БИООРГАНИЧЕСКАЯ ХИМИЯ
том 3 * № 3 * 1977

УДК 547.963.32.02

ОПТИМИЗАЦИЯ БЛОЧНОГО МЕТОДА ОПРЕДЕЛЕНИЯ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Рашин А. А., Юдман Б. Х., Киселев Л. Л., Мазо А. М.

Институт биологической физики Академии наук СССР, Пущино;
Институт молекулярной биологии Академии наук СССР, Москва

Рассмотрены возможности оптимизации блочного метода определения нуклеотидных последовательностей РНК.

Предложены строгие критерии единственности последовательности, построенной по данным полных и частичных гидролизов. Показано, что однозначное определение нуклеотидной последовательности по данным только полных гидролизов, специфичных к одному нуклеотиду каждый, возможно для 90±5% последовательностей длиной 15 нуклеотидов при двух, 30 нуклеотидов при трех и 70 нуклеотидов при четырех полных гидролизах.

Для случая трех или четырех полных гидролизов предложены детальные алгоритмы обработки биохимической информации, позволяющие резко сократить объем экспериментальной работы, необходимой для однозначного восстановления последовательности.

Сформулирована последовательность биохимических операций при секвенировании оптимизированным блочным методом, исходя из расщеплений по G (рибонуклеаза T_1), C и U (панкреатическая рибонуклеаза, действующая на рибонуклеотид, модифицированный карбодиимидом и метоксиаминбисульфитом соответственно).

Введение

За последние 10 лет установление нуклеотидных последовательностей в рибонуклеиновых кислотах прошло путь от первой работы Холли и сотр. [1], выполненной на $тРНК^{Ala}$ (дрожжи), имевшей длину 78 нуклеотидов, до расшифровки первичных структур высокомолекулярных РНК — вирусных [2—4] и рибосомальных [5—6], состоящих из сотен и даже тысяч нуклеотидов.

В подавляющем большинстве работ (см. [7]) по секвенированию применялся блочный метод, заключающийся в расщеплении молекулы полирибонуклеотида на блоки действием гуанил- и пиримидил-РНКаз (полный гидролиз), расшифровке последовательностей полученных олигонуклеотидов, получении неполных гидролизатов разными способами и затем восстановлении полной первичной структуры на основе перекрывания блоков, полученных в разных гидролизах.

Возможности этого общего подхода доказаны тем, что расшифровано общее количество нуклеотидных последовательностей, превышающее, по ориентировочным данным, 15 000 нуклеотидов [8].

Однако потребности молекулярной биологии и молекулярной генетики намного больше того, что к настоящему времени известно о первичных

структурах РНК. Расшифровка нуклеотидных последовательностей пока, пожалуй, больше искусство, чем использование рутинных методов, до настоящего времени она является делом немногих виртуозов, сконцентрированных в нескольких крупных лабораториях, и требует огромной затраты труда и времени, особенно для высокомолекулярных РНК.

Чтобы мог произойти скачок в количестве расшифровываемых последовательностей РНК, необходима либо разработка новых методов [9—10], либо существенное усовершенствование зарекомендовавшего себя блочного метода и переведение его в рамки рутинных процедур.

Учитывая огромный опыт и технику секвенирования, разработанную при использовании блочного метода, мы попытались пойти по второму пути, причем оптимизация блочного метода касалась исследования, с одной стороны, способов наиболее экономного получения необходимой структурной информации с применением трех исчерпывающих гидролизов (на основе новых приемов ферментативного расщепления химически модифицированных субстратов [11—14]) и, с другой — алгоритмов, позволяющих планировать эксперимент и резко уменьшить объем требующейся для определения структуры биохимической информации.

Постановка задачи и структура статьи

Биохимические процедуры, используемые при определении нуклеотидной последовательности блочным методом, можно расположить по степени возрастания их трудоемкости примерно следующим образом:

1—2) определение нуклеотидного состава и концевой анализ;

3) определение последовательности во фрагменте из 5—8 нуклеотидов;

4) получение и разделение продуктов полного гидролиза;

5) определение последовательности во фрагменте из 10—20 нуклеотидов;

6) получение и анализ продуктов частичного гидролиза.

Очевидно, что степень развития тех или иных методик в конкретной лаборатории может привести к несколько иному распределению процедур по степени их трудоемкости. Мы при оценке сравнительной сложности биохимических процедур руководствовались: а) наличием большого числа способов концевого анализа [7]; б) простотой и быстротой определения нуклеотидного состава методами хроматографии под давлением [15]; в) тем, что при полном гидролизе длинных фрагментов число разделяемых продуктов может составлять несколько десятков, а при анализе последовательности октануклеотида путем неполного гидролиза фосфодиэстеразой число разделяемых продуктов не превышает 7; г) тем, что определение последовательности во фрагментах из 10—20 нуклеотидов при использовании двух-трех специфических методов расщепления (см. ниже) оказывается существенно более простой задачей, чем подбор условий проведения неполных гидролизов и разделение и анализ их продуктов.

Кроме того, имеется трудоемкая не биохимическая процедура:

7) построение последовательности перекрывающихся продуктов полного гидролиза и проверка единственности такого построения при данной биохимической информации.

Если ясные критерии единственности построенной последовательности отсутствуют, то для уверенности в полученном результате довольно часто приходится лишний раз повторять наиболее сложную шестую процедуру.

Мы ставим перед собой следующие задачи:

А. Найти способы определения нуклеотидных последовательностей продуктов гидролизов при помощи наиболее простых из перечисленных биохимических процедур.

Б. Найти эффективные способы построения возможных последовательностей по имеющейся информации.

В. Найти критерии, позволяющие определять, достаточно ли имеющейся информации для однозначного восстановления первичной структуры.

Г. Уметь указывать, какую информацию необходимо получить дополнительно для достижения однозначности.

Д. Проанализировать, что может дать для оптимизации применение новых методов полного гидролиза РНК [11—14].

Математический анализ части поставленных задач проведен в работах [16—20]. На основе этого анализа построены алгоритмы, подробно описанные в статье, с ориентацией на использование трех полных гидролизов. Расщепление после гуаниновых звеньев достигается гидролизом T_1-рибонуклеазой. Специфического расщепления после цитидинов можно достичь действием панкреатической рибонуклеазы на полинуклеотид, модифицированный карбодиимидом [11]. Полный гидролиз после уридинов осуществляется, как показано недавно [12—14], действием панкреатической РНКазы на полинуклеотид, модифицированный смесью бисульфита с метоксиамином. Использование каждого алгоритма иллюстрируется примером. Ниже приведена общая структура статьи.

§ 1. Когда можно обойтись стандартными методами без использования перекрывающихся фрагментов?

§ 2. Как наилучшим образом использовать информацию нескольких полных гидролизов?

Iа. Какая часть продуктов нужна?

Iб. Нужно ли секвенировать все отобранные продукты?

IIа,б. Как построить одну из возможных последовательностей?

IIв. Единственна ли она?

§ 3. Путь к единственности — использование частичных гидролизов.

I. Какие фрагменты нужны?

II. Как это проверить?

III. Как построить последовательность, содержащую данный фрагмент?

IV. Единственна ли новая последовательность?

§ 4. Что это все даст для 5S-РНК и фрагмента РНК фага Q_β?

§ 5. Какова вероятность обойтись только полными гидролизами?

В настоящей статье не рассматриваются трудности, могущие возникнуть из-за не 100%-ной специфичности РНКазных расщеплений и возможной гетерогенности образца. Однако использование предложенных в статье алгоритмов помогает выявить возникающие из-за этих трудностей ошибки.

§ 1. Метод деления на три части

Определение первичной структуры РНК блочным методом использует перекрывание нуклеотидных последовательностей фрагментов, полученных в результате различных гидролизов. Нуклеотидная последовательность фрагментов устанавливается для каждого гидролиза, и, таким образом, последовательность нуклеотидов в участках перекрывания определяется несколько раз. Устранение такого дублирования является одним из интересующих нас вопросов.

Пусть несколько объектов произвольной природы следует упорядочить, причем два из них, начальный и конечный в искомой последовательности, известны. Очевидно, что если число объектов не больше трех, то третий объект однозначно располагается между начальным и конечным объектами последовательности. Под объектами мы можем понимать как отдельные нуклеотиды, так и нуклеотидные последовательности.

Разработанные в настоящее время методы полного энзиматического гидролиза по каждому из остатков гуаниловой, цитидиловой [21] и уридиловой кислот [13—14] и модификации 5′- и 3′-концов нуклеотидных пос-

ледовательностей [7] позволяют использовать эту простую идею деления на три части для определения нуклеотидных последовательностей длиной более 8 нуклеотидов (процедура 5). Причем модификация концов не обязательно должна быть полной (стехиометрической), достаточна любая степень модификации, позволяющая идентифицировать 5′- и 3′-концевые фрагменты.

Для применимости метода необходимо, чтобы в нуклеотидном составе рассматриваемого фрагмента с мечеными 5′- и 3′-концами один из нуклеотидов, по которым может быть проведен полный гидролиз (т. е. G, C и U), встречался не более двух раз, исключая его 3′-концевое положение. Путем полного гидролиза фрагмента по выбранному нуклеотиду его разбивают на олигонуклеотиды, число которых на 1 больше кратности выбранного нуклеотида, т. е. $\leqslant 3$, и которые, таким образом, однозначно упорядочиваются.

Пусть $n_1 \leqslant n_2 \leqslant n_3$ — соотношение между количествами 3 нуклеотидов, по каждому из которых может быть проведен полный гидролиз, в нуклеотидном составе рассматриваемого фрагмента, взятом без его 3′-концевого нуклеотида. Необходимым (но не достаточным!) условием последовательного использования «метода деления на три части» является соблюдение соотношения: $n_1 \leqslant 2$; $n_2 \leqslant (n_1 + 1) \times 2$; $n_3 \leqslant (n_2 + 1) \times 2$. Заметим, что имеющиеся методы расщепления позволяют проводить специфическое «деление на три части» по аденину только после гидролиза по G [13, 14, 22].

Если «метод деления на три части» применим для данного фрагмента и его частей, то в конце концов можно получить короткие олигонуклеотиды, последовательность которых определима при помощи стандартных процедур [7]. «Методом деления на три части» могут однозначно определяться последовательности значительной длины (50 и более нуклеотидов).

Для оценки максимальной длины фрагмента, к которому приложим «метод деления на три части», воспользуемся простым рассуждением. Предположим, что у нас имеется продукт частичного гидролиза по одному из нуклеотидов, например по остаткам цитидиловой кислоты, содержащий, например, всего лишь два уридиновых звена; пусть полученные делением на три части олигонуклеотиды содержат по два не 3′-концевых цитидина, а полученные их делением на три части олигонуклеотиды — по 2 остатка гуанозина; тогда олигонуклеотиды, полученные гидролизом по гуаниловой кислоте, будут в не 3′-концевом положении иметь только адениловые звенья, а последовательность всего фрагмента имеет вид:

$$\mathrm{p(A)^{k_1}\,G(A)^{l_1}\,G(A)^{m_1}\,C(A)^{k_2}\,G(A)^{l_2}\,G(A)^{m_2}\,C(A)^{k_3}\,G(A)^{l_3}\,G(A)^{m_3}\,U(A)^{k_4}\cdot}$$
$$\mathrm{\cdot G(A)^{l_4}\,G(A)^{m_4}\,C(A)^{k_5}\,G(A)^{l_5}\,G(A)^{m_5}\,C(A)^{k_6}\,G(A)^{l_6}\,G(A)^{m_6}\,U(A)^{k_7}\cdot}$$
$$\mathrm{\cdot G(A)^{l_7}\,G(A)^{m_7}\,C(A)^{k_8}\;G(A)^{l_8}\,G(A)^{m_8}\,C(A)^{k_9}\,G(A)^{l_9}\,G(A)^{m_9}C_{OH},}$$

где $(A)^{k_1}$ — гомонуклеотидный блок длиной k_1.

Очевидно, что если неконцевых продуктов полного гидролиза больше одного, но они все равны, то их перестановка не изменит нуклеотидной последовательности, и в этом случае деление также приведет к однозначному восстановлению последовательности [23].

Если «метод деления на три части» к фрагменту не приложим, то для определения его нуклеотидной последовательности необходим учет информации о перекрывающихся продуктах двух или более полных гидролизов этого фрагмента. Как будет показано ниже, в этом случае для ряда фрагментов, получаемых полным гидролизом, также можно избежать повторного определения нуклеотидной последовательности участков перекрывания.

§ 2. Обработка информации нескольких полных гидролизов

Оценки показывают (см. § 5), что два полных гидролиза по одному нуклеотиду в 80—90% случаев дают достаточно информации для однозначного определения последовательности фрагментов длиной до 15—20 нуклеотидов, а три полных гидролиза позволяют в 80—90% случаев однозначно восстановить последовательность фрагментов длиной до 30—40 нуклеотидов.

Процесс обработки информации, даваемой несколькими полными гидролизами, делится на несколько этапов.

1. Сведение к минимуму биохимической работы при анализе продуктов полных гидролизов. С этой целью

анализируется нуклеотидный состав каждого продукта полного гидролиза для отбора «внешних» олигонуклеотидов, характеризующихся тем, что они не содержатся ни в одном продукте других полных гидролизов; остальные олигонуклеотиды в дальнейшем не потребуются (см. замечание 1);

по нуклеотидным последовательностям «внешних» продуктов одного из полных гидролизов, определяемым описанными в работе [7] методами, и нуклеотидному составу остальных «внешних» олигонуклеотидов анализируется возможность избежать дублирования при определении нуклеотидной последовательности участков перекрывания;

по нуклеотидным составам «внутренних» продуктов контролируется правильность определения последовательностей во «внешних» продуктах.

2. Для решения задач Б, В и Г (см. «Постановка задач...») все «внешние» олигонуклеотиды после определения их нуклеотидной последовательности упорядочиваются в «схему», по которой мы, согласно строгим критериям, сможем отвечать на поставленные вопросы.

Замечание 1. «Внутренние» концевые продукты могут оказаться полезными при определении нуклеотидной последовательности «внешнего» концевого продукта, так как набор «внутренних» концевых продуктов, полученных в результате различных полных гидролизов, эквивалентен части всех продуктов неполного экзонуклеазного гидролиза «внешнего» концевого продукта [7]. В ряде случаев, чтобы определить последовательность во «внешнем» концевом продукте, достаточно знать лишь его нуклеотидный состав и составы соответствующих концевых «внутренних» продуктов (или их последовательностей, если они определимы при разделении продуктов по их положению на двумерной хроматограмме). Нуклеотидная последовательность «внутренних» неконцевых продуктов представляет интерес лишь в редких случаях для контроля правильности определения последовательности во «внешних» продуктах. Во избежание совершенно напрасных затрат труда специально секвенировать «внутренние» продукты не следует до окончания процесса определения последовательностей во «внешних» продуктах и контроля правильности этих последовательностей по нуклеотидному составу «внутренних» продуктов или их последовательностям, полученным при двумерной хроматографии. Если результаты не вызывают серьезных сомнений, секвенирование «внутренних» продуктов проводить не нужно.

Iа. Выделение «внешних» продуктов

Нас интересуют только такие продукты каждого полного гидролиза, каждый из которых не содержится ни в одном продукте других полных гидролизов, так как каждый такой продукт несет всю информацию о нуклеотидной последовательности содержащихся в нем олигонуклеотидов. Эти интересующие нас продукты полных гидролизов мы назовем «внешними».

Таблица 1

Нуклеотидный состав продуктов трех полных гидролизов * фрагмента РНК из цистрона белка оболочки бактериофага R17

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.
1	G, U, pC, A	1	+	9	U, pC, A	1		22	pC	1	
2	G	2		10	3G, U, C	1	+	23	2G, U, C, A	1	+
3	G, C	1		11	U	7		24	G, 2U, C	1	+
4	G, 2U, C	1	+	12	G, U, C	1	+	25	G, U, C, A	1	+
5	G, 5U, C, 5A	1	+	13	U, C, A	1		26	2G,6U, C,8A	1	+
6	G, 9U, 4C, 7A	1	+	14	U, 3A	1		27	3U, C, A	1	
7	G, U, 2C, A	1	+	15	U, A	2		28	C	2	
8	G_{OH}, U, 3C, 2A	1	+	16	2G, U, 2A	1		29	4U, C, 2A	1	
				17	U, C, 2A	1		30	G, C	1	
				18	U, 2C, 2A	1		31	U, C, A	1	
				19	G, U, 2C	1	+	32	G, C, 2A	1	
				20	G, U, 2C, 3A	1	+	33	U, C	1	
				21	G_{OH}, 2C	1		34	G_{OH}	1	

* Для каждого продукта полного гидролиза приведены: в графе «№» — его числовой код, в графе «кратн.» — количество продуктов с одинаковой структурой; в графе «внешн.» плюсом отмечены «внешние» продукты полных гидролизов.

Очевидно, что любой олигонуклеотид, кроме 3′-концевого, не содержащий нуклеотидов, по которым расщепляет один из гидролизов, составляет часть олигонуклеотида — продукта указанного гидролиза, т. е. не является «внешним».

Например, полученный G-гидролизом олигонуклеотид AUG не содержит цитидинов. В любой достаточно длинной последовательности найдется цитидин слева и справа (или хотя бы только справа) от олигонуклеотида AUG, т. е. найдется последовательность вида CXG-AUG-YC (где X и Y — последовательности, не содержащие цитидин). Олигонуклеотид XGAUGYC, получаемый C-гидролизом, окажется «внешним», содержащим «внутренний» олигонуклеотид AUG.

Таким образом, «внешними» *не 3′-концевыми являются* только те *не 3′-концевые продукты полного гидролиза, нуклеотидный состав которых содержит все нуклеотиды, по которым проводятся все используемые в работе полные гидролизы* (G, U, C). И только для этих продуктов следует определять нуклеотидную последовательность.

«Внешним» 3′-концевым продуктом является самый длинный из 3′-концевых продуктов.

В качестве примера здесь и далее будем рассматривать гипотетическое определение нуклеотидной последовательности участка цистрона белка оболочки бактериофага R17 [24] при помощи трех полных гидролизов:

$$pCAUGGCGUUCGUACUUAAAUAUGGAAUUAACUAUUCCAAU-$$
$$-UUUCGCUACGAACUCCG_{OH}.$$

Из табл. 1 видно, что из 34 различных продуктов полных гидролизов только 14 являются «внешними», нуклеотидную последовательность которых нужно определять. Остальные 20 видов продуктов полных гидролизов не нужны для определения нуклеотидной последовательности фрагмента (см. замечание 1).

16. Участки перекрывания и определение их последовательности

Построение нуклеотидной последовательности молекулы или фрагмента по перекрывающимся продуктам полных гидролизов с известной нуклеотидной последовательностью требует выделения 5′- и 3′-концевых участков перекрывания в каждом «внешнем» олигонуклеотиде.

Часть олигонуклеотида, примыкающая к его 5′-(3′)-концу, которая отщепляется символически проведенным полным гидролизом, является левым (правым) участком перекрывания данного олигонуклеотида с олигонуклеотидами гидролиза, тождественного проведенному символически.

Нас интересует не любое левое (правое) перекрывание, а только максимальной длины левое (правое) перекрывание, содержащее все левые (правые) перекрывания данного олигонуклеотида с продуктами других гидролизов. Например:

UACUUAAAUAUG →

→ U·AC*U·U·AAAU·AU·G →

UAC*UUAAAUAUG

Здесь знак · соответствует символическому U-гидролизу, знак * — символическому C-гидролизу. Во второй строке подчеркнуты все участки перекрывания, в последней — максимальные левое и правое перекрывания олигонуклеотида. Далее такие максимальные перекрывания будем называть «внешними».

При использовании трех полных гидролизов для выделения левого внешнего перекрывания следует двигаться вдоль олигонуклеотида слева направо от его 5′-конца до тех пор, пока не встретятся два из трех видов нуклеотидов, по которым проводятся гидролизы (не встретится тот, что на 3′-конце олигонуклеотида), и остановиться после первый раз встреченного нуклеотида второго вида; для выделения правого перекрывания следует двигаться справа налево вдоль олигонуклеотида от его 3′-конца до тех пор, пока не встретятся два из трех видов нуклеотидов, по которым проводятся гидролизы, и остановиться перед первый раз встреченным нуклеотидом третьего вида.

В 5′-концевом «внешнем» продукте выделяем только правый участок перекрывания, в 3′-концевом «внешнем» — только левый.

Поскольку участок перекрывания является общим для двух продуктов гидролизов, при известной нуклеотидной последовательности одного из них часто оказывается возможным установить для другого без секвенирования часть или всю его нуклеотидную последовательность.

Если участок перекрывания имеет характерный нуклеотидный состав, мы с большой вероятностью можем определить, в каком из «внешних» продуктов другого полного гидролиза содержится этот участок перекрывания, не определяя нуклеотидную последовательность продукта.

Пусть определена нуклеотидная последовательность продуктов одного гидролиза, которые будем называть просто олигонуклеотидами, а продукты других полных гидролизов заданы только их нуклеотидным составом. Задача состоит в определении нуклеотидной последовательности участков перекрывания (по крайней мере) продуктов всех гидролизов, не прибегая по возможности к дополнительным биохимическим исследованиям.

1. Выделяем в олигонуклеотидах левые и правые участки перекрывания. Каждому левому (правому) участку перекрывания с данной нуклеотидной последовательностью поставим в соответствие число олигонуклеотидов, имеющих левый (правый) участок перекрывания с той же нуклеотидной последовательностью. Составим список этих участков, располагая вначале более длинные и с более уникальным нуклеотидным составом.

2. Для каждого левого (правого) участка перекрывания найдем продукты другого гидролиза, возможно его содержащие, по признакам:

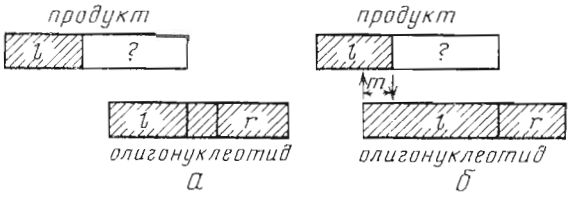

Рис. 1. Иллюстрация условий, при которых следует применять алгоритм 2в. Продукты полных гидролизов изображены прямоугольниками. Части продуктов с известной нуклеотидной последовательностью заштрихованы. Известные левые и правые перекрывания продуктов обозначены буквами l и r соответственно. Части продуктов с неизвестной последовательностью отмечены знаком «?». Нуклеотидные последовательности уже известных левых перекрываний продукта и олигонуклеотида не имеют общей части (*а*), имеют общую часть длиной m нуклеотидов (*б*)

а) если рассматривается левое перекрывание, то содержащий его продукт следует искать среди продуктов X-гидролиза, где X — крайний правый нуклеотид рассматриваемого перекрывания; если рассматривается правое перекрывание, то содержащий его продукт должен находиться среди продуктов Y-гидролиза, где Y — нуклеотид, предшествующий рассматриваемому перекрыванию в олигонуклеотиде;

б) нуклеотидный состав искомого продукта должен включать нуклеотидный состав рассматриваемого перекрывания;

в) пусть в рассматриваемом продукте одно из перекрываний, например левое (правое), уже известно; если сумма длин этого перекрывания продукта и левого (правого) перекрывания олигонуклеотида не превосходит длины рассматриваемого продукта (см. рис. 1, *а*), тогда нуклеотидный состав искомого продукта должен содержать сумму нуклеотидных составов его известного левого (правого) перекрывания и левого (правого) перекрывания олигонуклеотида; если же указанная сумма длин больше длины продукта на m (рис. 1, *б*), тогда последовательность m нуклеотидов на 5′-(3′)-конце левого (правого) перекрывания олигонуклеотида должна совпадать с 3′-(5′)-концевой последовательностью из m нуклеотидов в левом (правом) известном перекрывании искомого продукта и нуклеотидный состав этого продукта должен содержать сумму нуклеотидных составов его известного перекрывания и соответствующего перекрывания олигонуклеотида без m 5′-(3′)-концевых нуклеотидов рассматриваемого олигонуклеотида;

г) в случае идентификации по правому перекрыванию искомый продукт должен на 5′-конце иметь 5′-концевой нуклеотид участка перекрывания; этот признак можно использовать, если проведен 5′-концевой анализ соответствующих продуктов.

3. Если имеется перекрывание, для которого количество продуктов, удовлетворяющих указанным условиям, равно количеству олигонуклеотидов с данным перекрыванием, то все эти продукты действительно содержат рассматриваемое левое (правое) перекрывание и поэтому не содержат других левых (правых) перекрываний. Нуклеотидная последовательность их участков перекрывания, таким образом, определена. Чтобы учесть это, вычеркиваем из списков продуктов, возможно содержащих другие левые (правые) перекрывания, продукты, перекрывания которых мы определили. Возвращаемся к п. 2 и переходим к следующему перекрыванию. Если количество указанных продуктов больше, то нужна дополнительная проверка 4.

4. Для одного из продуктов, отобранных тестом 2, определяем нуклеотидную последовательность биохимически и, если она содержит одно из выписанных ранее перекрываний, вычеркиваем этот продукт из списков продуктов, возможно содержащих другие перекрывания, и возвращаемся к п. 3; если же она содержит новый тип перекрывания, то для этого перекрывания проводим процедуру 1, после чего снова производим проверку

условий 2, с учетом нового олигонуклеотида с известной последовательностью и т. д., пока не придем к однозначному ответу.

Использование проведенного в п. 1 иерархического распределения участков перекрывания желательно для уменьшения количества дополнительной работы (п. 4), так как вероятность однозначного результата поиска перекрывающихся продуктов значительно возрастает для более специфичных участков перекрывания.

Пример. Чтобы нагляднее продемонстрировать описанные выше алгоритмы, предположим, что нам известен только нуклеотидный состав каждого продукта и что мечеными являются только 5′- и 3′-концевые нуклеотиды молекулы.

Пусть вначале мы определили нуклеотидную последовательность «внешних» продуктов G-гидролиза (см. табл. 2). Последовательность в 5′-концевом продукте 1 определяется по нуклеотидному составу этого продукта и нуклеотидному составу 5′-концевого продукта 9 (см. замечание 1). Наиболее длинные участки перекрывания содержатся в олигонуклеотидах 5, 6, 8. Правое «внешнее» перекрывание продукта 5 и левое продукта 6 (обозначим их соответственно $5r$ и $6l$) должны содержаться в продуктах G-гидролиза (см. выше почему), а так как из табл. 1 следует, что длина каждого из этих перекрываний превосходит длину всех олигонуклеотидов G-гидролиза, кроме 26-го, то 26-й содержит оба эти перекрывания (так как 26-й не концевой продукт).

Длина (26) = длина ($5r$) + длина ($6l$) + 1,
нукл. состав (26) = нукл. состав ($5r$) + G + нукл. состав ($6l$)
поэтому 26 = $5r$-G-$6l$ = UUAAAUAUG·G·AAUUAAC.

Аналогично определяется последовательность 20 и 24:

длина (20) = длина ($7r$) + длина ($8l$),
нукл. состав (20) = нукл. состав ($7r$) + нукл. состав ($8l$),
20 = $7r$-$8l$ = ACG·AACU;
длина (24) = длина ($4l$) + 1,
нукл. состав (24) = G + нукл. состав ($4l$),
24 = G-$4l$ = GUUC.

Для продуктов 23 и 25 ситуация равнозначная по отношению к 1. Поэтому нуклеотидную последовательность для одного из них следует определить биохимически. Пусть она определена для 25. Тогда

длина (23) = длина ($1r$) + 2,
нукл. состав (23) = нукл. состав ($1r$) + G + C,
23 = AUGGC (так как 23 получен гидролизом по C);
длина (10) = длина ($23r$) + 2,
нукл. состав (10) = нукл. состав ($23r$) + G + U,
10 = $23r$-G-U = GGCGU и т. д.

Если вместо секвенирования продукта 25 провести определение его 5′-концевого нуклеотида, то использование признака «2г» сразу позволяет установить, что правое перекрывание олигонуклеотида 1 может являться левым перекрыванием только продукта 23. Таким образом, для нашего примера из 14 «внешних» олигонуклеотидов приходится определять нуклеотидную последовательность биохимическими методами *всего лишь для 5—6 «внешних» продуктов*. Все «внешние» олигонуклеотиды с подчеркнутыми участками перекрывания приведены в табл. 2. Теперь легко, используя для контроля табл. 1 и 2, проверить правильность последовательностей во «внешних» продуктах. Для этого во «внешних» продуктах (табл. 2) путем проведения символических полных гидролизов выделим «внутренние» и сравним их нуклеотидный состав и кратность с нуклеотидным составом и кратностью «внутренних» продуктов из табл. 1. Они должны совпадать.

Таблица 2

Нуклеотидная последовательность «внешних» продуктов полных гидролизов фрагмента РНК бактериофага R17

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.
1	pCAUG		*r*(9)	10	GGCGU		*l*(23)	23	AUGGC		*l*(1)
4	UUCG		*x*	12	CGU		*l*(4,6)	24	GUUC		*r*(4)
5	UACUUAAAUAUG		*x*!	19	CGCU		*r*(7,12) *l*(4,6)	25	GUAC		*xr*(5)
6	AAUUAACUAUUCCAAUUUUCG		*x*!	20	ACGAACU		*r*(8) *l*(7)	26	UUAAAUAUGGAAUUAAC		*r*(6) *l*(5)!
7	CUACG		*x*								
8	$AACUCCG_{OH}$		*x*								

Примечание. В нуклеотидной последовательности каждого «внешнего» продукта подчеркнут левый участок перекрывания, жирным шрифтом выделен правый участок. «*x*» означает, что секвенирование проводилось биохимическими методами; знаком «!» отмечены продукты длиной более 8 нуклеотидов, которые трудно секвенировать стандартными процедурами; *l*(4,6) означает, что левый участок перекрывания данного продукта восстанавливался по уже известной структуре продуктов, номера которых указаны в скобках. Аналогичен смысл обозначения *r*(7,12), где *r* — правое перекрывание данного продукта.

IIа. Построения «схемы»

Для каждого вида «внешнего» перекрывания составим два множества: M и S. В множество M включим числа, кодирующие все те (не 3′-концевые) «внешние» олигонуклеотиды, у которых рассматриваемое перекрывание является правым, а в множество S — кодирующие те (не 5′-концевые), у которых это перекрывание левое. В отдельные множества M_0 и S_0, состоящие из одного элемента, включаем соответственно 3′- и 5′-концевой «внешний» олигонуклеотид. Числа, кодирующие олигонуклеотиды множеств M и S, выписываем в две строки, причем числа из M выписываем в верхней строке, а из S — в нижней. Часть записи, соответствующую каждому перекрыванию, выделяем вертикальными линиями. (Если имеются кратные олигонуклеотиды с идентичными последовательностями, обозначаем их одним и тем же числом с индексами.)

Для нашего примера полученная запись имеет вид:

1	4, 6	5	7	10,12	19	20	23	24	25	26	8
1	23	12, 19	26	20	25,24	7	8	10	4	5	6	

Запись эту мы называем «схемой» молекулы (фрагмента), а каждую ее часть, выделенную вертикальными линиями,— «блоком» схемы. Соответствующее перекрывание назовем «перекрыванием блока». Схема содержит всю информацию, которую могут дать для восстановления нуклеотидной последовательности молекулы перекрывающиеся продукты полных гидролизов.

Любая построенная по указанным правилам схема будет обладать следующими свойствами (см. [16, 17, 25]):

все перекрывания блоков являются «внешними» перекрываниями, а все элементы блоков — «внешними» (по построению);

количества элементов в множествах M и S одного блока равны (исключая блоки, содержащие только начальный или конечный элементы последовательности);

все элементы, содержащиеся в верхней строке схемы (множествах M), содержатся и в нижней строке схемы (множества S), и наоборот;

множества M любых двух разных блоков не содержат общих элементов; то же верно относительно множеств S разных блоков;

для каждого блока всегда найдется другой блок, такой, что множество S первого и множество M второго имеют общие элементы (свойство связности).

Если схема не имеет указанных свойств, значит, она составлена неправильно.

IIб. Построение возможной последовательности

Совершенно ясно, что нуклеотидной последовательности, в которой выделены «внешние» продукты полных гидролизов, однозначно соответствует последовательность кодов этих продуктов (т. е. кодирующих эти продукты чисел). Верно и обратное. Например,

$$\underline{\underline{\text{CGCUACGAACUCCG}}}_{\text{OH}}$$
$$19—7—20—8$$

Поскольку мы работаем с кодами, построим последовательность элементов схемы — кодов «внешних» продуктов полных гидролизов, возможную при информации, заключенной в схеме.

Рассмотрим один произвольный блок (отличный от M_0 или S_0). Любой элемент верхней строки блока, т. е. множества M, кодирует олигонуклеотид, который перекрывается по перекрыванию блока с любым олигонуклеотидом, закодированным элементом множества S нижней строки того же блока. Какой элемент верхней строки перекрывается в искомой

последовательности с некоторым элементом нижней строки блока (соответствует ему), нам пока неизвестно. Именно многозначность соответствия и порождает неединственность структуры, возможной при данной информации (см. также ниже).

а) Установим *произвольным образом* однозначное соответствие между элементами верхних и нижних строк каждого блока так, чтобы каждому элементу из M соответствовал лишь один элемент из S и чтобы все элементы множества S исчерпывались. Удобно, например, условиться, что элементу верхней строки соответствует элемент, стоящий непосредственно под ним.

б) Найдем то множество M, в котором находится элемент, кодирующий 5′-концевой олигонуклеотид (beg) молекулы, и выпишем его в начале строки; элемент, стоящий под beg, выпишем в строке за ним; найдем этот последний элемент в одном из множеств M и выпишем за ним в строке элемент, стоящий в схеме под ним, и т. д., пока не выпишем в строку элемент, кодирующий 3′-концевой олигонуклеотид (end) молекулы. Отмечаем на схеме использованные элементы.

Если оказалось, что в строке выписаны все элементы схемы, значит, построенная последовательность элементов является одной из возможных при данной информации.

Если же часть элементов не вошла в строку, достраиваем ее «вставками» следующим образом:

в) начинаем писать новую строку, переписывая первую слева направо до тех пор, пока не дойдем до *элемента, находящегося в одном множестве S схемы с элементом, еще не использованным*;

г) подчеркнем этот элемент в старой строке, а в новой строке *вместо него выписываем* находящийся с ним в одном множестве S *неиспользованный элемент*;

д) находим то множество M в схеме, в котором находится крайний правый элемент новой строки, и элемент, стоящий под ним (являющийся неиспользованным), выписываем в качестве следующего элемента строящейся строки (по ходу дела отмечая в схеме использованные элементы), и т. д. до тех пор, пока не дойдем до неиспользованного элемента, под которым в схеме стоит элемент, с которого началось отличие второй строки от первой; тогда вместо этого элемента записываем во второй строке подчеркнутый элемент первой строки и продолжаем ее переписывать, пока снова не дойдем до элемента, находящегося в одном множестве S с еще не использованным, и тогда пункты «г» и «д» повторяются.

Если мы дошли до конечного элемента (end), но не все элементы схемы выписаны в новой строке, то повторяем по отношению к этой строке пункты «в», «г», «д» и т. д., пока не выпишем в строку все элементы схемы.

Построенная последовательность является одной из нескольких, которые можно построить по данной схеме. По-разному устанавливая соответствия между элементами множеств M и S, получаем разные допустимые последовательности. *Однако* по одной построенной последовательности *можно* построить все допустимые последовательности.

Для нашего примера в первой строке оказывается выписанной последовательность: 1-23-10-25-5-26-6-19-7-20-8, а во второй: 1-23-10-24-4-12-25-5-26-6-19-7-20-8.

IIв. Проверка единственности построенной структуры и выявление участков неоднозначности

Необходимым и достаточным условием единственности построенной нами последовательности элементов схемы является: отсутствие в последовательности двух пар элементов i и j, k и l, расположенных в ней в следующем порядке:

$$\ldots\ldots\underbrace{i\ldots}_{1}k\ldots\underbrace{j\ldots}_{2}l\ldots$$

(точки означают невыписанные элементы последовательности), и таких, что i и j находятся в одном множестве S и кодируемые ими олигонуклеотиды не равны; k и l также находятся в одном множестве S. Четверка элементов может вырождаться в тройку элементов из одного множества S при $k = j$.

В случае присутствия указанных двух пар элементов в построенной нами последовательности переставимы части последовательности, обозначенные скобками с цифрами 1 и 2. Выделением всех троек и четверок элементов с указанными свойствами в любой построенной последовательности и перестановкой соответствующих ее кусков можно из одной последовательности, если это желательно, получить все последовательности, допустимые при информации, заданной набором продуктов полных гидролизов молекулы.

Пример:

1-23-10-24-4-12-25-5-26-6-19-7-20-8,

перестановка приводит к

1-23-10-25-5-26-6-12-24-4-19-7-20-8.

Для предварительной проверки единственности последовательности, задаваемой схемой молекулы без построения последовательности, часто удобно пользоваться следующим достаточным условием неединственности: если в каком-либо множестве S содержится n элементов, кодирующих n олигонуклеотидов, и среди последних более чем один отличается от всех прочих, то последовательность *не единственна*. Если же только один олигонуклеотид в каждом таком множестве отличается от остальных, следует проверить *необходимое и достаточное* условие единственности.

§ 3. Путь к единственности — учет информации частичных гидролизов

I. Некоторые идеи и оценки

Ясно, что неединственность последовательности, задаваемой информацией полных гидролизов, определяется множествами S, содержащими элементы i, j, k, l (см. выше). Назовем эти множества множествами неоднозначности, а содержащие их блоки — блоками неоднозначности. Мы видели, что указанная четверка (или тройка) элементов порождает две последовательности. В одной из них элементу i предшествует некоторый элемент a, а в другой за элементом a следует уже элемент j. Если бы у нас была дополнительная информация о том, какой из элементов, i или j, следует за a в реальной последовательности, это устранило бы неоднозначность, заключенную в данном множестве неоднозначности. Формально мы можем учесть такую информацию, выделив в блоке неоднозначности новый блок, содержащий элементы a и i. Последовательное использование информации такого рода приведет к дроблению блоков неоднозначности схемы молекулы до тех пор, пока новая «модифицированная схема» будет определять только одну последовательность.

Анализ схемы молекулы длиной более 50 нуклеотидов для трех полных гидролизов часто указывает на неединственность нуклеотидной последовательности, восстанавливаемой по данной информации. Возникает необходимость в дополнительной информации, которую можно получить в виде продуктов частичных гидролизов.

Так как большинство случайно выбранных последовательностей длиной 30—40 нуклеотидов однозначно определяется по продуктам трех их полных гидролизов, имеет смысл рассматривать продукты частичного гидролиза именно такой длины. Продукты частичного гидролиза должны быть достаточно длинными (30—40 нуклеотидов) также и для того, чтобы,

будучи случайно локализованными в последовательности длиной 100—150 нуклеотидов, они могли с большой вероятностью включить в себя олигонуклеотиды, кодируемые элементами блоков неоднозначности. В дальнейшем для краткости будем говорить об олигонуклеотидах блока неоднозначности или схемы. Короткие продукты частичного гидролиза могут включать в себя лишь очень малое число «внешних» продуктов полного гидролиза из-за большой длины последних (в среднем около 8 нуклеотидов) и поэтому обычно почти не несут дополнительной информации. В ряде случаев можно с достаточной определенностью указать, какой именно продукт частичного гидролиза требуется для достижения однозначности.

Например, *в построенной нами* последовательности один из элементов блока неоднозначности, а именно 19, начинается на расстоянии 14 нуклеотидов от 3′-конца последовательности:

pCAUGGCGUUCGUACUUAAAUAUGGAAUUAACUAUUCCAAUUUUCGCUACG-

-AACUCCG$_{OH}$.

Поэтому, если мы выделим продукт частичного гидролиза нашего полинуклеотида, включающий его 3′-конец, длиной в 25—30 нуклеотидов, то наверняка получим информацию о том, какой именно олигонуклеотид перекрывается в последовательности с олигонуклеотидом 19. Среди пяти полученных авторами работы [24] продуктов только один (T_{II3}) удовлетворяет описанным требованиям.

II. Идентификация олигонуклеотидов фрагмента

Нас *должны интересовать* продукты всех тех полных гидролизов фрагмента, с помощью которых были получены олигонуклеотиды блоков неоднозначности. Однако проводить все эти полные гидролизы не обязательно.

1. Пусть получен некоторый продукт частичного гидролиза и нас интересует, содержит ли он олигонуклеотиды из блоков неоднозначности. Проводим один из тех гидролизов интересующего нас фрагмента, с помощью которого были получены олигонуклеотиды блоков неоднозначности.

2. Выделим среди не содержащих концевые метки фрагмента продуктов полных гидролизов «внешние» (см. Iа) и проведем идентификацию их с олигонуклеотидами схемы. В большинстве случаев такую идентификацию можно провести сравнением только нуклеотидных составов или положений продуктов на фингерпринте. Отмечаем на схеме молекулы идентифицированные олигонуклеотиды. Некоторые сложности могут возникнуть на этом этапе, если среди продуктов полного гидролиза молекулы имеются равные и не все они принадлежат фрагменту (см. замечание 2, ниже).

Переходим к анализу продуктов, содержащих 5′-(3′)-концевые метки.

3. Если среди 5′-(3′)-концевых продуктов полных гидролизов фрагмента имеется содержащий два из трех нуклеотидов, по которым проводились полные гидролизы молекулы, и полученный гидролизом не по тому нуклеотиду, которым оканчивается фрагмент, то будем считать этот продукт левым (правым) концевым продуктом фрагмента и попытаемся проидентифицировать его с олигонуклеотидами схемы (см. п. 4, ниже).

Если такой продукт отсутствует или его идентификацию провести не удалось, но имеется продукт с той же концевой меткой фрагмента и тремя видами нуклеотидов, по которым проводились гидролизы, то считаем концевым его. Если идентификацию провести удалось, то продукт с тремя видами нуклеотидов идентифицируем, но в дальнейших рассмотрениях считаем не концевым. Идентификация такого продукта проводится как в п. 2.

Может встретиться ситуация, когда 5′-(3′)-концевой продукт фрагмента содержит менее двух видов нуклеотидов, по которым проводились пол-

ные гидролизы молекулы. В этом случае пытаемся провести его идентификацию (см. п. 4).

4. Внешний олигонуклеотид молекулы, с которым идентифицируется левый (правый) концевой продукт длины l фрагмента, должен: а) быть получен тем же гидролизом, что и концевой продукт фрагмента; б) иметь длину не меньше l; в) иметь на l-м месте от 3′-(5′)-конца 5′-(3′)-концевой нуклеотид фрагмента и на $(l + 1)$-м месте для левого продукта нуклеотид, равный 3′-концевому нуклеотиду фрагмента; г) иметь нуклеотидный состав 3′-(5′)-концевой части длины l, совпадающий с нуклеотидным составом идентифицируемого концевого продукта фрагмента.

5. При идентификации концевых олигонуклеотидов возникают следующие ситуации:

а) левый и правый концевые олигонуклеотиды фрагмента идентифицируются с двумя внешними продуктами полных гидролизов молекулы;

б) левый (правый) концевой продукт фрагмента не идентифицируется, но идентифицируется его правое (левое) перекрывание, т. е. известен блок схемы молекулы, которому концевой продукт должен принадлежать;

в) левый (правый) концевой продукт и его участок перекрывания не идентифицируются.

Идентифицированный правый концевой продукт фрагмента отмечаем только в S-множестве схемы, левый — только в M-множестве.

6. Если проведены не все три полных гидролиза фрагмента, то для определения того, какие из продуктов еще не проведенных гидролизов принадлежат фрагменту, используются следующие правила:

а) если проведен один (два) полный гидролиз фрагмента и его продукты отмечены на схеме молекулы, то фрагменту будут принадлежать также те продукты двух (одного) других полных гидролизов, которые являются общими для каких-нибудь двух блоков, содержащих продукты *уже проведенных* полных гидролизов фрагмента;

б) если для каких-нибудь двух блоков, содержащих продукты уже проведенных полных гидролизов фрагмента, общим является более чем один продукт еще не проведенного гидролиза фрагмента, то в случае 5а на схеме молекулы в соответствующих блоках отмечается лишь часть этих продуктов, такая, что число отмечаемых продуктов еще не проведенного гидролиза должно быть равно числу уже отмеченных продуктов проведенного гидролиза. При этом, если все упомянутые продукты еще не проведенного гидролиза равны, отмечаем нужное их количество, если же не все равны, то для выяснения принадлежности каждого из них фрагменту используем тот факт, что «внутренние» продукты этих олигонуклеотидов, получаемые символическим гидролизом, соответствующим проведенному, должны встретиться среди продуктов этого уже проведенного гидролиза фрагмента.

Считается, что только отмеченные таким образом олигонуклеотиды принадлежат фрагменту. В случае 5б делается то же, что и в случае 5а, исключая блоки, могущие содержать неидентифицированные концы фрагмента.

В случае 5в возвращаемся к п. 1 и проводим полный гидролиз фрагмента по нуклеотиду, которым он оканчивается;

в) чтобы проверить, все ли продукты еще не проведенных гидролизов фрагмента выявлены при использовании правила 6а, следует построить схему фрагмента (§ 2, IIб) по внешним продуктам проведенного гидролиза и внешним продуктам, выявленным по правилу 6а или 6б; если по такой схеме удается построить непрерывную нуклеотидную последовательность с длиной, равной длине фрагмента, то уже найдены все входящие в фрагмент внешние продукты непроведенных гидролизов фрагмента; если такую последовательность построить не удается, возвращаемся к п. 1 и проводим дополнительный полный гидролиз фрагмента.

Для примера рассмотрим учет информации, содержащейся в фрагмен-

Таблица 3

Нуклеотидный состав продуктов полных гидролизов олигомеров, полученных частичным гидролизом фрагмента РНК бактериофага R17

	G-гидролиз					U-гидролиз				
№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ	№ в фр.	нукл. состав	кратн.	внешн.	№ в мол. / способ
				T_{II_3}						
Ф1	G, 9U, 4C, 6A, pA		+	6	*C*				+	19 G
Ф2	G, U, 2C, A		+	7	*C*				+	20 G
Ф3	G_{OH}, U, 3C, 2A		+	8	*C*					

Плюсом обозначены внешние продукты полных гидролизов; в графе «№ в мол.» указан номер данного продукта полного гидролиза в молекуле; *С* означает, что продукты полного гидролиза олигомера и фрагмента идентифицированы по составу. В графе «способ» знаки G, U, C или их сочетание указывают, что внешние продукты соответствующего гидролиза (наименование колонки), номера которых указаны в графе «№ в мол.», найдены по «схеме» согласно алгоритму по продуктам одного из гидролизов (G, U, C) или их комбинации.

те T_{II_3} РНК бактериофага R 17. Нас интересуют продукты только тех его полных гидролизов, которыми были получены олигонуклеотиды блоков неоднозначности. Для нашего примера таких блоков два. Рассмотрим блок неоднозначности схемы молекулы (§ 3, IIа)

$$\left|\begin{matrix} 4, & 6 \\ 12, & 19 \end{matrix}\right|,$$

содержащий олигонуклеотиды 4, 6, 12, 19. Олигонуклеотиды 4 и 6 получены G-гидролизом. Поэтому проводим G-гидролиз фрагмента. Нуклеотидный состав полученных продуктов приведен в табл. 3.

Сравнение нуклеотидных составов продуктов G-гидролиза фрагмента и всей молекулы показывает, что Ф1 = 6, Ф2 = 7, Ф3 = 8. Таким образом, идентификация проведена только по нуклеотидным составам и нам удалось идентифицировать концевые продукты фрагмента.

Отмечаем на схеме молекулы олигонуклеотиды фрагмента 6, 7 и 8. Олигонуклеотид 6, как начальный, отмечаем лишь в верхней, а 8, как конечный, — лишь в нижней строке схемы. Отмеченные элементы содержатся в блоках схемы

$$\left|\begin{matrix} 4, & \dot{6} \\ 12, & 19 \end{matrix}\right| \left|\begin{matrix} \dot{7} \\ 20 \end{matrix}\right| \left|\begin{matrix} 19 \\ \dot{7} \end{matrix}\right| \left|\begin{matrix} 20 \\ \dot{8} \end{matrix}\right|.$$

Каждый из олигонуклеотидов 19 и 20 оказывается общим для двух блоков, содержащих отмеченные элементы. Отмечаем олигонуклеотиды 19 и 20 на схеме молекулы.

Проверяем возможность построения связной последовательности фрагмента Ф1-19-7-20-Ф3, т. е. связная последовательность существует. Строим ее нуклеотидную последовательность:

$$AAUUAACUAUUCCAAUUUUCGCUACGAACUCCG_{OH}.$$

Длина этой последовательности равна длине фрагмента. Значит, мы определили (идентифицировали) все олигонуклеотиды, входящие в фрагмент. Это 6, 7, 8, 19, 20.

III. Учет информации частичного гидролиза и модификация схемы

На схеме молекулы отмечены идентифицированные олигонуклеотиды фрагмента. При этом в случае II.5а количества отмеченных элементов в верхней и нижней строках каждого блока равны. Подмножества отмеченных элементов выделяем в отдельный блок, разбивая старый блок на

две части. В случае II.5б нас не интересует блок, в котором должны содержаться левый (правый) концевой продукт фрагмента, а остальные блоки дробим так же, как в случае II.5а. В случае II.5в игнорируем блоки с разным числом отмеченных элементов в верхней и нижней строках, а с остальными блоками поступаем так же, как в случае II.5а. Если в случае II.5в во всех блоках содержится равное число отмеченных элементов, то информацию, содержащуюся в фрагменте, мы учесть не можем.

Пример. В одном из блоков неоднозначности схемы молекулы оказывается по одному отмеченному элементу в множествах M и S. Выделяем эти элементы в отдельный блок и строим модифицированную схему молекулы. Для нашего примера модифицированный блок неоднозначности имеет следующий вид:

$$\left|\begin{array}{c|c}4 & 6\\ 12 & 19\end{array}\right|,$$

а остальные блоки схемы остаются без изменения.

Легко видеть, что после этого в модифицированной схеме молекулы уже не найдется двух пар элементов, попарно принадлежащих к одному множеству S, и будет выполняться необходимое и достаточное условие единственности построенной последовательности. Построением легко также показать, что модифицированной схеме соответствует только первая из двух уже построенных нами последовательностей.

IV. Условие единственности

Для модифицированной схемы кроме проверки условий единственности, которая проводилась для схемы полных гидролизов (см. § 2, IIв), необходимо проверить отсутствие четверок или троек элементов, отличающихся от уже рассмотренных лишь тем, что элементы i, j кодируют равные олигонуклеотиды и входят в одно множество S, но в разные множества M. Если такие тройки или четверки элементов в последовательности находятся, но при перестановке местами кусков

$$\underbrace{\quad}_{1} \text{ и } \underbrace{\quad}_{2}$$

в исходной построенной последовательности получаем последовательность, кодирующую тот же олигонуклеотид, что и исходная, то построенная нами последовательность единственна.

Замечание 2. Если среди «внешних» продуктов полных гидролизов имеются равные, и только часть (скажем, k штук) из них принадлежит рассматриваемому фрагменту, и если это первый фрагмент, в котором содержатся такие олигонуклеотиды, то отмечаем на схеме молекулы любые k из этих олигонуклеотидов. Если же мы уже установили, что часть этих равных олигонуклеотидов принадлежит некоторому фрагменту (§ 3, II.6б), то для следующего содержащего такие олигонуклеотиды фрагмента мы уже не можем отмечать на уже модифицированной схеме молекулы *любые* k из таких равных олигонуклеотидов. Общий алгоритм состоит в этом случае в рассмотрении всех возможных способов выбора k равных олигонуклеотидов в модифицированной схеме молекулы и проверке возможности построения связной последовательности всей молекулы и рассматриваемого фрагмента по схеме, модифицируемой в соответствии с каждым возможным выбором. Для всех способов выбора, удовлетворяющих указанному тесту, следует проводить дальнейшее рассмотрение модифицированных схем молекулы. Часть из этих схем не будет удовлетворять тесту при анализе следующих фрагментов, содержащих указанные равные олигонуклеотиды. Таким образом, в конце концов будет достигнута однозначность [20]. В большинстве практически встречающихся случаев идентификацию равных олигонуклеотидов, принадлежащих фрагменту, можно провести значительно проще [19, 20]. Идея, на которой осно

вано большинство таких упрощений, состоит в том, что если один и тот же из равных олигонуклеотидов отмечается на схеме при анализе двух различных фрагментов, то он принадлежит им обоим и, значит, эти фрагменты в молекуле должны перекрываться. Если фрагменты действительно перекрываются, то в участок перекрывания обычно попадают также и олигонуклеотиды, встречающиеся один раз. Наличие в каждом из фрагментов таких олигонуклеотидов, общих для обоих фрагментов, проверить уже очень легко. Среди «внешних» продуктов трех полных гидролизов РНК длиной 100—200 нуклеотидов равных довольно мало. Поэтому практически всегда легко установить, перекрываются ли фрагменты в последовательности, путем простого сравнения олигонуклеотидов, входящих в оба фрагмента. При этом важно проверять вхождение концевого продукта одного фрагмента в другой для каждого из фрагментов.

Замечание 3. Все описанные в настоящей работе алгоритмы легко могут быть использованы для анализа данных, полученных четырьмя полными гидролизами. В этом случае «внешними» олигонуклеотидами будут олигонуклеотиды, содержащие все четыре нуклеотида, «внешние» перекрывания будут содержать три из четырех видов нуклеотидов и во всех алгоритмах под словами «все полные гидролизы» следует понимать четыре, а не три гидролиза. Описанные алгоритмы могут быть использованы с аналогичными поправками и для анализа результатов полученных традиционными гидролизами, но эффективность идентификации перекрываний будет существенно меньшей, чем для трех или четырех гидролизов, а идентификация олигонуклеотидов фрагментов может сильно усложниться из-за большого числа равных «внешних» продуктов двух полных гидролизов (см. замечание 2).

§ 4. Проверка предложенных методов на ранее установленных последовательностях 5S-РНК из *E. coli* и РНК фага Q_β

В этом параграфе при помощи предложенных в работе методов проанализированы нуклеотидные последовательности 5S-РНК из *E. coli* [26] и 5′-концевого фрагмента РНК фага Q_β [2]. Результаты этого анализа суммированы в таблицах. В табл. 4 и 7 указан нуклеотидный состав продуктов трех полных гидролизов этих полирибонуклеотидов и отмечены «внешние» продукты этих гидролизов. В табл. 5 и 8 приведены нуклеотидные последовательности «внешних» продуктов трех полных гидролизов и указаны методы, которыми могли бы быть определены эти последовательности. Под табл. 5 и 8 даны «схемы» соответствующих молекул с отмеченными блоками неоднозначности. Четверки элементов, порождающие неединственность последовательности, отмечены также на последовательности элементов, кодирующих «внешние» олигонуклеотиды. В табл. 6 и 9 приведен нуклеотидный состав продуктов полных гидролизов фрагментов, полученных частичными гидролизами 5S-РНК и фрагмента РНК фага Q_β, и указан способ идентификации этих продуктов с соответствующими продуктами полных гидролизов всей молекулы.

Продукты частичных гидролизов отбирались нами из числа приведенных авторами по признаку максимальности их длины, т. е. сначала анализировался самый длинный продукт случайно выбранного частичного гидролиза, затем самый длинный из оставшихся продуктов этого гидролиза и т. д.

Под табл. 6 и 9 даны модифицированные схемы молекул, учитывающие информацию использованных продуктов частичных гидролизов, и указано, в результате учета каких фрагментов произошло дробление блоков исходной схемы.

Таблица 4

Нуклеотидный состав продуктов трех полных гидролизов 5S-РНК

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.
1	G, pU	1		19	pU	1		38	G, pU, C	1	+
2	G, U, 2C	1	+	20	G, U, 2C	1	+	39	C	14	
3	G	11		21	5G, U, 3C	1	+	40	2G, U, C	1	+
4	G, C	5		22	4G, U, 2C, A	1	+	41	2G, C	1	
5	G, 2C	4		23	2G, U	2		42	2G, U, C, A	2	+
6	G, U, A	4		24	U, 4C, A	1		43	G, C	3	
7	G, U	5		25	G, U, 4C, 2A	1	+	44	4G, 2U, C	1	+
8	G, 2U, 5C, A	1	+	26	2G, U, 3C, 2A	1	+	45	C, A	1	
9	G, U, 4C, 2A	1	+	27	2G, U, C, 3A	1	+	46	G, U, C, A	1	+
10	G, U, 2C, 3A	1	+	28	G, U	2		47	G, U, C, A	2	+
11	G, 2A	1		29	3G, U, 3C, 3A	1	+	48	G, C, 2A	1	
12	G, C, 2A	1		30	3G, U, 3C, 2A	1	+	49	U, C	2	
13	G, U, A	1		31	G, U, A	1		50	3G, U, C, 6A	1	+
14	G, 3U, 5C, A	1	+	32	4G, U	1		51	9G, 5U, C, 2A	1	+
15	G, A	2		33	U, C	1		52	6G, U, C, 5A	1	+
16	G, U, C, 2A	1	+	34	U, 4C, A	1		53	G, U, C	1	+
17	G, 2C, A	1		35	4G, U, C, 2A	1	+	54	2G, C, A	1	
18	U_{OH}, C, A	1		36	3G, U, C, 3A	1	+	55	U_{OH}, A	1	
				37	3G, U_{OH}, 3C,2A	1	+				

Обозначения те же, что и в табл. 1.

§ 5. Можно ли обойтись только полными гидролизами?

Для любой лаборатории, занимающейся определением нуклеотидных последовательностей, ответ на вопрос «Какова вероятность того, что нуклеотидная последовательность длины N может быть однозначно восстановлена по данным только k полных гидролизов?» представляет практический интерес.

Если ответ на этот вопрос известен, то можно оценить время, которого потребует эксперимент в условиях конкретной лаборатории, и сразу использовать оптимальные методы расщепления. Если какие-нибудь методики недоступны или задачу нужно решить за ограниченное время, то можно ограничиться анализом последовательностей только определенной длины.

Для ответа на поставленный вопрос нами были использованы 30 известных последовательностей РНК [8] и описанные в настоящей работе критерии единственности (§ 2, IIв).

Для каждой последовательности проводилось символическое разбиение на продукты полного гидролиза отдельно по каждому из нуклеотидов, G, C, U и A, и эти продукты кодировались (см. § 2). Рассматривались следующие наборы полных гидролизов: два гидролиза — по G и C; три гидролиза — по G, C и U; четыре гидролиза — по G, C, U и A. Для каждого из наборов выделялись «внешние» олигонуклеотиды, строилась последовательность их кодов, соответствующая исходной нуклеотидной последовательности, и на ней отмечались олигонуклеотиды, порождающие неоднозначность (§ 2, IIв). Каждая последовательность разбивалась на фрагменты, соответствующие N нуклеотидам, двумя способами, различавшимися сдвигами на $N/2$ нуклеотидов, и для каждого фрагмента по критерию

Таблица 5

Нуклеотидная последовательность внешних продуктов полных гидролизов 5S-РНК

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.
2	CCUG	1	*l* (8, 14, 20)	20	GCCU	1	*x*	38	pUGC	1	СК
8	UCCCACCUG	1	*x*!!	21	GGCGGCCGU	1	*x*!	40	UGGC	1	*l* (2, 8, 4, 6)
9	ACCCCAUG	1	*l* (25)	22	AGCGCGGU	1	*x*	42	GUAGC	2	*r* (22, 30, 44) *l* (21, 10, 29)
10	AACUCAG	1	*r* (16, 27) *l* (9, 26, 36)	25	GACCCCAU	1	*x*	44	GGUGGUC	1	*r* (14, 38) *l* (22, 42)
14	UCUCCCCAUG	1	*x*!!	26	GCCGAACU	1	*x*	46	UGAC	1	*r* (25, 42)
16	AACUG	1	*l* (9, 26, 36)	27	CAGAAGU	1	*x*	47	AUGC	2	*l* (9)
				29	GAAACGCCGU	1	*x*!	50	AGAAGUGAAAC	1	*r* (29, 52) *l* (27)
				30	AGCGCCGAU	1	*x*!		AAC		
				35	GCGAGAGU	1	*x*	51	GAUGGUAGUGUGGGGUC	1	*x*
				36	AGGGAACU	1	*x*				
				37	$GCCAGGCAU_{OH}$	1	*x*	52	GAGAGUAGGGAAC	1	*r* (36, 8, 50) *l* (35, 51)
								53	UGC	1	*l* (2, 8, 46)

Примечание. Обозначения те же, что и в табл. 2. «!!» означает, что продукт полного гидролиза длиной более 8 нуклеотидов может быть секвенирован «методом деления на три части». «СК» означает, что последовательность определена по составу и данным концевого анализа.

К табл. 4 и 5

37	2, 8, 16 40, 46, 53	9, 14 47_1, 47_2	10 27	20 2	21, 29 42_1, 42_2	22 44	25 9	27 50	30 51	35 52	40 21	42_1, 42 22, 30	44, 51 8, 14	46 25	47_1, 47_2, 38, 53 26, 35, 20, 37	50 29	52 36	36, 26 16, 10	38

38-20-2-40-21-42_1-22-44-8-46-25-9-47_1-26-10-27-50-29-42_2-30-51-14-47_2-35-52-36-16-53-37

• — = + * — ~ • × = + * ~ • × — •

Примечание. Одинаковыми знаками отмечены элементы, входящие в одно множество *S*.

Таблица 6

Нуклеотидный состав продуктов полных гидролизов олигомеров, полученных частичными гидролизами 5S-РНК

G-гидролиз						U-гидролиз						C-гидролиз					
№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ	№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ	№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ
T_{20}																	
			+	10	U	1	G, U, 2C, pC, 2A		+	26	***C***				+	50	U
						2	2G, U, C, 3A		+	27	***C***				+	42_2	U + *i*
						3	3G, U, 3C, 3A		+	29	***C***						
						4	G_{OH}, A										
T_{15}																	
1	G, C(G, pC)	2				6	3G, U, C, pC		+	22	***C***				+	44	G, U
2	G, C	2				7	2G, U								+	46	G, U
3	G, U					8	U, 5C, A										
4	G, 2U, 5C, A		+	8	***C***	9	G, U, 4C, 2A		+	25	***C***						
5	G_{OH}, U, 4C, 2A		+	9	***C***	10	G_{OH}										
T_5																	
						1	pU					8	8G, 3U, pU, C, A		+	51	***C***
						2	G, U, A					9	U, C				
			+	14	U, C	3	G, U					10	C	3			
						4	4G, U					11	G, U, C, A		+	47_2	***C***S
						5	U, C					12	G_{OH}, G, A				
						6	U, 4C, A										
						7	G_{OH}, 2G, C, A		+	35	***C***						

T_6

						1	pU, A					9	8G, 4U, C, 2A		+	51	*C*
						2	2G, U					10	U, C				
						3	G, U, A					11	C	3			
						4	G, U					12	G_{OH}, U, A		+	47_2	*C*S
			+	14	U	5	4G, U										
						6	U, C										
						7	U, 4C, A										
						8	G_{OH}										

T_9

						1	U (pU)	2				5	G, pU, C		+	38	*C*
			+	2	C, U	2	G, U, 2C		+	20	*C*	6	C	2			
						3	4G, U, 3C					7	2G, U, C		+	40	*C*
						4	G_{OH}, A					8	2G, C				
												9	G_{OH}, 2U, A		+	?	

T_{19}

						1	2G, U, 3C, 2A, pA		+	29	*C*	3	C, pA, 2A				
						2	G_{OH}, 2G, 3C, A		+	30	*C*	4	G, C	2			
												5	C	2			
												6	2G, U, C, A		+	42_2	U
												7	G_{OH}				

Примечание. Обозначения те же, что и в табл. 3. *C*S означает, что продукты полного гидролиза олигомера и молекулы идентифицированы по составу при помощи «схемы».

	T_{15}	T_9		T_5			T_{20}									T_{19}		T_{15}			T_5	T_9				T_{20}			
37	8	2	16	9	14	10	20	21	29	22	25	27	30	35	40	42_1	42_2	44	51	46	47_2	38	47_1, 53	50	52	36	26	
	46	40	53	47_1	47_2	27	2	42_1	42_2	44	9	50	51	52	21	22	30	8	14	25	35	20	37, 26	29	36	16	10	38

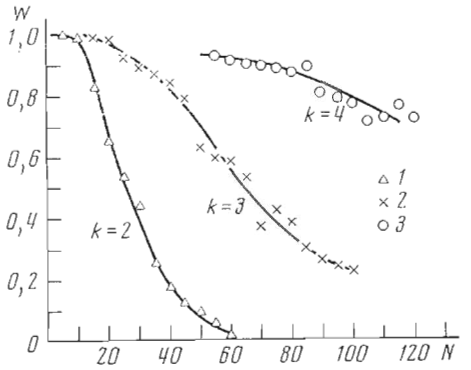

Рис. 2. Доли последовательностей длины N, однозначно восстановленные только по данным двух (*1*), трех (*2*), четырех (*3*) полных гидролизов

проводилась проверка единственности. Для каждой длины N_i подсчитывалось, какая доля W_k (N_i) последовательностей данной длины восстанавливается однозначно по данным k полных гидролизов. При достаточно большом числе фрагментов величины W_k (N_i) достаточно близки к вероятности однозначного восстановления по данным k полных гидролизов случайно выбранной из биохимического материала последовательности длиной N_i.

Результаты проведенного анализа суммированы в рис. 2.

Следует заметить, что при случайном изменении числа проанализированных последовательностей в 1,5 раза положение точек на сглаженной кривой W_2 (N) двух гидролизов и на первой половине приведенных участ-

Таблица 7

Нуклеотидный состав продуктов трех полных гидролизов 5′-концевого фрагмента РНК фага Q_β

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.	№	нукл. состав	кратн.	внешн.
1	pG	1		26	pG, 3G, U, 6C, A	1	+	53	5G, 4U, C, A	1	+
2	G	13		27	U	16		54	C, A	6	
3	G, 3U, 6C, 2A	1	+	28	5G, U, A	1		55	U, C	2	
4	G, 2U, 6C, 4A	1	+	29	U, 5C, 3A	1		56	G, C, A	1	
5	G, C, A	1		30	2G, U, 2C, 2A	1	+	57	G, U, C, 2A	1	+
6	G, 4U, 3C, 2A	1	+	31	U, C, A	1		58	2U, C	2	
7	G, A	4		32	U, 2C, A	1		59	5G, 3U, C, 6A	1	+
8	G, 2U, 3A	1		33	2G, U, A	1		60	G, 2U, C, 2A	1	+
9	G, 2U, C, 3A	1	+	34	U, A	3		61	3U, C, 4A	1	
10	G, 3U, 4C, 4A	1	+	35	3G, U, C, 5A	1	+	62	G, C	2	
11	G, C	2		36	G, U, 2C	1	+	63	3G, 2U, C	1	+
12	G, U	2		37	U, 3A	1		64	G, U, C	1	+
13	G, 2U, C	1	+	38	2G, U, 3C, A	1	+	65	G, 3U, C	1	+
14	G, 3U, C	1	+	39	2G, U	2		66	3G, C, A	1	
15	G, 2C	1		40	U, C	5		67	2G, 4U, C, 6A	1	+
16	G, 2U, 3A	1		41	2G, U, C	1	+	68	G, 7U, C, 3A	1	+
17	G, 5U, C, 5A	1	+	42	4G, U, 3C, 2A	1	+	69	3G, C, 2A	1	
18	G, 4U, C, 2A	1	+	43	U, 2A	2		70	2G, 4U, C	1	+
19	G, 2U, 2C	1	+	44	G, U, A	1		71	G, C, 2A	1	
20	G, 3U, 2C, A	1	+	45	3G, U, 2C, 2A	1	+	72	3U, C	1	
21	G, 5U, 4C, 2A	1	+	46	G, U, 4C, 2A	1	+	73	2U, C, A	1	
22	G, 3U, 3C, 2A	1	+	47	G, U, C, 2A	1	+	74	G, U, C, 2A	1	+
23	G, 2C, 2A	1		48	2G, U, 5C, 3A	1	+	75	G, C, A	1	
24	G, U, 3C, A	1	+	49	G, U, 2C	1	+	76	U, C, A	1	
25	G, G_{OH}, U	1		50	G_{OH}, G	1		77	G_{OH}, 2G, U	1	+
				51	pG, 3G, C, A	1					
				52	C	14					

Обозначения те же, что в табл. 1.

Таблица 8

Нуклеотидная последовательность внешних продуктов полных гидролизов 5′-концевого фрагмента РНК фага Q_β

G-гидролиз				U-гидролиз				C-гидролиз			
№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.	№	нукл. послед.	кратн.	способ опред.
3	ACCCCCCUUUAG		*x*!	26	pGGGGACCCCCCU		*r*(3)	53	UUUAG^{5}UC		*r*(4,13)
4	UCAC(AC)(AC)(CUC)AG		*x*!	30	CAGCAGU		*x*	57	AGUAC		*r*(6,74)
6	UACUUCACUG		*x*!!	35	AAGAGGACAU		*l*(59)	59	UGAGUAUAAGAGGAC		*x*
9	ACAUAUG		*x*	36	GCCU		*r*(10,38)	60	AUAUGC		*l*(9,68)
10	CCUAAAUUACCG		*x*!	38	ACCGCGU		*r*(63,65,41) *l*(10)				
13	UCUG		*x*	41	GCGU		*l*(36,60,64)	63	GUGGUC		*r*(4,13)
14	UUUCG		*x*	42	CGGAGCCGAU		*x*!	64	UGC		*l*(6,13)
17	AAAUUCUUAAUG		*x*!	45	CAGGAGCU		*x*!!	65	CUUUC		*r*(14,20,67)
18	AUUUUCAG		*x*	46	CCAGACCU		*r*(21,30) *l*(20,30)	67	GAUAAUGAAAUUC		*r*(18,68) *l*(42)U
19	CUCUG		*x*	47	CGAAU		*r*(74)	68	UUAAUGAUUUUC		*r*(17,59) *l*(17,59)
20	UUUCCAG		*x*	48	CCGACACGCAU		*x*!!	70	UGGUUUC		*r*(14,20,67) *l*(6,13)
21	ACCU[(U)(U)(C)]UAUCG		*x*!!	49	CCGU		*r*(22,46,24)	74	CAAUC		*r*(22)
22	AAUCUUCCG		*x*!!					77	CUGG$_{OH}$		*r*(38,41,49, 63,65)
24	CAUCCG		*x*								

Примечание. Обозначения те же, что в табл. 5.

К табл. 7 и 8

77	3 53	4 ,18 30,45	6 ,13,19 59,64,70	9 60	10 38	14,21 42,77	17 68	20 46	22,24 48,49	26 3	30 57	35 9	35 10	38,41,49 63,65,77	42 67	45 19	46 21	47 74	48 24	53,63 4,13	57 6	59 35	60,64 36,41	65,70 14,20	67 17	68 18	74 22	26

26-3-53-4-30-57-6-59-35-9-60-36-10-38-63-13-64-41-65-14-42-67-17-68-18-45-19-70-20-46-21-47-74-22-48-24-49-77

Таблица 9

Нуклеотидный состав продуктов полных гидролизов олигомеров, полученных частичными гидролизами фрагмента РНК фага Q_β

G-гидролиз						U-гидролиз						C-гидролиз					
№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ	№ в фр.	нукл. состав	кратн	внешн.	№ в мол.	способ	№ в фр.	нукл. состав	кратн.	внешн.	№ в мол.	способ
							T_{XI}										
1	G, 2U, 2A, pA	1		67	S										+	67	G
2	G, 5U, C, 5A		+	17	C										+	68	G
3	G, 4U, C, 2A		+	18	C												
4	G	3															
5	G, A														+	70	G
6	G, 2U, 2C		+	19	C												
7	G, 3U, 2C, A		+	20	C				+	45	G						
8	G_{OH}, 4U, 4C, 2A		+	21													
									+	46	G						
							T_X										
1	G, 2U, C, 2A, pA		+	9	C	7	U, C, A, pA										
2	G, 3U, 4C, 4A		+	10	C	8	U, A								+	60	G, U
3	G, C					9	C, U, 2C		+	36	C						
4	G, U					10	U, 3A										
5	G					11	U										
6	G_{OH}, 2U, C		+	13	C	12	2G, U, 3C, A		+	38	C				+	63	G, U
						13	2G, U										
						14	U, C										
						15	G_{OH}										

Примечание. Обозначения те же, что и в табл. 6.

T_{XI}			T_{XI}							T_{XI}										T_X			T_X		T_{XI}						
77	3	4	18	6,13	19	9	10	14,21	17	20	22,24	26	30	35	36	38	41,49	42	45	46	47	48	53	63	57	59	60	64	65	70	67	68	74	
	53	30	45	59,64	70	60	38	42,47	68	46	48,49	3	57	9	10	63	65,77	67	19	21	74	24	4	13	6	35	36	41	14	20	17	18	22	26

ков кривых $W_3(N)$ и $W_4(N)$ трех и четырех гидролизов изменяется в пределах 10%. Остальная часть двух последних кривых становится с ростом N значительно менее устойчивой к изменению числа анализируемых последовательностей. Приведенные на рис. 2 сглаженные кривые позволяют ответить на поставленный в начале этого параграфа вопрос с точностью $\sim \pm 5\%$ в указанном выше интервале длин N.

Обсуждение

Несмотря на то что блочный метод с преимущественным использованием пиримидил- и гуанил-РНКазных гидролизов привел к установлению последовательности в длинных полирибонуклеотидах, этот успех был достигнут ценой огромного труда и секвенирование длинных полирибонуклеотидов остается работой, доступной ограниченному кругу лабораторий.

Использование традиционных методов позволяет однозначно восстанавливать по данным только полных гидролизов подавляющую часть олигонуклеотидов длиной 10—12 мономеров. Проведенный нами анализ показывает, что проведение С-гидролиза вместо гидролиза панкреатической РНКазой лишь незначительно увеличивает эту длину. Но использование трех полных гидролизов, специфичных к одному нуклеотиду каждый, позволяет однозначно восстанавливать по данным только полных гидролизов $90 \pm 5\%$ всех последовательностей длиной уже в 30 нуклеотидов. Если провести четвертый специфический полный гидролиз, то по данным только полных гидролизов можно однозначно восстанавливать $90 \pm 5\%$ последовательностей длиной 70 (!) нуклеотидов.

Описанные в работе алгоритмы касаются оптимизации секвенирования при применении трех специфических к одному нукяеотиду способов расщепления. Наш опыт модификации 23S-рибосомных и вирусных РНК смесью метоксиамина и бисульфита показывает, что можно провести количественный полный гидролиз этих длинных молекул панкреатической РНКазой только по остаткам уридина [13, 14]. Большой опыт использования T_1-РНКазы гарантирует также возможность проведения избирательного полного гидролиза по гуанозину. Метод модификации РНК карбодиимидом, позволяющий проводить полный избирательный С-гидролиз модифицированной РНК панкреатической РНКазой [21], также представляется достаточно апробированным.

В настоящее время неизвестен прямой способ проведения полного гидролиза по аденозиновым звеньям для достаточно длинных нуклеотидных последовательностей. Но благодаря комплементарности оснований в двойной спирали данные о продуктах четырех полных гидролизов могут быть получены для двухтяжевой РНК. Для этого достаточно разделить две нити и для каждой из них провести гидролиз по остаткам гуаниловой и уридиловой кислот. Последовательности в продуктах полных гидролизов по гуанозину и уридину одной из нитей легко могут быть перекодированы в последовательности продуктов полных гидролизов по цитидиловой и адениловой кислотам второй нити. Синтез второй нити для однотяжевых РНК может быть проведен, например, методом, описанным в работе [4].

Проведенный нами на примере трех известных последовательностей РНК теоретический анализ возможностей оптимизированного метода показывает, что использование этого метода позволяет резко сократить объем экспериментальной работы, необходимой для однозначного восстановления первичной структуры.

Использование данных нуклеотидного состава и разработанных алгоритмов позволяет ограничиться биохимическим определением последовательности лишь примерно для $^1/_4$ всех продуктов полных гидролизов. Однако обычно для разделения продуктов полных гидролизов применяется двумерная хроматография, позволяющая определить последовательности нуклеотидов длиной до пяти мономеров по их положению на двумерной

Таблица 10

Сравнение числа фрагментов, использованных при установлении структуры традиционными методами, и числа фрагментов, необходимых для однозначного определения последовательности предлагаемым оптимизированным методом

Объект	Фрагмент цистрона из РНК фага R_{17}		5S-РНК *E. coli*		5′-Концевой фрагмент РНК фага Q_β	
Длина	57		120		175	
Метод	Adams et al. (1969) Nature, 223, 1009	Оптимизированный	Brownlee et al. (1968) J. Mol. Biol., 34, 379	Оптимизированный	Billeter et al. (1969) Nature, 224, 1083	Оптимизированный
Продукты полных гидролизов *	22 9 3	34 6 3	40 16 8	55 14 13	47 26 18	77 19 16
Продукты частичных гидролизов	5	1	108	6	14	2

* В графе «Продукты полных гидролизов» число в верхней строке равно количеству различных продуктов полных гидролизов; количество всех секвенируемых продуктов длиной более 3, последовательность нуклеотидов в которых не может быть установлена по данным нуклеотидного состава и концевого анализа, указано во второй строке; в третьей строке указано количество секвенируемых продуктов длиной более 5, последовательность нуклеотидов в которых не может быть установлена по положению на двумерной хроматограмме (фингерпринте).

хроматограмме. С учетом этого для сравнения эффективности двух методов лучше сравнивать количества продуктов полных гидролизов длиной более пяти нуклеотидов, в которых при использовании каждого из методов последовательность нуклеотидов должна определяться биохимически.

В табл. 10 приведены данные, позволяющие сравнивать количество биохимической информации, необходимой для однозначного восстановления структуры трех РНК традиционными методами, использующими два полных гидролиза, и оптимизированным методом, использующим три полных гидролиза. Приведенные данные позволяют проводить сравнение обоих методов при различных способах разделения продуктов гидролизов. Из верхней строки графы «продукты полных гидролизов» табл. 10 видно, что число различных продуктов трех полных гидролизов больше числа различных продуктов двух традиционных гидролизов почти в 1,5 раза. Ниже показано, что, несмотря на это, число олигонуклеотидов, структуру которых надо определять биохимическими методами, при использовании оптимизированного метода остается примерно тем же, что и при использовании традиционного.

После установления нуклеотидной последовательности «внешних» продуктов полных гидролизов разработанные алгоритмы позволяют построить возможную последовательность всего фрагмента (или молекулы). Предложенные критерии единственности позволяют найти участки неоднозначности в построенной гипотетической последовательности.

Если такие участки выявлены, то необходимо провести частичный гидролиз, однако схема анализа позволяет отобрать лишь те немногие олигомеры, которые нужны для устранения неоднозначности. Разработанные алгоритмы, упрощающие обычную интуитивную работу биохимика, позволяют обойтись минимумом экспериментальной работы и при анализе этих олигомеров.

Из табл. 10 видно, что если для установления структуры фрагментов РНК фагов R 17 и Q_β и молекулы 5S-РНК понадобилось 5,14 и 108 про-

дуктов неполных гидролизов (во втором случае авторы имели много дополнительной информации, полученной другими методами), то проведенное нами символическое секвенирование потребовало анализа лишь 1, 2 и 6 олигомеров соответственно, т. е. в 5—18 (!) раз меньше.

Поскольку получение и анализ продуктов частичного гидролиза является одной из наиболее сложных и трудоемких биохимических операций, то столь значительное уменьшение их числа свидетельствует о большой эффективности и целесообразности разработанного метода. Поэтому можно надеяться, что использование оптимизированного метода может резко ускорить важные работы по определению первичной структуры нуклеиновых кислот.

В заключение приводим краткую схему оптимизированной процедуры секвенирования полирибонуклеотидов блочным методом.

1. В зависимости от длины полинуклеотидной цепи проводят два, три или четыре полных гидролиза.

2. Анализируют нуклеотидный состав полученных продуктов и выделяют «внешние». На этой и последующих стадиях имеет существенное значение точность определения нуклеотидного состава.

3. Определяют экспериментально любым из методов нуклеотидные последовательности во «внешних» продуктах одного из гидролизов.

4. По нуклеотидным составам остальных «внешних» продуктов и участков перекрывания продуктов с уже определенной последовательностью находят последовательность в этих остальных продуктах.

5. Все «внешние» олигонуклеотиды упорядочивают в схему, по которой строят нуклеотидную последовательность, возможную при данной информации, и указывают, какие продукты необходимы для достижения однозначности по алгоритмам построения последовательности и проверки единственности.

6. Получают продукт частичного гидролиза с требуемыми характеристиками, проводят необходимые полные гидролизы его (§ 2, II).

Анализируют нуклеотидный состав полученных продуктов и выделяют «внешние» продукты полных гидролизов фрагмента.

7. Проводят идентификацию «внешних» продуктов фрагмента с «внешними» олигонуклеотидами всей молекулы по их нуклеотидным составам и схеме молекулы (§ 3, II).

8. Определяют, вносит ли фрагмент дополнительную информацию, и проводят учет такой информации модификацией схемы.

9. По схеме опять определяют характеристики дополнительной информации, если однозначность не достигнута, и т. д., пока однозначность не будет достигнута.

Алгоритмические этапы процесса просты в употреблении и могут быть использованы как вручную, так и на компьютере, если обращение к таковому удобно в работе конкретной лаборатории. Алгоритмический анализ всей биохимической информации о последовательности длиной $\sim$100 нуклеотидов вручную требует нескольких часов работы одного человека.

Если на каком-либо этапе работы описанные алгоритмы не срабатывают, то это может служить указанием на погрешности в исходных биохимических данных.

ЛИТЕРАТУРА

1. Holly R. W., Apgar J., Everett G. A., Madison J. T., Marguisse M., Merrill S. H., Penswick J. R., Zamir A. (1965) Science, **147**, 1462—1465.
2. Billeter M. A., Dahlberg J. E., Goodman H. M., Hidley J., Weissman C. (1969) Nature, **224**, 1083—1086.
3. Min Jou W., Haegeman G., Fiers W. (1972) Nature, **237**, 82—90.
4. Mills D. R., Cramer F. R., Spiegelman S. (1973) Science, **180**, 916—927.
5. Rubin G. M. (1974) Eur. J. Biochem., **41**, 197—201.
6. Ehressman C., Stiegler P., Mackie R. A., Zimmerman R. A., Ebel J. P., Fellner P. (1975) Nucleic Acids Res., **2**, 265—278.

7. Mandeles S. (1972) Nucleic Acids Sequence Analysis, Colambia Univ. Press, N. Y.—London (русский перевод: Манделес С. (1975) Установление первичной структуры нуклеиновых кислот, «Мир», М.).
8. Barrell B. G., Clark B. F. C. (1974) Handbook of Nucleic Acid Sequences Joynson-Bruwers Ltd, England.
9. Sanger F., Coulson A. R. (1975) J. Mol. Biol., 94, 441—448.
10. Uziel M. (1975) Nucleic Acid. Res., 2, 469—173.
11. Lee J. C., Ho N. W. Y., Gilham P. T. (1965) Biochim. et biophys. acta, 95, 503—504.
12. Budowski E. I., Sverdlov E. D., Monastyrskaya G. S. (1972) FEBS Lett., 25, 201—204.
13. Mazo A. M., Kisselev L. L. (1975) FEBS Lett., 59, 177—179.
14. Mazo A. M., Scheinker V. S., Kisselev L. L. (1975) Mol. Biol. Rep., 2, 233—239.
15. Sen G. C., Ghosh H. P. (1974) Anal. Biochem., 58, 578—591.
16. Юдман Б. Х. (1973) Тезисы докладов конференции «Анализ и синтез конечных автоматов», с. 48—52, Изд. Саратовского ун-та.
17. Юдман Б. Х., Рашин А. А. (1974) Биофизика, 19, 405—409.
18. Юдман Б. Х., Рашин А. А. (1974) Биофизика, 19, 613—615.
19. Юдман Б. Х. (1974) Тезисы докладов IV совещания по проблемам автоматизации анализа изображений микроструктур, с. 49, Пущино.
20. Юдман Б. Х. (1974) Тезисы докладов IV совещания по проблемам автоматизации анализа изображений микроструктур, с. 49—50, Пущино.
21. Gilham P. T. (1970) Ann. Rev. Biochem., 39, 227—250.
22. Uchida T., Arima T. (1970) J. Biochem., 67, 91—95.
23. Mosimann J. E., Shapiro M. B., Merrill C. R., Bradley D. F., Vinton J. E. (1966) Bull. Math. Biophys., 28, 235—260.
24. Adams J. M., Jeppesen P. J. N., Sanger F., Barrell B. G. (1969) Nature, 223, 1009—1014.
25. Юдман Б. Х., Рашин А. А. (1974) Биофизика, 19, 805—809.
26. Brownlee G. G., Sanger F., Barrell B. G. (1968) J. Mol. Biol., 34, 379—412.

Поступила в редакцию
9.VII.1976

OPTIMIZATION OF THE BLOCK METHOD OF NUCLEIC ACID SEQUENCE DETERMINATION

RASHIN A. A., YUDMAN B. Kh., KISSELEV L. L., MAZO A. M.

Institute of Biological Physics, Academy of Sciences of the USSR, Pushchino; Institute of Molecular Biology, Academy of Sciences of the USSR, Moscow

The possibilities of optimization of the block method of RNA sequence determination are considered. Rigorous criteria to assess the uniqueness of a reconstructed sequence are found. Unambiguous determination of a nucleotide sequence if it relies upon complete hydrolyses (each specific for one nucleotide) is shown to be possible for about 90% sequences of nucleotides having the following chain-length: 15 nucleotides with two, 30-with three and 70 with four complete hydrolyses. The detailed algorithms for processing biochemical information are proposed for the cases of three or four complete hydrolyses thus significantly reducing the experimental work needed for the sequence determination. The order of biochemical operations for sequence determination by the optimized block metod is formulated based on the use of RNA hydrolyses at G (T_1 RNase), C (pancreatic RNase affecting carbodiimide-modified RNA) and U (pancreatic RNase acting on methoxiamine-bisulphite modified RNA). The utility of the proposed method is exemplified with some known sequences.