УДК 577.212+519.756.2+593.174.1

С. А. Залетова

МАТЕМАТИЧЕСКИЕ МОДЕЛИ СБОРКИ ГЕНОВ У РЕСНИЧНЫХ

Введение

Мысль о том, что живая клетка, выполняя свои функции, перерабатывает не только химические вещества, но и информацию, возникла в третьей четверти XX в., когда была раскрыта роль ДНК как носителя генетического кода. Яркой иллюстрацией здесь может служить чрезвычайно интересная с вычислительной точки зрения процедура сборки генов при половом размножении одноклеточных класса ресничных (к нему принадлежит знакомая по школьным урокам зоологии инфузория-туфелька, см. рис. 1 и 2). У ресничных функционируют два ядра: макроядро и микроядро, наследственная информация в которых организована по-разному. Микроядро является покоящимся ядром, оно служит для передачи наследственного материала от поколения к поколению. Хромосомы микроядра находятся в компактном состоянии. В макроядре же хромосомы находятся в активном состоянии – они поставляют информацию для процессов жизнедеятельности организма. На начальных этапах развития инфузории хромосомный материал испытывает серию сложных превращений. Происходит своеобразная «разархивация» генов, биологи обычно говорят о *сборке* генов.

Обсудим процесс сборки гена немного подробнее. ДНК хромосом микроядра имеет большие размеры. Гены в них собраны в группы с длинными промежутками между ними, заполненными разнообразными уникальными и повторяющимися последовательностями ДНК. Важная особенность генов микроядра заключается в том, что все они прерваны особыми элементами, так называемыми *IES* (internal eliminated sequences). Кодирующие участки генов, разделенные *IES*, называются *MDS* (macronuclear destined sequences). Отметим, что мы используем не самые распространенные обозначения *MDS* и *IES* вслед за авторами рассматриваемых в данной работе моделей [1]. Каждая *IES* уникальна, т. е. все они разные.

В процессе сборки фрагменты, соответствующие *IES*, удаляются. Но что совсем удивительно – происходит перестановка участков, соответствующих

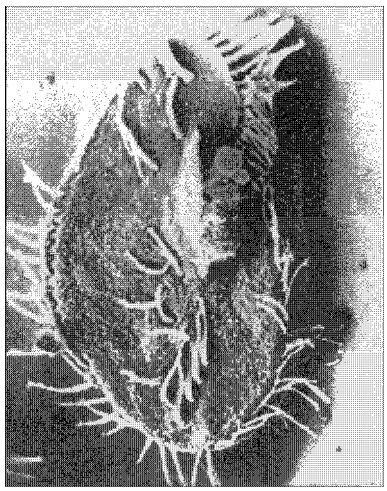

Рис. 1. Инфузория-туфелька

MDS. Например, в микроядерной ДНК девять кодирующих участков гена актина I у инфузории Oxytricha nova расположены в следующей последовательности: $3, 4, 6, 5, 7, 9, 2, 1, 8$. После преобразования геномной ДНК эти участки занимают положение в ряду с 1-го по 9-й, и только такая их последовательность дает функциональный ген. Замечательно, что хранение генетической информации в микроядре инфузории и размещение файлов на жестком диске современного компьютера организовано по одному и тому же принципу. При этом сложный процесс извлечения фрагментов гена из ДНК микроядра с последующим соединением их в нужном порядке происходит в несколько раз быстрее, точнее и энергетически выгоднее, чем гораздо более примитивные операции в опытах по молекулярным вычислениям in vitro (подробное описание таких опытов см. в [2, 3]).

Неясно, почему информация, кодирующая белок, хранится в микроядре в одном порядке, а функционирует в макроядре – в другом. Непонятно также, по какой программе происходит правильное соединение *MDS* в макроядре и с помощью каких молекулярных механизмов осуществляется эта программа. Различными авторами предложены разные математические модели процесса сборки генов у ресничных. Пока неизвестно, реализуется ли какая-либо из них в живых организмах, а если реализуется, то какая именно. В настоящей работе мы анализируем и сравниваем две такие модели, предлагавшиеся в [1]. Целью нашего анализа было, в частности, нахождение таких особенностей каждой из моделей, проявления которых в живой клетке могли бы

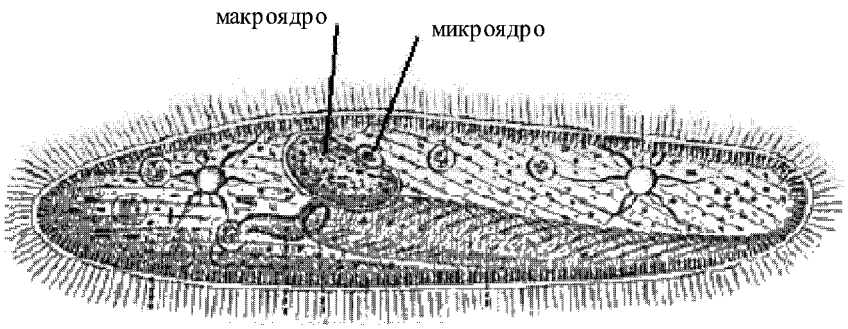

Рис. 2. Инфузория-туфелька. Схематическое изображение

быть обнаружены экспериментально. Мы показываем, что исследуемые модели отличаются тем, что при сборке гена по одной из них все промежуточные молекулы имеют одну и ту же длину, в то время как процесс, реализующий другую модель, необходимо проходит через фазы, в которых промежуточные молекулы существенно различаются по длине. Так как различия в длинах молекул хорошо обнаруживаются методами современной биохимии, наше исследование может служить теоретической основой для экспериментальной проверки того, какая из рассматриваемых моделей процесса сборки генов у ресничных реализуется в природе.

В первом разделе речь пойдет о биологической стороне процесса сборки генов, и его чтение не является обязательным для понимания математических объектов, рассматриваемых в следующих разделах. Во втором разделе мы рассмотрим структуру микроядерных и макроядерных генов. В следующих двух разделах описываются две существующие модели процесса сборки, внутримолекулярная и межмолекулярная. В четвертом разделе мы проанализируем и сравним между собой молекулярные операции, на которых основываются предложенные модели.

1. Биологическая сторона процесса

В данном разделе нам придется оперировать некоторыми биологическими терминами. Для более подробного ознакомления с используемыми понятиями можно рекомендовать [4].

При бесполом размножении инфузории диплоидное микроядро делится путем митоза, а макроядро – с помощью прямого деления. Но это не может продолжаться бесконечно: постепенно макроядро стареет, что приводит к ослаблению метаболической активности клетки. Необходимо обновление

макроядра. Поэтому время от времени у ресничных происходит конъюгация – своеобразная форма полового процесса. Две клетки сцепляются друг с другом и плавают вместе 10–12 часов, а затем расходятся. За время конъюгации их макроядра начинают разрушаться, а микроядра делятся путем мейоза на четыре гаплоидных ядра каждое. Ход последующих событий варьирует в деталях у разных видов ресничных, но принципиальная схема общая: клетки партнеров обмениваются гаплоидными ядрами, по одному от каждой клетки, затем каждое сливается с местным (стационарным) гаплоидным ядром, т. е. происходит оплодотворение. К этому моменту все лишние ядра дегенерируют, и в каждой клетке остается по одному диплоидному ядру – продукту оплодотворения. После расхождения партнеров ядро делится митотически на два. Одно из дочерних ядер останется микроядром, другое превратится в макроядро. Развитие макроядра занимает несколько дней и сопровождается полной реорганизацией генома микроядра-предшественника. Именно в этот период происходит сборка генов. Схематически процесс полового размножения изображен на рис. 3.

2. Структура генов

Микроядерный ген устроен как последовательность *MDS*, разделенных *IES*. Каждая *MDS*, кроме первой и последней, состоит из трех последовательностей нуклеотидов: «приходящего» указателя, тела и «уходящего» указателя. *MDS* имеет следующую структуру: $M_i = (\pi_i, \mu_i, \pi_{i+1})$, где π_i – «приходящий» указатель *MDS* M_i, μ_i – его тело, а π_{i+1} – «уходящий» указатель. Уходящий указатель M_i совпадает с приходящим указателем M_{i+1} (одинаковые последовательности нуклеотидов). $M_1 = (b, \mu_1, \pi_2)$ – первая *MDS*, $M_k = (\pi_k, \mu_k, e)$ – последняя *MDS*, где b, e – «непарные указатели», которые мы будем называть маркерами. У ресничных *MDS* находятся в перемешанном порядке: по-видимому, случайные перестановки с некоторыми инвертированными *MDS*.

В макроядерном гене все *MDS* сцеплены вместе в правильном порядке: каждая M_i склеена с M_{i+1} по перекрывающемуся указателю π_{i+1}. Таким образом, «программа» сборки гена должна:

- удалить *IES*;
- выстроить в правильном порядке *MDS*;
- склеить *MDS*;
- вырезать ген из хромосомы;
- размножить молекулу ДНК до нужного числа копий (в зависимости от вида).

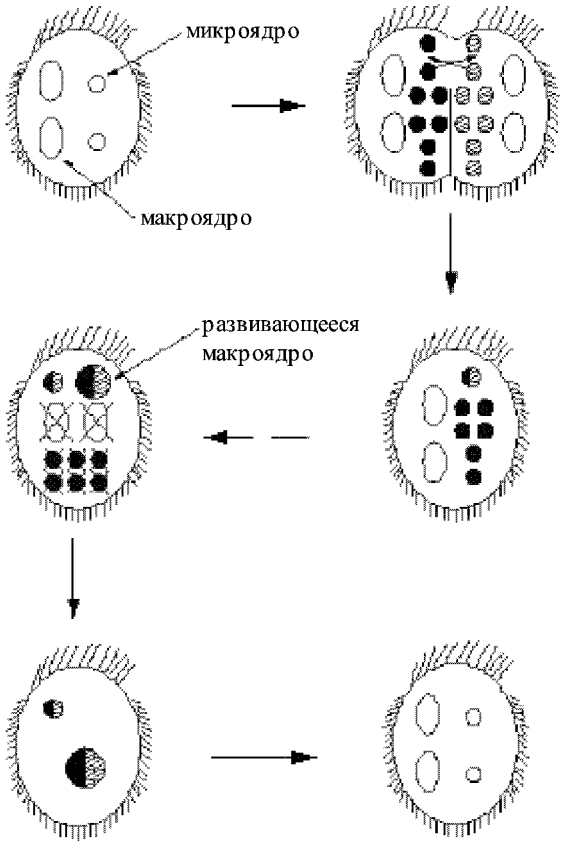

Рис. 3. Процесс полового размножения

Сложность проблемы заключается в том, что часто для одного-единственного гена должно быть упорядочено, сцеплено и затем вырезано с предельной точностью более 50 *MDS*.

3. Внутримолекулярная модель

В рамках этой модели есть три молекулярные операции, которые применяются к одной молекуле. Молекула ДНК складывается и происходит перекомбинация, изменяющая последовательность нуклеотидов. Только одна операция вырезает кольцевую молекулу из оригинальной последовательности, в остальных случаях ни один нуклеотид не теряется из ДНК.

Опишем молекулярные операции.

ld-вырезание (от англ. **l**oop – петля, **d**irect repeat – прямой повтор). Эта операция применяется к последовательностям нуклеотидов, которые имеют повторяющиеся в прямом порядке образцы указателя p. Молекула складывается в виде петли так, чтобы два повтора p расположились на одной линии. В результате одна кольцевая и одна линейная молекула вырезаются из оригинальной молекулы ДНК (рис. 4).

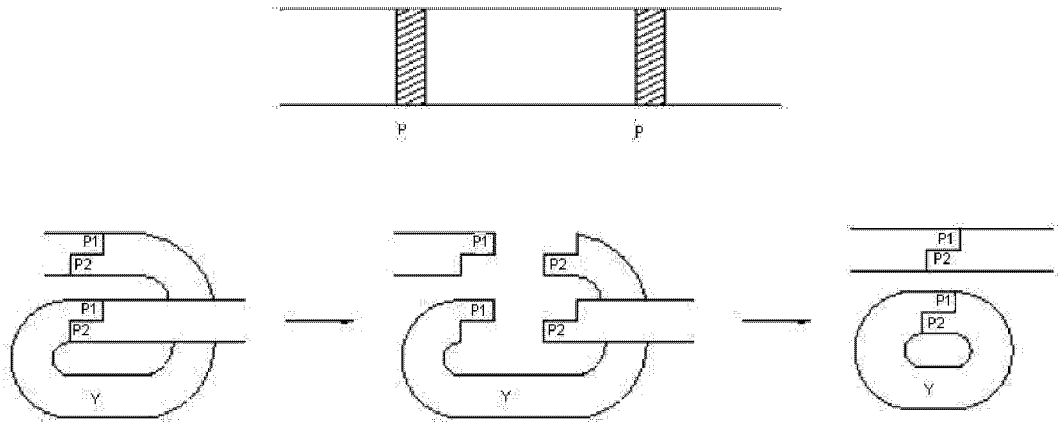

Рис. 4. Схематическое изображение операции **ld**

hi-вырезание/перестановка (от англ. **h**airpin – шпилька, **i**nverted repeat – обратный повтор). Эта операция применяется к последовательностям ДНК, которые имеют повторяющиеся в обратном порядке образцы указателя p. Молекула складывается в виде шпильки так, чтобы два повтора p расположились на одной линии. В результате инвертируется участок молекулы, находящийся между двумя повторами указателя. Последовательность нуклеотидов меняется, но ни одна молекула не вырезается из оригинальной ДНК (рис. 5).

dlad-вырезание/перестановка (от англ. **d**ouble **l**oop – двойная петля, **a**lternating **d**irect repeat – чередующийся прямой повтор). Эта операция применяется к последовательностям ДНК, которые имеют чередующиеся прямые повторы указателей p и q. Молекула складывается в виде двойной петли так, чтобы два повтора p и два повтора q расположились на одной линии. В результате участки молекулы, находящиеся между указателями p и q, меняются местами (рис. 6). На рисунках 4–6 показано, что указатели p и q

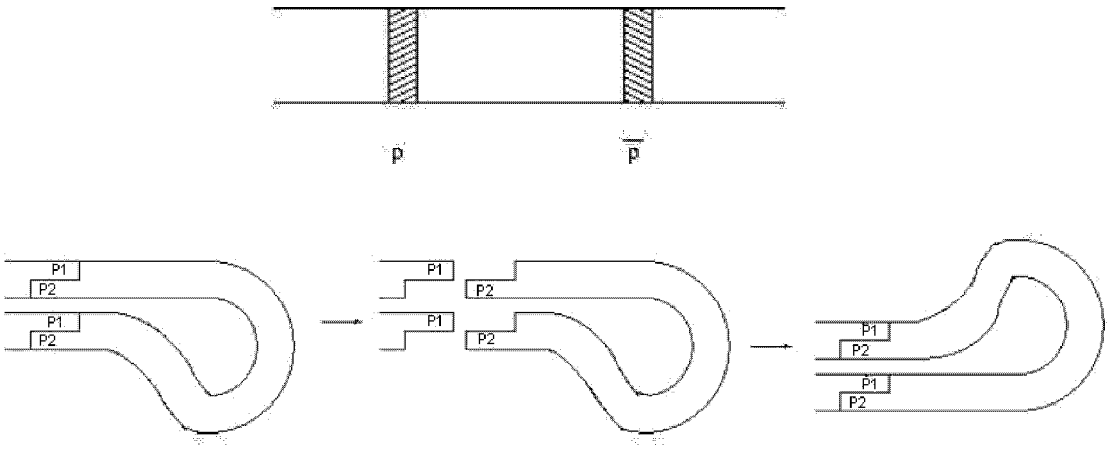

Рис. 5. Схематическое изображение операции **hi**

разрезаются на части p_1, p_2 и q_1, q_2 соответственно. Мы знаем только, что эти части составляют указатель, но не знаем, каким образом.

Указатели перестают выступать в качестве указателей после завершения операции. Последовательность нуклеотидов, которая формирует указатель, может встречаться в молекуле ДНК много раз, но она является указателем только тогда, когда расположена на границе *MDS* и *IES*.

Будем называть удачным набор, в котором присутствуют все *MDS*. Между двумя повторами указателя либо не содержится ни одной *MDS*, либо содержатся все. Таким образом, после применения к удачному набору **ld**-операции в первом случае кольцевая молекула не содержит *MDS*, а в последнем – содержит все *MDS* и ген будет собран на кольцевой молекуле.

В макроядре все молекулы линейные, даже если ген собран на кольцевой молекуле, то при помощи энзимов и теломер она будет разрезана и преобразована в линейную молекулу.

3.1. Процесс сборки гена

Сборка гена – это в основном процесс упорядочения и собирания *MDS*.

Изначально есть последовательность из k *MDS*. С каждой операцией число *MDS* уменьшается на 1 (**ld**, **hi**) или на 2 (**dlad**). В конце остается только одна большая *MDS*, состоящая из всех k оригинальных *MDS*, сцепленных вместе в правильном порядке.

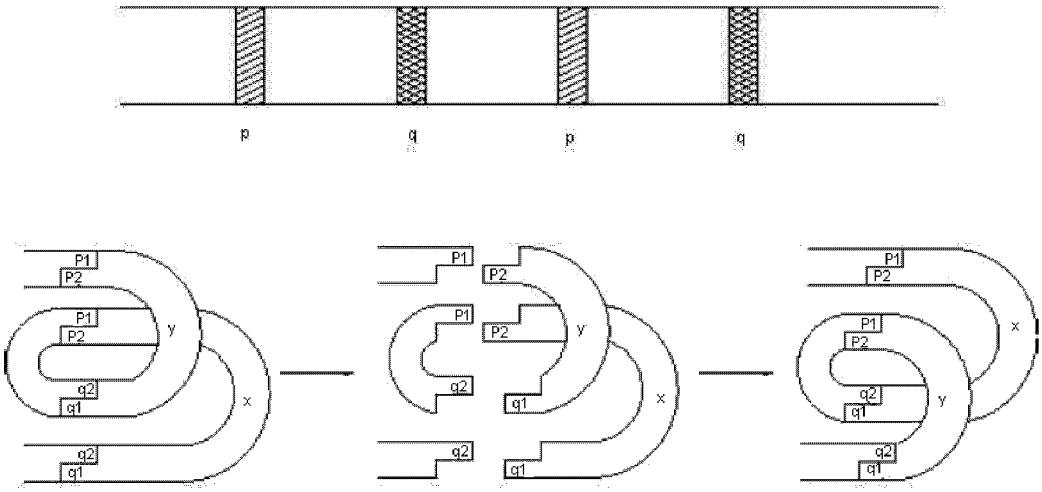

Рис. 6. Схсматичсскос изображснис опсрации **dlad**

В рамках рассматриваемых нами моделей сборки генов полная структурная информация о микроядерном гене и промежуточных молекулах заложена только в последовательности *MDS*, поэтому для упрощения «забудем» о *IES* и будем работать только с *MDS*. Информация о конечном гене и его предшественниках, таким образом, будет сохранена.

3.2. Формализация процесса

Каждая *MDS* – это последовательность ДНК. Структура *MDS*:

$$M_1 = (b, \mu_1, \pi_2), \ldots, M_i = (\pi_i, \mu_i, \pi_{i_1}), \ldots, M_k = (\pi_k, \mu_k, e),$$

где π_{i+1} – уходящий указатель M_i – идентичен приходящему указателю M_{i+1}.

Для сборки гена важна только пара приходящего и уходящего указателей каждой *MDS*, поэтому можно представить каждую *MDS* как пару ее указателей (π_i, π_{i+1}).

Для нас знать последовательность ДНК каждого указателя не так важно, как его позицию в гене. Поэтому для простоты обозначим каждый указатель с помощью целых чисел: $M_1 = (b, 2), \ldots, M_i = (i, i+1), \ldots, M_k = (k, e)$. Составные *MDS* формируются из сцепленных в правильном порядке нескольких элементарных (микроядерных) *MDS*. Обозначим через $M_{i,j}$, $i < j$, составную *MDS*, сформированную из сцепленных M_i, $M_{i+1}, \ldots, M_j$. И для простоты введем также обозначения $M_{i,j} = (i, j+1)$, $M_{i,i} = M_i$. Далее вве-

дем понятие *MDS*-дескрипторов, переведем молекулярные операции на язык *MDS*-дескрипторов и, опираясь на эту формальную систему, докажем общий результат:

Теорема 1. *Любой микроядерный ген может быть собран с помощью последовательности операций* **ld**, **hi** *и* **dlad**.

Еще раз отметим, что мы рассматриваем уже существующую модель, предложенную в [1], и используем для ее описания систему обозначений, с помощью которых она представлена авторами.

3.3. Последовательности MDS

Шаг 1. Опустим *IES*, обозначим *MDS* через $M_{i,j}$. (Мы работаем с геном, состоящим из k *MDS* в их микроядерной версии.)

Мы используем алфавит $\Theta_k = \{M_{i,j} \mid 1 \leqslant i \leqslant j \leqslant k\}$. Последовательности *MDS* – строки над алфавитом Θ_k, некоторые *MDS* могут быть инвертированы: $\overline{M}_{i,j}$.

Последовательность *MDS* назовем *ортодоксальной*, если она представлена в форме M_{1,i_2-1}, $M_{i_2,i_3-1}, ..., M_{i_n,k}$, т. е. все *MDS* в правильном порядке, нет инверсий, некоторые из *MDS* уже собраны.

Последовательность *MDS* назовем *реальной*, если она является направленной перестановкой ортодоксальной последовательности *MDS*.

Шаг 2. Закрепим за каждым *MDS* пару его указателей/маркеров; обозначим каждый указатель целым числом.

Алфавиты:

$$\Delta_k = \{2, 3, ..., k\}, \qquad \overline{\Delta}_k = \{\overline{2}, \overline{3}, ..., \overline{k}\},$$

где $\Pi_k = \Delta_k \cup \overline{\Delta}_k$ – множество указателей, $\bar{\bar{i}} = i$, $\Psi = \{b, e, \overline{b}, \overline{e}\}$ – множество маркеров.

Мы работаем сейчас с последовательностями (строками) над следующим алфавитом:

$$\Gamma_k = \left\{(b,e), (\overline{e},\overline{b})\right\} \cup \{(i,j), (\overline{j},\overline{i}) \mid 2 \leqslant i \leqslant j \leqslant k\} \cup \\ \cup \left\{(b,j), (\overline{j},\overline{b}) \mid 2 \leqslant j \leqslant k\right\} \cup \{(i,e), (\overline{e},\overline{i}) \mid 2 \leqslant i \leqslant k\}.$$

Каждую строку над алфавитом Γ_k будем называть *MDS*-дескриптором. *MDS*-дескриптор назовем *реальным*, если он описывает реальную последовательность *MDS*.

Факт. *Для реального MDS-дескриптора δ и указателя p, δ имеет либо* 0, *либо* 2 *вхождения из множества* $\{p, \overline{p}\}$.

Если p и $\overline{p}$ встречаются в δ, тогда p назовем *положительным* указателем, в противном случае p назовем *отрицательным* указателем. В дальнейшем будем говорить о вхождении указателя p, имея в виду вхождение символа из множества $\{p, \overline{p}\}$.

Переведем молекулярные операции на язык реальных *MDS*-дескрипторов. Рассмотрим для каждой из операций ситуации, когда задействован входящий или уходящий указатель.

Существует два случая применения операции **ld**:

Случай 1. Простая **ld** (только одна *IES* разделяет два вхождения указателя p):

$$\mathbf{ld}_p(\delta_1(q,p)(p,r)\delta_2) = \delta_1(q,r)\delta_2.$$

Случай 2. Граничная **ld** (ген ограничен двумя вхождениями p):

$$\mathbf{ld}_p((p,q)\delta(r,p)) = (r,q)\delta.$$

Существует два случая применения операции **hi**:

Случай 1. Первое вхождение p – входящий указатель (тогда и второе тоже):

$$\mathbf{hi}_p(\delta_1(p,q)\delta_2(\overline{p},\overline{r})\delta_3) = \delta_1\overline{\delta}_2(\overline{q},\overline{r})\delta_3.$$

Случай 2. Первое вхождение p – уходящий указатель (тогда и второе тоже):

$$\mathbf{hi}_p(\delta_1(q,p)\delta_2(\overline{r},\overline{p})\delta_3) = \delta_1(q,r)\overline{\delta}_2\delta_3.$$

Существует четыре основных случая применения операции **dlad**: первое вхождение p – входящий/уходящий указатель, первое вхождение q – входящий/уходящий. Результат: две последовательности нуклеотидов, ограниченные p и q, поменяются местами.

1. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(q,r_2)\delta_3(r_3,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_2)\delta_3(r_3,r_1)\delta_2\delta_5.$
2. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(r_2,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_1)\delta_2(r_2,r_4)\delta_5.$
3. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,r_3)\delta_4(r_4,q)\delta_5) = \delta_1(r_1,r_3)\delta_4(r_4,r_2)\delta_3\delta_2\delta_5.$
4. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(r_2,q)\delta_3(p,r_3)\delta_4(q,r_4)\delta_5) = \delta_1(r_1,r_3)\delta_4\delta_3\delta_2(r_2,r_4)\delta_5.$
5. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(q,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_1)\delta_2\delta_5.$
6. $\mathbf{dlad}_{p,q}(\delta_1(p,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_4)\delta_5.$
7. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,q)\delta_5) = \delta_1(r_1,r_2)\delta_3\delta_2\delta_5.$

Отметим, что при использовании граничного варианта операции **ld** «содержательная» часть ДНК оказывается в циклической молекуле. Очевидно, что нет существенной разницы, применяются рассмотренные операции к линейным фрагментам ДНК или к циклическим – все операции будут работать и на циклических фрагментах. Поэтому не будем отдельно описывать применение операций к циклическим молекулам.

Далее, от молекулярных операций мы перейдем к формальным вычислениям на парах букв. Процесс сборки гена сейчас – формальный перезаписывающий процесс, основанный на трех правилах. На входе – реальный *MDS*-дескриптор, на выходе – последовательность правил, которая превращает его в (b, e) или $(\overline{e}, \overline{b})$; любую такую последовательность будем называть *успешной редукцией*. Теперь можно доказать теорему 1 ([1]).

Доказательство. Для любого *MDS*-дескриптора существует операция, применимая к нему (таким образом, сокращающая его длину). Если δ имеет положительный указатель, то применяем операцию **hi** относительно этого указателя. В противном случае все указатели отрицательные. Если δ имеет чередующиеся, повторяющиеся в прямом порядке участки $(\dots p \dots q \dots p \dots q \dots)$, тогда применяем **dlad** относительно p и q. В противном случае рассмотрим (отрицательный) указатель p в δ такой, что расстояние (число указателей) между двумя вхождениями p минимально. Тогда это расстояние должно быть равно нулю и к δ можно применить **ld**.

4. Межмолекулярная модель

Рассмотрим теперь межмолекулярную модель. Операции этой модели привлекают взаимодействия между двумя различными молекулами ДНК. Основной биологический механизм тот же самый: выравнивание и сшивание по указателям. До фактического начала процесса сборки гена микроядерный геном умножается в несколько раз (число копий зависит от вида и обычно это степень 2). Процесс сборки гена происходит параллельно во всех копиях генома. Внутримолекулярная модель предполагает, что идентичные копии генома собираются параллельно, без взаимодействия друг с другом. Межмолекулярная модель предполагает, что эти молекулы ДНК взаимодействуют друг с другом, чтобы получилось в конце то же число собранных генов.

Модель состоит из трех операций:

- внутримолекулярная рекомбинация;
- кольцевая внутримолекулярная рекомбинация;
- межмолекулярная рекомбинация.

Отметим, что описание операций межмолекулярной модели мы даем не в обозначениях, предлагаемых автором [1], а в обозначениях, использованных для описания операций внутримолекулярной модели, т. е. на языке MDS-дескрипторов.

Внутримолекулярная рекомбинация

$$\delta_1(p, r_1)\delta_2(r_2, p)\delta_3 \to \delta_1\delta_3 + [(r_2, r_1)\delta_2],$$

$$\delta_1(r_1, p)\delta_2(p, r_2)\delta_3 \to \delta_1(r_1, r_2)\delta_3 + [\delta_2],$$

когда первое вхождение p – входящий и уходящий указатель соответственно. Эта операция обратима, т. е. мы также имеем следующую (межмолекулярную!) операцию:

$$\delta_1(p, r_1)\delta_3 + [\delta_2(r_2, p)] \to \delta_1\delta_2(r_2, r_1)\delta_3,$$

$$\delta_1(r_1, p)\delta_3 + [\delta_2(p, r_2)] \to \delta_1(r_1, r_2)\delta_2\delta_3.$$

Заметим, что это обобщение операции **ld**: кольцевая молекула ДНК вырезается при таком же типе образца, сворачивания и сращивания и не имеет ограничений на то, что вырезанная молекула не должна содержать MDS. Можно сказать, что это обратимая «не простая» операция **ld**.

Кольцевая внутримолекулярная рекомбинация

$$[\delta_1(p, r_1)\delta_2(r_2, p)\delta_3] \to [\delta_1\delta_3] + [(r_2, r_1)\delta_2],$$

$$[\delta_1(r_1, p)\delta_2(p, r_2)\delta_3] \to [\delta_1(r_1, r_2)\delta_3] + [\delta_2],$$

эта операция также обратима, и ее инверсия есть

$$[\delta_1(p, r_1)\delta_3] + [\delta_2(r_2, p)] \to [\delta_1\delta_2(r_2, r_1)\delta_3],$$

$$[\delta_1(r_1, p)\delta_3] + [\delta_2(p, r_2)] \to [\delta_1(r_1, r_2)\delta_2\delta_3].$$

Как и предыдущую, эту операцию можно рассматривать как обратимую «не простую» **ld**, примененную к кольцевой молекуле.

Межмолекулярная рекомбинация

$$\delta_1(p, r_1)\delta_2 + \delta_3(r_2, p)\delta_4 \to \delta_1\delta_4 + \delta_3(r_2, r_1)\delta_2,$$

эта операция также обратима.

5. Анализ операций и сравнение процессов сборки

Попробуем сопоставить операции двух моделей и сравнить процесс сборки гена во внутри- и межмолекулярной моделях.

Как уже говорилось, внутримолекулярная рекомбинация (прямая операция) является обобщением операции **ld** без ограничения на то, что кольцевая молекула либо не должна содержать MDS, либо должна содержать все MDS. Если между повторяющимися в прямом порядке указателями содержатся все MDS либо не содержится ни одной MDS, то применение этих двух операций эквивалентно. Однако если между повторяющимися указателями содержится часть MDS (операция **ld** в этом случае неприменима) и, следовательно, кольцевая молекула будет содержать часть MDS после применения прямой внутримолекулярной рекомбинации, то вслед за ней должна быть применена обратная операция, иначе часть кодирующей последовательности будет утеряна. Рассмотрим эту ситуацию подробнее.

Случай 1. Есть повторяющиеся в прямом порядке чередующиеся указатели p и q.

1. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(q,r_2)\delta_3(r_3,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_2)\delta_3(r_3,r_1)\delta_2\delta_5.$
2. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(r_2,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_1)\delta_2(r_2,r_4)\delta_5.$
3. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,r_3)\delta_4(r_4,q)\delta_5) = \delta_1(r_1,r_3)\delta_4(r_4,r_2)\delta_3\delta_2\delta_5.$
4. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(r_2,q)\delta_3(p,r_3)\delta_4(q,r_4)\delta_5) = \delta_1(r_1,r_3)\delta_4\delta_3\delta_2(r_2,r_4)\delta_5.$
5. $\mathbf{dlad}_{p,q}(\delta_1(p,r_1)\delta_2(q,p)\delta_4(r_4,q)\delta_5) = \delta_1\delta_4(r_4,r_1)\delta_2\delta_5.$
6. $\mathbf{dlad}_{p,q}(\delta_1(p,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5) = \delta_1\delta_4\delta_3(r_3,r_4)\delta_5.$
7. $\mathbf{dlad}_{p,q}(\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,q)\delta_5) = \delta_1(r_1,r_2)\delta_3\delta_2\delta_5.$

1. $\delta_1(p,r_1)\delta_2(q,r_2)\delta_3(r_3,p)\delta_4(r_4,q)\delta_5 \to \delta_1\delta_4(r_4,q)\delta_5 + [\delta_2(q,r_2)\delta_3(r_3,r_1)] \to$
$\to \delta_1\delta_4(r_4,r_2)\delta_3(r_3,r_1)\delta_2\delta_5.$
2. $\delta_1(p,r_1)\delta_2(r_2,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5 \to \delta_1\delta_4(q,r_4)\delta_5 + [\delta_2(r_2,q)\delta_3(r_3,r_1)] \to$
$\to \delta_1\delta_4\delta_3(r_3,r_1)\delta_2(r_2,r_4)\delta_5.$
3. $\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,r_3)\delta_4(r_4,q)\delta_5 \to \delta_1(r_1,r_3)\delta_4(r_4,q)\delta_5 + [\delta_2(q,r_2)\delta_3] \to$
$\to \delta_1(r_1,r_3)\delta_4(r_4,r_2)\delta_3\delta_2\delta_5.$
4. $\delta_1(r_1,p)\delta_2(r_2,q)\delta_3(p,r_3)\delta_4(q,r_4)\delta_5 \to \delta_1(r_1,r_3)\delta_4(q,r_4)\delta_5 + [\delta_2(r_2,q)\delta_3] \to$
$\to \delta_1(r_1,r_3)\delta_4\delta_3\delta_2(r_2,r_4)\delta_5.$
5. $\delta_1(p,r_1)\delta_2(q,p)\delta_4(r_4,q)\delta_5 \to \delta_1\delta_4(r_4,q)\delta_5 + [\delta_2(q,r_1)] \to \delta_1\delta_4(r_4,r_1)\delta_2\delta_5.$

6. $\delta_1(p,q)\delta_3(r_3,p)\delta_4(q,r_4)\delta_5 \to \delta_1\delta_4(q,r_4)\delta_5 + [(r_3,q)\delta_3] \to \delta_1\delta_4\delta_3(r_3,r_4)\delta_5.$

7. $\delta_1(r_1,p)\delta_2(q,r_2)\delta_3(p,q)\delta_5 \to \delta_1(r_1,p)\delta_5 + [\delta_2(q,r_2)\delta_3] \to \delta_1(r_1,r_2)\delta_3\delta_2\delta_5.$

Последовательное применение прямой и обратной (по указателям p и q соответственно) внутримолекулярной рекомбинации преобразует последовательность так же, как и операция **dlad**.

Случай 2. Повторы отрицательного указателя чередуются с повторами положительного указателя $(\ldots p\ldots \bar{q}\ldots p\ldots q\ldots)$:

$$\delta_1(p,r_1)\delta_2(\overline{r}_2,\overline{q})\delta_3(r_3,p)\delta_4(r_4,q)\delta_5 \to \delta_1\delta_4(r_4,q)\delta_5 + [\delta_2(\overline{r}_2,\overline{q})\delta_3(r_3,r_1)] =$$
$$= \delta_1\delta_4(r_4,q)\delta_5 + [(q,r_2)\overline{\delta}_2(\overline{r}_1,\overline{r}_3)\overline{\delta}_3] \to \delta_1\delta_4(r_4,r_2)\overline{\delta}_2(\overline{r}_1,\overline{r}_3)\overline{\delta}_3\delta_5.$$

Во внутримолекулярной модели:

$$\delta_1(p,r_1)\delta_2(\overline{r}_2,\overline{q})\delta_3(r_3,p)\delta_4(r_4,q)\delta_5 \to [\mathbf{hi}_q] \to$$
$$\to \delta_1(p,r_1)\delta_2(\overline{r}_2,\overline{r}_4)\overline{\delta}_4(\overline{p},\overline{r}_3)\overline{\delta}_3\delta_5 \to [\mathbf{hi}_p] \to \delta_1\delta_4(r_4,r_2)\overline{\delta}_2(\overline{r}_1,\overline{r}_3)\overline{\delta}_3\delta_5.$$

Нетрудно проверить, что и для остальных возможных случаев чередования указателя p (отрицательного) и указателя q (положительного) получим аналогичный результат: последовательное применение прямой и обратной внутримолекулярной рекомбинации (к указателям p и q соответственно) преобразует последовательность так же, как последовательное применение двух операций **hi** к указателям q и p в первой модели.

Рассмотрим межмолекулярную рекомбинацию

$$\delta_1(p,r_1)\delta_2 + \delta_3(r_2,p)\delta_4 \to \delta_1\delta_4 + \delta_3(r_2,r_1)\delta_2,$$

после однократного применения этой операции по крайней мере одна из участвующих в ней молекул потеряет часть кодирующей последовательности. Поэтому следует повторно применить операцию. Межмолекулярная рекомбинация применяется, когда есть положительный указатель:

$$\delta_1(p,q)\delta_2(\overline{p},\overline{r})\delta_3 + \overline{\delta}_3(r,p)\overline{\delta}_2(\overline{q},\overline{p})\overline{\delta}_1 \to$$
$$\to \delta_1(p,q)\delta_2\overline{\delta}_1 + \overline{\delta}_3(r,p)\overline{\delta}_2(\overline{q},\overline{r})\delta_3 \to \delta_1\overline{\delta}_2(\overline{q},\overline{r})\delta_3 + \overline{\delta}_3(r,q)\delta_2\overline{\delta}_1.$$

Во внутримолекулярной модели в этой ситуации применяется операция **hi**:

$$\mathbf{hi}_p(\delta_1(p,q)\delta_2(\overline{p},\overline{r})\delta_3) = \delta_1\overline{\delta}_2(\overline{q},\overline{r})\delta_3.$$

Нетрудно проверить, что для второго случая (первое и второе вхождение p – уходящий указатель) результат будет аналогичным: последовательное применение межмолекулярных рекомбинаций к указателям p и $\overline{p}$ преобразует

последовательность так же, как и операция **hi**. Заметим, что результатом однократного применения межмолекулярной рекомбинации являются две молекулы, различные по длине (если, конечно, указатель не находился ровно посередине молекулы). Причем эта разница тем существенней, чем ближе к краю молекулы находился указатель. Другими словами, если в живой клетке реализуется именно межмолекулярная модель, то в процессе сборки гена есть вероятность появления молекул, сильно различающихся по длине.

Сравним две модели по количеству операций, необходимых для сборки гена, считая, что все операции имеют одинаковую сложность. (Это предположение существенно упрощает ситуацию, так как даже внутри одной модели одни операции сложнее других. Совершенно ясно, например, что **ld** – самая простая из операций внутримолекулярной модели, а **dlad** – самая сложная из них.) Во внутримолекулярной модели операции **ld** и **hi** уменьшают количество MDS на 1, а операция **dlad** – на 2. В межмолекулярной модели каждая операция уменьшает количество MDS на 1.

1. Чередование двух отрицательных указателей. Во внутримолекулярной модели возможно применение операции **dlad**, в межмолекулярной – последовательное применение прямой и обратной внутримолекулярной рекомбинации.

2. Есть положительный указатель, но нет его чередования с отрицательным. В первой модели применяется операция **hi**, во второй – необходимо последовательное применение двух межмолекулярных рекомбинаций. Хотя межмолекулярная рекомбинация также уменьшает количество MDS на 1, но в операции участвуют две молекулы и количество MDS уменьшается только в одной из них. Таким образом, в этих двух случаях одна операция в первой модели соответствует двум операциям во второй. В остальных случаях идентичные преобразования последовательности в обеих моделях происходят за одинаковое количество операций.

6. Заключение

Анализ операций обеих моделей и сравнение процессов сборки в каждой из них показали, что в двух случаях (наличие двух чередующихся отрицательных указателей либо наличие такого положительного указателя, который не чередуется с отрицательным) в межмолекулярной модели требуется вдвое большее количество операций, чем во внутримолекулярной модели. В остальных случаях расположения указателей преобразование последовательности происходит за одинаковое количество операций в обеих моделях.

Неизвестно, какая из моделей описывает реальный процесс сборки генов. Заметим, однако, что при применении операций внутримолекулярной модели все молекулы остаются одинаковой длины, тогда как в межмолекулярной модели при применении межмолекулярных рекомбинаций появляются существенно различные по длине промежуточные молекулы, отвечающие одному гену. Допустим, удастся обнаружить во время процесса реорганизации генов у ресничных такие молекулы. Это обстоятельство исключит возможность реализации в клетке внутримолекулярной модели и будет весомым аргументом в пользу межмолекулярной модели.

Представим, как мог бы выглядеть такой эксперимент. Остановим процесс сборки гена, например, заморозив клетку в тот момент, когда в ядре происходит реорганизация генетического материала. Возьмем образцы молекул ДНК, находящиеся в этот момент в развивавшемся макроядре, и применим известный в биохимии метод *гель-электрофореза*, который используют для измерения длин молекул ДНК. Подробное описание техники электрофореза см. в [1].

Литература

1. Ehrenfeucht A., Harju T., Petre I. et al. Computation in Living Cells: Gene Assembly in Ciliates. Berlin: Springer, 2004.
2. Adleman L. M. Molecular computation of solutions to combinatorial problems // Sciencc. 1994. Vol. 226. P. 1021–1024.
3. Паун Г., Розенберг Г., Саломаа А. ДНК-компьютер. Новая парадигма вычислений. М.: Мир, 2004.
4. Высоцкая Л. В., Глаголев С. М. Дымшиц Г. М. и др. Общая биология. М.: Просвещение, 1995.