Естественно, возможен и иной подход, при котором проверяется значимость каждого отдельно взятого пика или впадины.

Действительно, высокие пики и глубокие впадины редко встречаются на профилях изменчивости искусственных гомологичных семейств. Например, значение изменчивости, не уступающее амплитуде пика в районе 10-го а.к.о., в 1000 профилях изменчивости семейства искусственных фосфолипаз, получаемых перестановкой столбцов выравненных природных фосфолипаз (рис. 1), возникает менее 40 раз при длине сегмента сканирования 14 а.к.о. Тем самым пик в районе 10-го а.к.о. (рис. 2*а*) выходит за уровень 95%-ного неслучайного возникновения, и его можно считать уникальным по амплитуде. Напротив, остальные пики и впадины на рис. 2*а* не являются уникальными по амплитуде. Это не кажется удивительным, так как уникальность амплитуд пиков и впадин профиля изменчивости очень зависит от длины сегмента сканирования, и при сканировании сегментом в 10 а.к.о. уже ни один из экстремумов не является уникальным по амплитуде.

Очевидно также, что уникальность амплитуд пиков и впадин профилей изменчивости действительна только на ограниченной длине белковой последовательности. В связи с этим заметно преимущество предлагаемого метода «лишней площади», для идентификации консервативных и вариабельных участков в гомологичных белках. Этот метод применим к белкам любой длины (см. ниже), и результат идентификации не зависит в широких пределах от длины сегмента сканирования.

Нами были исследованы также профили изменчивости аминокислотных последовательностей еще пяти семейств гомологичных белков: родопсинов [7, 8], аспартатаминотрансфераз [9], α-субъединиц Na^+, K^+-АТР-аз [10—13], L- и M-субъединиц фотореакционного центра фотосинтезирующих бактерий [6, 14—19]. Данные о видовом составе семейств, а также о местоположении делеций в выравненных аминокислотных последовательностях приведены в табл. 1. Оказалось, что наличие рельефных профилей межгрупповой изменчивости с высоким значением общей неравномерности аминокислотных замен характерно и для всех остальных исследованных семейств гомологичных белков (табл. 2). Например, общая неравномерность профиля межгрупповой изменчивости аспартатаминотрансферазы из *E. coli* (первая группа последовательностей) [9] и митохондриальных аспартатаминотрансфераз трех млекопитающих (вторая группа последовательностей) [9] составляет 2,4σ; в профиле присутствуют шесть пиков и четыре впадины большой интенсивности (рис. 4*а*), отвечающие вариабельным и консервативным участкам белковых последовательностей. В то же время профиль внутригруппового изменчивости митохондриальных аспартатаминотрансфераз млекопитающих (рис. 4*б*) имеет невысокое значение общей неравномерности распределения аминокислотных замен — 0,9σ. Это делает затруднительной идентификацию достоверных участков

Таблица 1

Видовой состав семейств и групп гомологичных белковых последовательностей и местоположение делеций в них

Семейство гомологичных белков	Длина выравненных последовательностей	Первая группа организмов	Местоположение делеций	Вторая группа организмов	Местоположение делеций
Фосфолипазы А2 змей	127	*N. n. kaouthia*	109, 121–127	*B. gabonica*	15–16, 62–64, 79–80, 126–127
		N. n. atra	109, 122–127	*B. nasicornis*	15–16, 62–64, 79, 127
		N. melanoleuca I	83, 109, 121–127	*B. caudalis*	17, 62–64, 79
		N. melanoleuca III	109, 121–127		
		N. m. mossambica	58, 109, 121–127		
		N. m. pallida	58, 109, 121–127		
		N. n. oxiana	109, 121–127		
Аспартатаминотрансферазы	401	Человек	–	*E. coli*	1–2, 119, 224
		Свинья	–		
		Крыса	–		
α-Субъединицы Na^+, K^+-АТР-азы	1023	Человек	7, 13–19, 27, 379	Скат	499
		Овца	22, 27		
		Свинья	22, 27		
L-Субъединицы фотореакционного центра фотосинтезирующих бактерий	320	*C. aurantiacus*	244–246, 297, 315–320	*Rps. viridis*	1–15, 30, 32–40, 44–53, 87, 97–98, 100, 116, 313–320
				Rh. capsulatus	1–15, 30, 32–40, 44–53, 92, 98–100, 116
				Rs. rubrum	1–15, 30, 32–40, 44–53, 90, 96–98, 116, 312, 314, 317–320
				Rb. sphaeroides	1–15, 30, 32–40, 44–53, 92, 98–100, 116
M-Субъединицы фотореакционного центра фотосинтезирующих бактерий	328	*C. aurantiacus*	1–8, 17–18, 107, 319–328	*Rps. viridis*	16, 44, 107, 291, 307
				Rb. capsulatus	37, 44, 107, 291, 307, 312–328
				Rs. rubrum	37, 44, 107, 291, 307, 312–328
				Rb. sphaeroides	44, 291, 307, 311–328
Родопсины * человека	367	Родопсин	1–16, 349, 352, 355	Красный	352, 362–363
				Зеленый	352, 362–363
				Синий	1–19

* Все четыре гомологичных родопсина принадлежат человеку. Разбивка на две группы произведена по функциональному признаку.

Таблица 2

Результаты межгруппового сравнения семейств гомологичных белковых последовательностей и местоположение участков высокой консервативности и высокой вариабельности

Семейство гомологичных белков	% аминокислотных замен	Общая неравномерность (ед. стандарт. отклонения)	Участки высокой консервативности	Участки высокой вариабельности
Фосфолипазы А2 змей *Naja* и *Bitis*	54–64	4,73	24–34, 41–51, 90–101	10–23, 52–66, 75–89, 102–118
Аспартатаминотрансферазы (*E. coli* и митохондрий млекопитающих)	59–60	2,36	99–116, 177–197, 213–223, 247–261	3–12, 56–66, 117–130, 163–176, 198–212, 230–246, 338–349
α-Субъединицы Na^+, K^+-АТР-азы (рыба и млекопитающие)	13–22	14,52	70–118, 181–208, 211–229, 235–256, 265–408, 447–461, 479–493, 504–519, 538–557, 583–652, 685–791, 799–867, 899–912, 929–970	5–30, 119–180, 230–234, 257–264, 416–426, 434–442, 462–478, 494–503, 520–537, 558–582, 673–684, 881–898, 971–976, 1004–1012
L-Субъединицы фотореакционного центра фотосинтезирующих бактерий (зеленая и пурпурные)	58–61	4,25	56–69, 192–205, 226–235	80–95; 238–250, 275–290, 292–308
M-Субъединицы фотореакционного центра фотосинтезирующих бактерий (зеленая и пурпурные)	56–63	3,81	47–54, 102–116, 195–226, 243–261	9–28, 55–68, 71–82, 143–153, 227–242, 275–292, 300–310
Родопсины человека (черно-белый и цветные)	57–60	6,87	82–94, 117–130, 142–158, 194–204, 259–270, 307–319	20–38, 52–70, 100–116, 131–141, 165–185, 207–230, 236–249, 289–306, 333–349

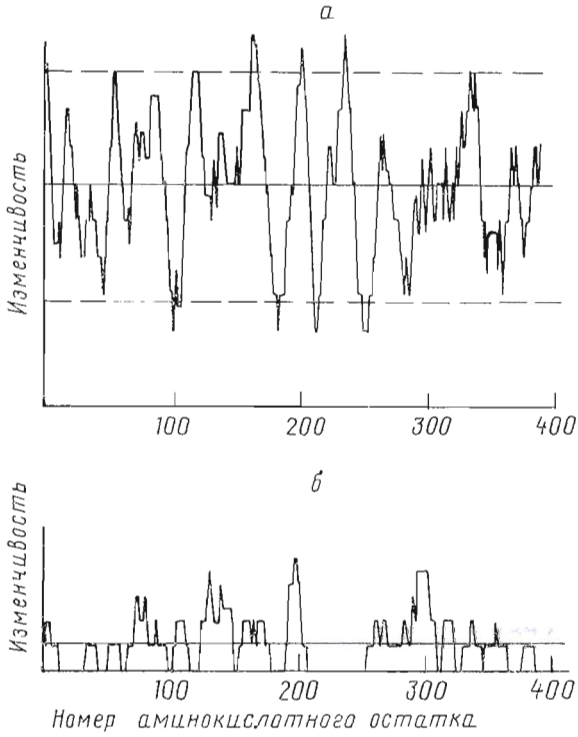

Рис. 4. Профили изменчивости аспартатаминотрансфераз. Обозначения те же, что и на рис. 2. *а* — межгрупповой профиль изменчивости аспартатаминотрансферазы из *E. coli* против трех аминотрансфераз млекопитающих; *б* — внутригрупповой профиль изменчивости трех аспартатаминотрансфераз млекопитающих

высокой консервативности или вариабельности в аминокислотных последовательностях аспартатаминотрансфераз группы млекопитающих.

Большинство результатов, представленных в табл. 2, по межгрупповому сравнению аминокислотных последовательностей выполнено для гомологичных белков, принадлежащих различным организмам. Единственное внутривидовое семейство последовательностей образуют четыре родопсина человека, обеспечивающие черно-белое и цветное зрение [7, 8]. В этом случае логично сравнивать черно-белый родопсин с группой из трех цветных родопсинов. Действительно, получаемый при этом профиль сравнения характеризуется высоким значением общей неравномерности замен (6,2σ) и содержит ряд интенсивных пиков и впадин, отвечающих вариабельным и консервативным участкам белковых последовательностей (табл. 2).

Следует отметить, что абсолютной величины локальной изменчивости участка гомологичных аминокислотных последовательностей недостаточно для отнесения участка к вариабельным или консервативным. Действительно, в случае профиля межгруппового сравнения фосфолипаз *Naja* и *Bitis* (рис. 2*а*) значение изменчивости менее 0,40 говорит о высокой консервативности соответствующего участка. В то же время при внутригрупповом сравнении фосфолипаз змей *Naja* (рис. 3*а*) значение изменчивости более 0,43 отвечает высокой вариабельности. Это неудивительно, так как степень консервативности или вариабельности тех или иных участков последовательностей гомологичных белков связана только с величиной отклонения значения изменчивости данного участка от среднего значения изменчивости по всей длине сравниваемых последовательностей.

При построении профилей межгрупповой изменчивости важно правильно выделять группы наиболее близких последовательностей в изучаемом семействе гомологичных белков. Самый простой способ, по-видимому,— предварительное построение дихотомичного филогенетического дерева, вершина которого делит семейство нисходящих ветвей на две группы [20]. Например, семейство последовательностей фосфолипаз А2 змей рода

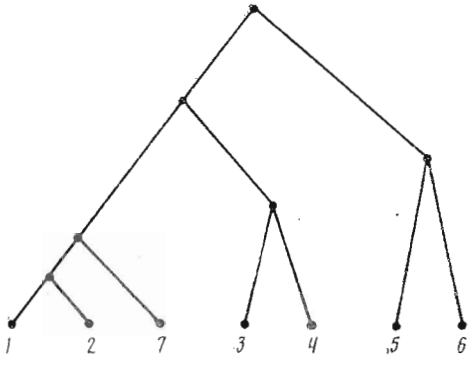

Рис. 5. Филогенетическое дерево змей рода *Naja*, построенное исходя из аминокислотных последовательностей фосфолипаз А2. Номера в нижней части рисунка соответствуют видам змей рода *Naja* и совпадают с обозначениями рис. 1

Naja делится таким образом (рис. 5) на группу из двух фосфолипаз видов *N. m. mossambica* и *N. m. pallida* и группу из пяти фосфолипаз остальных видов *Naja*. В результате разбиения семейства *Naja* на две группы оказалось возможным получить рельефный профиль межгрупповой изменчивости (рис. 3*в*). Этот профиль отличается более высоким значением общей неравномерности (2,7σ против 2,3σ) и большим числом достоверных консервативных и вариабельных участков по сравнению с внутригрупповым профилем изменчивости семи фосфолипаз змей *Naja* (рис. 3*а*).

При филогенетическом способе формирования сравниваемых групп аминокислотных последовательностей идентифицируемые консервативные и вариабельные участки характеризуют особенности накопления аминокислотных замен на определенном этапе эволюции. Что касается профиля внутригрупповой изменчивости, то его можно рассматривать как суперпозицию ряда межгрупповых профилей. Например, внутриродовой профиль изменчивости семи фосфолипаз А2 *Naja* (рис. 3*а*) является суммой шести межгрупповых профилей, что соответствует числу вершин филогенетического дерева (рис. 5).

Для семейства Na^+, K^+-АТР-аз и родопсинов известны также и нуклеотидные последовательности соответствующих структурных генов. В связи с этим представлялось интересным изучить профили изменчивости семейств указанных генов. Оказалось, что профиль межгрупповой изменчивости генов α-субъединиц Na^+, K^+-АТР-аз ската [10] и группы из трех млекопитающих [11—13] сильно отличается от шума (длина сегмента сканирования равнялась 30 нуклеотидам), о чем говорит высокое значение общей неравномерности нуклеотидных замен (12,6σ). Отличным от шума ($S - S_р = 10{,}2\sigma$) оказался и профиль межгруппового сравнения гена черно-белого родопсина человека [7] с группой генов трех цветных родопсинов человека [8]. Эти результаты достаточно интересны, так как аналогичные профили межгруппового сравнения белковых семейств также резко отличаются от шума и имеют высокое значение общей неравномерности аминокислотных замен: 14,5σ в случае Na^+, K^+-АТР-аз и 6,9σ в случае родопсинов (табл. 1). Кроме того, в профилях изменчивости белков и кодирующих их мРНК наблюдается взаимное соответствие ряда значимых пиков и впадин, что может служить дополнительным свидетельством неслучайности их возникновения.

Таким образом, в настоящей работе предложен метод идентификации вариабельных и консервативных участков в семействах последовательностей гомологичных белков. При этом в соответствии с филогенетическим деревом формируют две группы выравненных аминокислотных последовательностей, каждая из которых состоит из наиболее близких организмов. По описанному выше способу получают профиль межгрупповой изменчивости. Если общая неравномерность распределения аминокислотных замен, оцениваемая с помощью численного эксперимента на искусственных семействах гомологичных белков, превышает случайный уровень, выполняют идентификацию экстремальных пиков и впадин. Для этого проводят горизонтальные прямые, отсекающие в экстремальных участках профиля изменчивости «излишек» площади, равный $S - (S_р + 2\sigma)$, где S — наблюдаемая площадь между кривой изменчивости и прямой ее среднего значения, $S_р$ — средняя расчетная площадь для большого числа искусственных семейств гомологичных белковых последовательностей. Иден-

тифицируемые таким образом на кривых профилей изменчивости пики и впадины отвечают наиболее вариабельным и консервативным участкам аминокислотных последовательностей гомологичных белков.

Естественно предположить, что обнаруживаемые предлагаемым методом наиболее вариабельные и консервативные участки аминокислотных последовательностей гомологичных белков могут отвечать за проявление структурных, функциональных, видовых и иных особенностей белков каждого семейства. В таком случае знание местоположения этих участков в сочетании с информацией о трехмерной структуре окажется полезным при конструировании генно-инженерными методами белков и пептидов с заданными свойствами.

Все вычисления и построение графиков выполняли с помощью программ, написанных на языке ФОРТРАН применительно к ЭВМ «Hewlett-Packard-3000» (США). Данные об аминокислотных последовательностях гомологичных белков брали из оригинальных работ и 19-го выпуска банка белковых последовательностей PIR [9]. При построении филогенетического дерева последовательностей фосфолипаз А2 змей использовали матрицу попарных различий аминокислотных последовательностей. Элементами этой матрицы являлись числа аминокислотных замен, приходящиеся на 100 позиций, с учетом скрытых вырожденных замен [21]. Профили изменчивости строили только для общей части выравненных последовательностей гомологичных белков. Общая часть последовательностей формировалась таким образом, чтобы ни одна из последовательностей не начиналась и не оканчивалась делецией.

Площадь фигуры, характеризующую общую неравномерность замен в семействе белков, оценивали по формуле

$$\sum_{i=1}^{N} |I_i - I|,$$

где N — число точек профиля изменчивости, I_i — значение изменчивости в i-й точке, $I = \sum_{i=1}^{N} \frac{1}{N} I_i$ — среднее значение изменчивости. Искусственные семейства белковых последовательностей получали произвольной перестановкой столбцов однобуквенных символов а.к.о. исходного семейства выравненных белков, используя датчик случайных чисел. Ординаты горизонтальных линий, идентифицирующих вариабельные и консервативные участки белковых последовательностей, определяли с точностью 0,05, так, чтобы площадь между кривой изменчивости и этими линиями составляла $S - (S_p + 2\sigma)$. Для последовательного приближения площади к заданной величине применяли метод деления отрезка пополам.

Авторы выражают признательность Н. Б. Флоровой (биофак МГУ) за обсуждение отдельных результатов настоящей работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Панков Ю. А., Поздняков В. И., Туманян В. Г.* // Молекуляр. биология. 1976. Т. 10. № 2. С. 423—436.
2. *Maloy W. L., Coligan J. E.* // Immunogenetics. 1982. V. 16. № 1. P. 11—22.
3. *Wieland B., Tomasselli A. G., Nodal L. H., Frank R., Schulz G. E.* // Eur. J. Biochem. 1984. V. 143. № 2. P. 331—339.
4. *Мещерякова Е. А., Айанян А. Е., Костецкий П. В., Мирошников А. И.* // Биоорган. химия. 1982. Т. 8. № 3. С. 349—363.
5. *Kostetsky P. V., Vladimirova R. R., Archipova S. F., Abdulaev N. G.* // WATOC World Congress Abstracts. Budapest, 1987. P. 363.
6. *Belanger G., Berard J., Corriveau P., Gingras G.* // J. Biol. Chem. 1988. V. 263. № 16. P. 7632—7638.
7. *Nathans J., Hogness D. S.* // Proc. Nat. Acad. Sci. USA. 1984. V. 81. № 15. P. 4851—4855.
8. *Nathans J., Hogness D. S.* // Science. 1986. V. 232. № 4747. P. 203—210.
9. *Sidman K. E., George D. C., Barker W. C., Hunt L. T.* // Nucl. Acids. Res. 1988. V. 16. № 5. P. 1869—1871.
10. *Kawakami K., Noguchi S., Noda M., Takahashi H., Ohta T., Kawamura M.,*

Nojima H., Nagano K., Hirose T., Inayama S., Nayachida H., Miyata T., Numa S. // Nature. 1985. V. 316. № 6030. P. 733—736.
11. *Shull G. E., Schwartz A., Lingrel J. B.* // Nature. 1985. V. 316. № 6030. P. 891—895.
12. *Ovchinnikov Yu. A., Modyanov N. N., Broude N. E., Petrukhin K. E., Grishin A. V., Arzamasova N. M., Aldanova N. A., Monastyrskaya G. S., Sverdlov E. D.* // FEBS Lett. 1986. V. 201. № 2. P. 237—245.
13. *Kawakami K., Ohta T., Nojima H., Nagano K.* // J. Biochem. (Tokyo). 1986. V. 100. № 2. P. 389—398.
14. *Ovchinnikov Yu. A., Abdulaev N. G., Zolotarev A. S., Shmukler B. E., Zargarov A. A., Kutuzov M. A., Telezhinskaya I. N., Levina N. B.* // FEBS Lett. 1988. V. 231. № 1. P. 237—242.
15. *Ovchinnikov Yu. A., Abdulaev N. G., Shmukler B. E., Zargarov A. A., Kutuzov M. A., Telezhinskaya I. N., Levina N. B., Zolotarev A. S.* // FEBS Lett. 1988. V. 232. № 2. P. 364—368.
16. *Youvan D. C., Bylina E. J., Alberti M., Begush H., Hearst J. E.* // Cell. 1984. V. 37. № 4. P. 949—957.
17. *Williams J. C., Steiner L. A., Odgen R. C., Simon M. I., Feher G.* // Proc. Nat. Acad. Sci. USA. 1983. V. 80. № 21. P. 6505—6509.
18. *Williams J. C., Steiner L. A., Feher G., Simon M. I.* // Proc. Nat. Acad. Sci. USA. 1984. V. 81. № 23. P. 7303—7307.
19. *Michel H., Weyer K. A., Gruenberg H., Dunger I., Oesterhelt D., Lottspeich F.* // EMBO J. 1986. V. 5. № 6. P. 1149—1158.
20. *Fitch W. M., Margoliach E.* // Science. 1967. V. 155. № 3760. P. 279—284.
21. *Dickerson R. E.* // J. Molec. Evolution. 1971. V. 1. № 1. P. 26—45.

Поступила в редакцию
25.IX.1989

После доработки
13.IV.1990

P. V. KOSTETSKY, R. R. VLADIMIROVA

A METHOD OF LOCALIZATION OF CONSTANT AND VARIABLE REGIONS IN HOMOLOGOUS PROTEIN SEQUENCES

M. M. Shemyakin Institute of Bioorganic Chemistry,
Academy of Sciences of the USSR, Moscow

A set of aligned homologous protein sequences is divided into two groups consisting of m and n sequences. Each group contains sequences from the most related organisms. Value of the position dissimilarity of proteins from different groups of m and n sequences is defined as a number of mismatches in comparison of all possible $m \times n$ pairs of amino acid residues in the position (each from different group) divided by $m \times n$. Ten position average of dissimilarity values is plotted vs. the first position number. Area of the figure between the profile of dissimilarity values and its mean value line characterizes the overall irregularity of amino acid substitutions along the protein sequences. If the area is greater than the average area for 1000 random profiles by more than two standard deviation units, the profile extrema containing the «surplus» of area are cut off. The cut-off stretches are likely to be variable and constant regions. If necessary, each of stretches may be separately tested and statistically estimated using a standard size sample of artificial protein families.

Intergroup comparison of protein sequences reveals high overall irregularity of amino acid substitutions and identifies variable and conservative regions for all considered families of proteins: phospholipases A2, aspartate aminotransferases, alpha-subunits of Na^+, K^+-ATPase, L- and M-subunits of photosynthetic bacteria photoreaction centre, human rhodopsins.