Вычисления с использованием ДНК

Чутков Ростислав и Петров Александр

октябрь 2006 г.

Содержание

1. Введение

Вычисления на ДНК - это раздел области молекулярных вычислений, нового междисциплинарного направления исследований на границе молекулярной биологии и компьютерных наук. Основная идея ДНК-вычислений - построение новой парадигмы вычислений, новых моделей, новых алгоритмов на основе знаний о строении и функциях молекулы ДНК и операций, которые выполняются в живых клетках над молекулами ДНК при помощи различных ферментов. Основные надежды, которые возлагаются на

область ДНК-вычислений в практическом смысле - это новые методы синтеза веществ и объектов на молекулярном уровне.
Область ДНК-вычислений несет новые идеи для специалистов по нанотехнологиям, идеи, связанные с программируемым синтезом структур на наноуровне, со сборкой методами “снизу вверх"с использованием механизмов самоорганизации и самоформирования на молекулярном уровне.
Для специалистов в области компьютерных наук, теории вычислений, парадигма ДНК-вычислений интересна новыми открывающимися возможностями: новыми моделями вычислений, новыми алгоритмами, возможностью решения задач, не решаемых в рамках классической парадигмы вычислений, возможностью исследования процессов массового параллелизма, которые средствами классической парадигмы даются трудно.
Эти новые идеи в дальнейшем будут использованы в построении Биологического нанокомпьютера, который будет способен хранить террабайты информации при объеме всего несколько микрометров,совершать миллиарды операций в секунду при затратах энергии не более одной миллиардной доли ватта. Низкая стоимость “материалов", использующихся для создания и обслуживания компьютера и его способность внедрять в клетку живого организма откроет новые горизонты для развития науки.

2. Эелементарные операции с ДНК

Разберем основные “команды", которые доступны нам при работе с ДНК в лабораторном опыте, и на которые должны опираться и теоретические разработки в области компьютерных наук.
Молекула ДНК (рис. 1) представляет собой двойную ленту, составленную из четырех оснований: **А** (аденин), **Т** (тимин), **Г** (гуанин), **Ц** (цитозин). На рис. 2 изображен ДНК под электронным микроскопом.

Рис. 1.

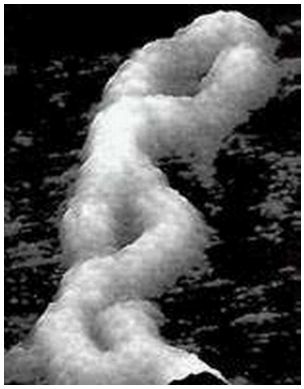

Рис. 2.

Диаметр двойной спирали ДНК - 2нм, расстояние между соседними парами оснований - 0.34нм. Полный оборот двойная спираль делает через 10 пар. ДНК простейших типов вирусов содержит всего несколько тысяч звеньев, бактерий - несколько миллионов, а высших организмов - миллиарды.

2.1. Ренатурация. Денатурация

Комплиментарность. Комплементарность оснований заключается в том, что образование водородных связей при соединении одинарных цепочек ДНК в двойную цепочку возможно только между парами **A-T**

и **Г-Ц**(рис. 3). Этот же рисунок иллюстрирует операции ренатурации и денатурации. Ренатурация - это соединение двух одинарных цепочек ДНК за счет связывания комплементарных оснований. Денатурация - разъединение двойной цепочки и получение двух одинарных цепочек. Денатурация и ренатурация происходят при нагревании и охлаждении раствора с ДНК соответственно. Плавление ДНК происходит в диапазоне температур 85-95°C. Некоторые катализаторы позволяют понизить температуру этого процесса.

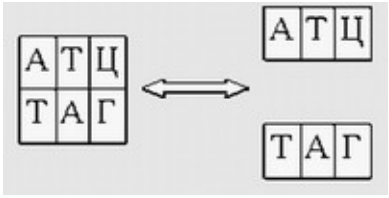

Рис. 3.

2.2. Дополнение цепочки

ополнение цепочки ДНК происходит при воздействии на исходную молекулу ферментов - полимераз(рис. 4). Для работы полимеразы необходимо наличие:

1. Одноцепочечной матрицы, которая определяет добавляемую цепочку по принципу комплементарности;
2. Праймера - двухцепочечного участка, который присоединен к матрице, и к которому присоединяются свободные нуклеотиды;
3. Свободных нуклеотидов в растворе.

Существуют полимеразы, которым не требуются матрицы для удлинения цепочки ДНК. Например, терминальная трансфераза обавляет одинарные цепочки ДНК к обоим концам двухцепочечной молекулы.

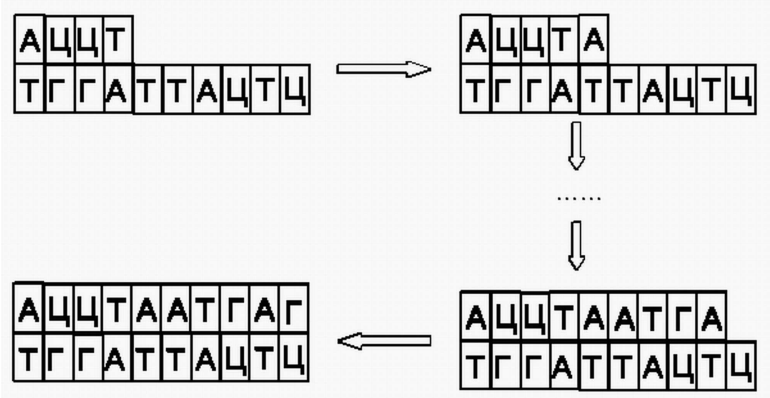

Рис. 4.

2.3. Удлинение

уществуют полимеразы, которым не требуются матрицы для удлинения цепочки ДНК. Например, терминальная трансфераза добавляет одинарные цепочки ДНК к обоим концам двухцепочечной молекулы.

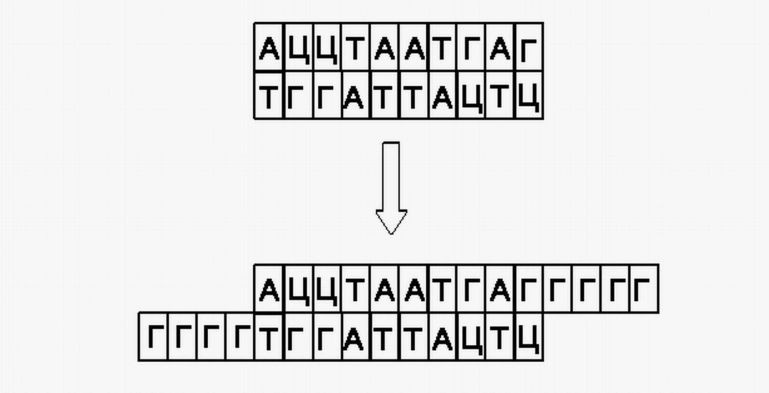

Рис. 5.

2.4. Укорочение

За укорочение и разрезание молекул ДНК отвечают ферменты - нуклеазы. Различают эндонуклеазы и экзонуклеазы. Экзонуклеазы осуществляют укорочение молекулы ДНК с концов, эндонуклеазы же разрушают внутренние фосфодиэфирные связи в молекуле ДНК. Экзонуклеазы могут укорачивать одноцепочечные молекулы и двухцепочечные, с одного конца или с обоих.

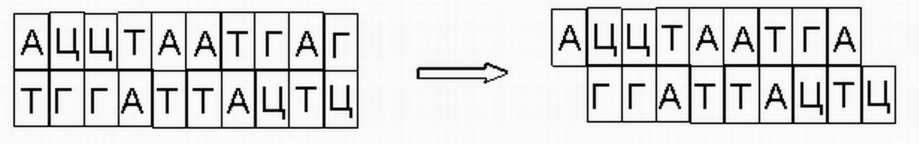

Рис. 6.

2.5. Разрезание

Эндонуклеазы могут быть весьма избирательными в отношении того, что они разрезают, где они разрезают и как они разрезают. Сайт-специфичные эндонуклеазы - рестриктазы - разрезают молекулу ДНК в определенном месте, которое закодировано последовательностью нуклеотидов - сайтом узнавания. Разрез может быть прямым, или несимметричным, как на рис. 7. Разрез может проходить по сайту узнавания, или же вне его.

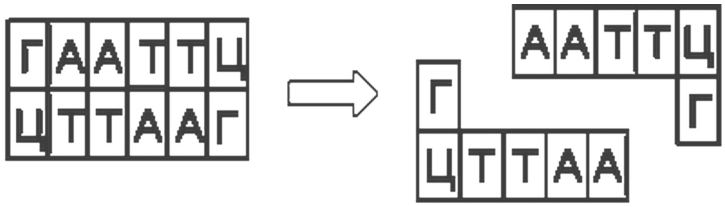

Рис. 7.

2.6. Сшивка

Сшивка - операция, обратная операции разрезания, происходит под воздействием ферментов - лигаз. Когда двухцепочечные молекулы ДНК имеют комплементарные одноцепочечные концы, то говорят,

что это “липкие концы". Липкие концы соединяются вместе с образованием водородных связей, однако при этом остаются “ зазоры", называемые насечками, т.е. отсутствующие фосфодиэфирные связи между соседними нуклеотидами одинарной цепочки. Фосфодиэфирные связи гораздо сильнее водородных, поэтому, в частности, при нагревании сначала разрушаются водородные связи, что приводит к образованию двух одинарных цепочек. Лигазы как раз и служат для того, чтобы закрыть насечки, т.е. способствовать образованию в нужных местах фосфодиэфирных связей.

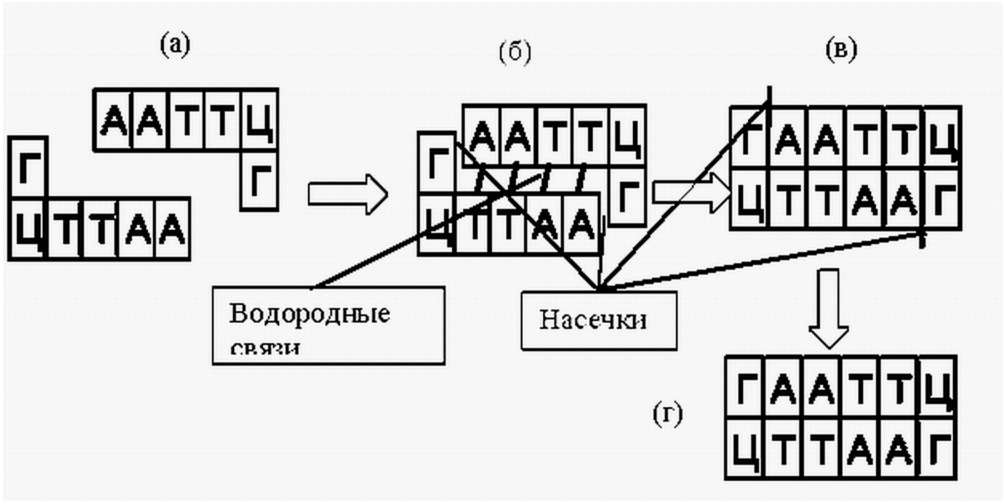

Рис. 8.

2.7. Модификация

Потребность в модификации может возникнуть, например, для того, чтобы исключить молекулу из какой-либо операции. В живой клетке рестриктазы играют роль защитников от агрессии - например, в клетке бактерии, рестриктазы разрушают ДНК вируса-агрессора. Чтобы собственная ДНК не подверглась разрезанию, она модифицируется (рис. 9). Существует несколько типов модифицирующих ферментов - метилазы, фосфатазы и т.д. Метилаза имеет тот же сайт узнавания, что и соответствующая рестриктаза. При нахождении нужной молекулы, метилаза модифицирует участок с сайтом так, что рестриктаза уже не сможет идентифицировать эту молекулу.

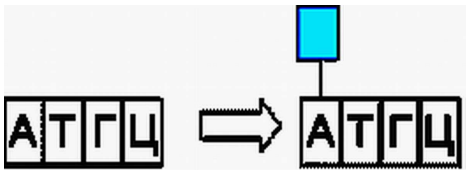

Рис. 9.

2.8. Полимеразная цепная реакция

Иногда необходимо увеличить количество определенных фрагментов ДНК. Это делается при помощи метода полимеразной цепной реакции. Этот метод позволяет получить миллионы копий желаемой молекулы, даже если начать всего лишь с одного ее экземпляра. Метод применяется для двойных цепочек длиной от 100 до 35000 звеньев.
Каждая итерация метода требует трех стадий: денатурации, праймирования и удлинения. На каждой

итерации количество молекул (теоретически) увеличивается в 2 раза.
Денатурация происходит при нагревании, поэтому полимеразы, находящиеся в растворе, должны быть устойчивыми к воздействию температуры. Такие полимеразы выделены из бактерий, живущих в горячих источниках.
Праймирование заключается в том, что к концам обеих одноцепочечных молекул, полученных в результате денатурации, присоединяются праймеры - заранее синтезированные короткие одноцепочечные молекулы. Праймирование необходимо для корректной работы полимераз. На третьем этапе итерации происходит дополнение праймированных одноцепочечных молекул до двухцепочечных под действием полимераз.
В результате всей итерации из одной молекулы мы получили две ее копии,как видно на рис. 10.

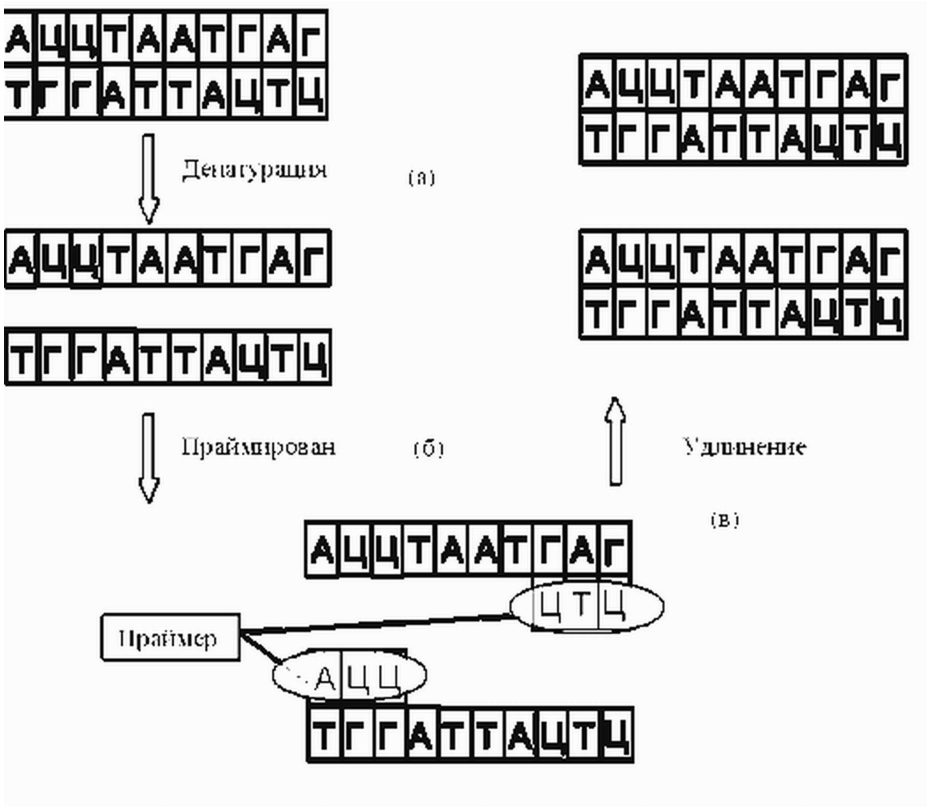

Рис. 10.

2.9. Спелтение

Операция сплетения(рис. 11) представляет собой последовательное разрезание двух молекул ДНК и сшивания полученных одноцепочечных молекул между собой.

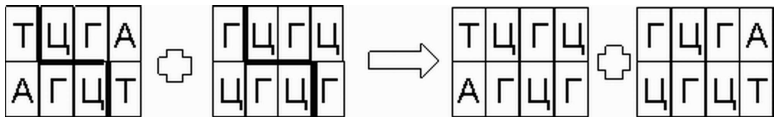

Рис. 11.

2.10. Секвенирование

Секвенирование - это определение последовательности нуклеотидов в ДНК. Для секвенирования цепочек различной длины применяют различные методы. При помощи метода праймер-опосредованной прогулки удается на одном шаге секвенировать последовательность в 250-350 нуклеотидов. Отдельные шаги этого метода были автоматизированы, что позволило с легкостью секвенировать последовательности длиной в десятки тысяч пар нуклеотидов.Естественно, после открытия рестриктаз стало возможным секвенировать длинные последовательности по частям. Как известно, не так давно был закончен проект “Геном человека секвенирование всего генома человека.

2.11. Гель-электрофорез

Гель-электрофорез используется для разделения молекул ДНК по длине. Молекулы ДНК имеют отрицательный заряд, поэтому, если их поместить в гель и приложить постоянное электрическое поле, то они будут двигаться по направлению к аноду, причем молекулы меньшей длины будут двигаться быстрее. Когда первые, самые короткие молекулы достигают анода, процесс останавливают. Для маркировки молекул используют либо методы окрашивания, либо радиоактивные маркеры. Часто используют калибровочные молекулы (молекулы известной длины).
Пример снимка, полученного в результате гель-электрофореза представлен на рис. 12. Молекулы двигались слева направо. После остановки видно, что молекулы одинаковой длины двигаются единым фронтом, образуя в геле дискретные полосы. По первой дорожке пущены калибровочные молекулы.

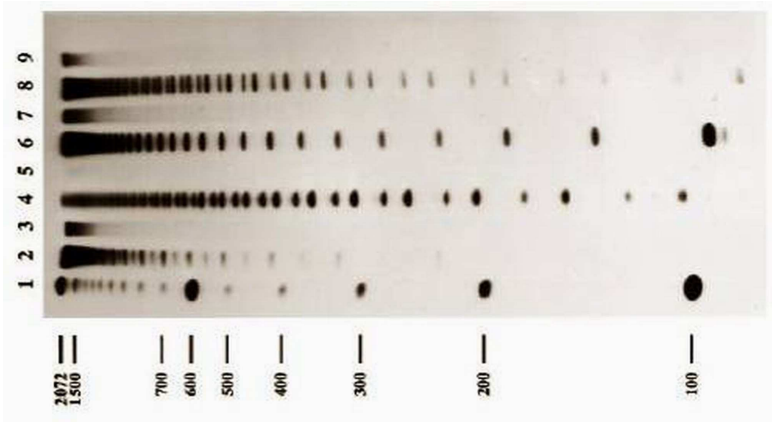

Рис. 12.

2.12. Синтез

Современные приборы автоматического химического синтеза позволяют быстро создавать одноцепочные последовательности длиной ≤ 50 звеньев. Более длинные цепочки (≥ 300) не удается синтезировать теми же методами, что и короткие. Дело в том, что синтез ДНК происходит последовательно, нуклеотид за нуклеотидом, при этом, естественно, на каждом шаге происходят потери: не все введенные в реакционную среду нуклеотиды присоединяются к нужным цепочкам. Доля успешно присоединившихся нуклеотидов на одной итерации цикла синтеза называется эффективностью цикла. Как правило, реальная эффективность составляет 95%, хотя иногда удается достичь и 99% эффективности. Следовательно, при средней эффективности 95% в результате синтеза последовательности длиной 100 звеньев выход составит 0,6%, что нельзя считать удовлетворительным. Поэтому при синтезе длинных цепочек сначала синтезируют их короткие составляющие, а затем из них при помощи операций сшивания получают длинную цепочку. Таким образом получают последовательности длиной более 1000 звеньев.

3. Эксперименты с ДНК

3.1. Эксепримент Эдлмана

В 1994 г. Л.Эдлман продемонстрировал решение задачи о доказательстве существования гамильтонова пути в графе при помощи ДНК-вычислителя. Задача формулируется следующим образом: существует ли в данном направленном графе G, в котором выделена начальная и конечная вершины, гамильтонов путь, т.е. путь, который проходит через каждую вершину в точности один раз (рис. 13). Для решения задачи был применен следующий алгоритм:

- Шаг 1. Вход. Ориентированный граф G с n вершинами, среди которых выделены 2 вершины - v_{in} и v_{out};
- Шаг 2. Породить большое количество случайных путей в G;
- Шаг 3. Отбросить все пути, которые не начинаются с v_{in} или не заканчиваются на v_{out};
- Шаг 4. Отбросить все пути, которые не содержат точно n вершин;
- Шаг 5. Для каждой из n вершин v отбросить пути, которые не содержат v;
- Шаг 6. Выход. Да, если есть хоть один путь, нет - в противном случае.

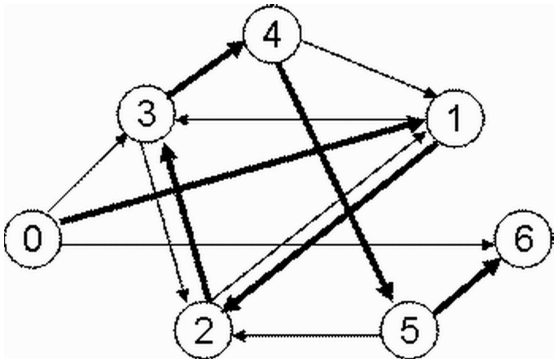

Рис. 13.

Реализация алгоритма:
Каждая вершина графа кодируется последовательностью 20 нуклеотидов. На рисунке 14 даны коды вершин V_0 и V_1.
Ребра кодируются так: берется вторая половина цепочки для начальной вершины, и первая половина цепочки для конечной вершины, эти цепочки соединяются в одну, затем берется комплементарная к полученной цепочке последовательность нуклеотидов, которой и кодируется соответствующее ребро графа. Общее правило кодирования ребер: дан оператор комплементарности, затем показан процесс кодирования ребра U_{01}: взята вторая половина цепочки для вершины V_0 - V_0'', первая половина цепочки для вершины V_1 - V_1', эти цепочки последовательно соединены в одну, и к получившейся цепочке применен оператор комплементарности.
Ключевой момент опыта:
Пусть элементарный объем реакционной среды содержит три молекулы - кодирующие соответственно два различных ребра $U_{i,j}$, $U_{j,k}$ и общую для них вершину V_j. Тогда произойдет соединение этих молекул в одну длинную цепочку $U_{i,k}$, кодирующую путь из вершины i в вершину k (рис. 15). Полученная цепочка будет способна соединяться с другой подходящей цепочкой, кодирующей вершину, или другой фрагмент пути на графе. В ходе опыта сначала синтезируются цепочки, которые кодируют ребра и вершины графа, затем они в необходимом количестве запускаются в реакционную среду. Через некоторое время в среде

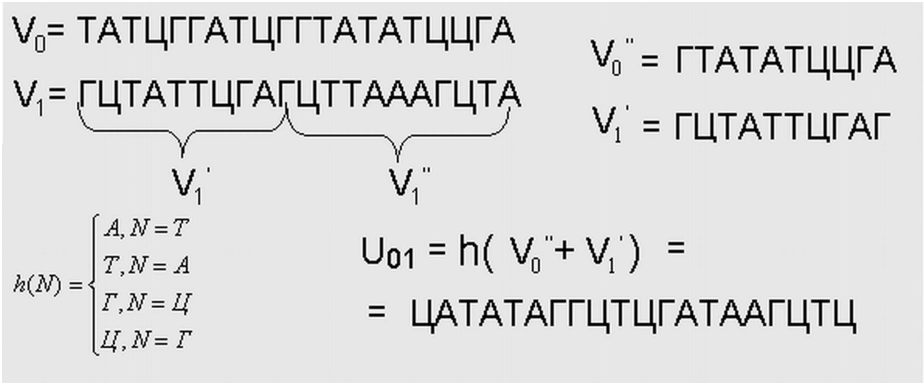

Рис. 14.

образуются молекулы, которые соответствуют всем возможным путям на графе. Далее вопрос только в том, чтобы отыскать среди всех возможных путей гамильтонов путь, что и делается при помощи трех шагов фильтрации, описанных в алгоритме. Опыт Эдлмана занял 7 дней, больше всего времени заняла процедура фильтрации на шаге 4.

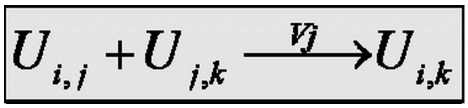

Рис. 15.

Итог:
Эксперимент Эдлмана показал,что, пользуясь вычислениями на ДНК, можно эффективно решать задачи переборного характера.
Обозначил технику, которая, в дальнейшем послужила основой для создания модели параллельной фильтрации.
Важно то, что построив эффективную реализацию Алоритма Эдлмана мы научимся решать NP-полные задачи за полиномиальное время.

3.2. Эксперимент Шапиро

Опыт, осуществленный в 2001 г. группой Э. Шапиро, принципиально отличается от опыта Эдлмана тем, что и “исходные данные", и “программа"описываются молекулами ДНК, в то время как в опыте Эдлмана “программа это, по существу, последовательность реакций, задаваемых человеком.В опыте Э. Шапиро был реализован конечный автомат, т.е. система, состоящая из множества состояний, алфавита (множества символов, которые могут поступать на вход), начального состояния, множества заключительных состояний и функции переходов.Данный автомат изображен на рис. 16.
Автомат может находиться в двух состояниях - S0 и S1. Алфавит автомата состоит из двух символов - a и b. На вход автомату подается последовательность символов a и b. Автомат отвечает на вопрос - четное или нечетное количество символов a содержится во входной последовательности. Автомат может отвечать на 765 подобных “вопросов". Программирование автомата заключается в задании функции переходов на рис 16. И переходам, и состояниям, и входной последовательности в опыте Шапиро отвечают молекулы ДНК. “Программа"для этого автомата (правила переходов) записывается следующим образом:

$$S0, a \rightarrow S1$$

$$S0, b \rightarrow S0$$

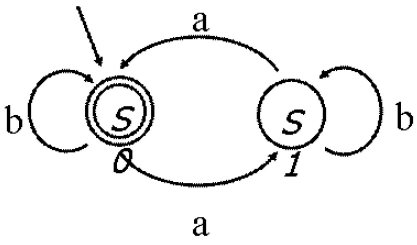

Рис. 16.

$$S1, a \to S0$$

$$S1, b \to S1$$

Если по окончании обработки входной последовательности автомат находится в состоянии S0 - это означает, что во входной последовательности было четное количество символов a, если же он находится в состоянии S1 - количество символов a было нечетным. Вычисления происходят по перечисленным правилам, причем воспринимать их следует “буквальным образом", т.е. как замену строки S0a на строку S1 для первого правила. Тогда процесс вычислений будет проходить, например, так: Так же просто работает и

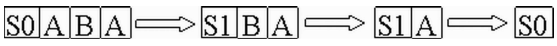

Рис. 17.

автомат на ДНК. Символы алфавита - a и b - кодируются молекулами ДНК (рис. 18). Далее кодируются “полные"состояния автоматов, т.е. состояние автомата + символ на входе. Таких “полных"состояний получается 4: S0,A; S0,B; S1,A; S1,B.
В нашем случае “программа"автомата содержит 4 перехода. Их коды показаны на рис. 19.

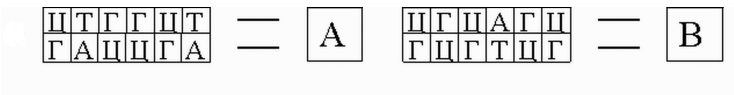

Рис. 18.

Забегая вперед, отметим, что каждый шаг работы автомата выполняется за два “молекулярных шага": к закодированной входной последовательности присоединяется нужный переход, образовавшиеся насечки закрываются посредством действия лигазы, затем необходимо отделить от полученной цепочки ненужную часть так, чтобы конец оставшейся цепочки кодировал следующее “полное"состояние автомата: следующий входной символ и собственно состояние. Это и происходит при помощи рестриктазы. Для того, чтобы рестриктаза работала корректно, необходимо так закодировать переходы, чтобы они содержали в себе сайты узнавания - точку отсчета для рестриктазы. Для перехода $S0, A \to S1$ сайт узнавания показан на рис. 20. В конце входной цепочки располагается символ-терминатор (рис. 21). По окончании работы автомата получается одна из молекул- S0,T, или S1,T (рис. 21), к которым присоединяется одна из молекул - индикаторов конечного состояния, различных по длине, что позволяет выяснить конечное состояние при помощи гель-электрофореза. Непосредственно опыт Э.Шапиро на примере простой входной последовательности показан на рис. 22 и 23. Опыт начинается с синтеза молекул, соответствующих символам алфавита, переходам, полным состояниям, символу-терминатору и молекулам - индикаторам конечного состояния. Далее все эти молекулы в необходимом количестве помещаются в реакционную среду, в которую дополнительно помещаются и необходимые рестриктазы и лигазы.
Пусть на вход автомата подается последовательность ABA и он работает по уже описанным правилам. Начальное состояние автомата - $S0$. Исходная цепочка ДНК состоит из фрагментов $S0, A, B, A$ и выглядит

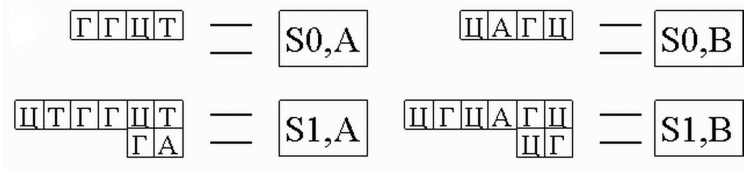

Рис. 19.

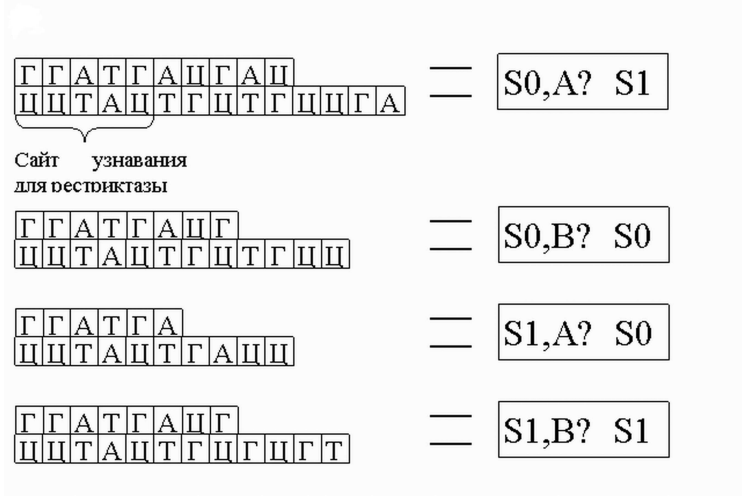

Рис. 20.

так, как на рис. 22а. Очевидно, что, в силу принципа комплементарности, из четырех возможных вариантов, к молекуле, кодирующей входящую последовательность и начальное состояние может присоединиться только переход $S0, A \to S1$ (рис. 22б). Молекулы соединяются липкими концами, далее при помощи лигазы закрываются насечки, т.е. образуются более прочные фосфодиэфирные связи между соседними нуклеотидами на местах стыков в одинарных цепочках. В результате получается молекула, показанная на рис. 22в, которая содержит сайт узнавания для рестриктазы. Рестриктаза, определив сайт узнавания, разрезает молекулу строго в местах, отмеченных на рис. 22в, т.е. отступая на 9 нуклеотидов в верхней цепочке и на 13 в нижней от границы сайта узнавания. Говорить о верхней и нижней цепочке можно потому, что молекула, в силу особенностей на уровне химических связей, имеет направление. После разрезания молекула становится такой, как на рис. 22г. Отметим, что конец молекулы кодирует теперь не просто следующий входной символ - В, а “полное"состояние, т.е. $S1, B$.

Далее, к полученной молекуле может присоединиться только переход $S1, B \to S1$ и никакой другой. Затем происходит сшивка и разрезание (рис. 22д) таким же образом, как и на предыдущем шаге. Обратим внимание на то, что переход $S1, B \to S1$ содержит тот же сайт узнавания, что и переход $S0, A \to S1$, да и любой другой переход, что позволяет обойтись в опыте рестриктазой одного типа.

В результате последней итерации получилась молекула, показанная на рис. 22е, которая кодирует “полное"состояние $S1, A$, т.е. автомат находится в состоянии S1 и на входе символ А. Следующая итерация аналогична двум предыдущим - прилипание перехода $S1, A \to S0$, сшивка и разрезание (рис. 23е). Для того, чтобы не загромождать рисунок, символ-терминатор был опущен, поэтому можно представить, что полученная в ходе третьей итерации молекула (рис. 23з) заканчивается символом-терминатором. Полученная молекула соответствует состоянию S0,Т. Теперь к ней может присоединиться только молекула-индикатор, с липким концом АГЦГ, имеющая определенную длину. Под действием лигазы происходит сшивка. Затем при помощи калибровочных молекул и гель-электрофореза мы можем выяснить, что в ходе вычислений получен результат - автомат обработал входную последовательность полностью, до символа-терминатора и находится в состоянии S0, а, значит, входная последовательность содержала четное количество симво-

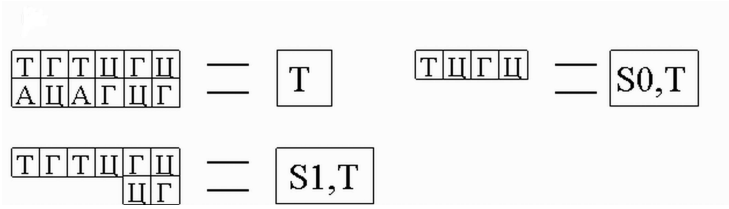

Рис. 21.

лов А.
В опыте одновременно работали 10^{12} автоматов с одинаковым “программным обеспечением". Входные данные, в принципе, для автоматов могут быть различными. “Вычислительная мощность"составляла 10^9 переходов в секунду с вероятностью больше, чем 99,8%. На вопрос о том, смогут ли ДНК-вычислители конкурировать в будущем с существующими процессорами, Э. Шапиро отвечает, что такой вопрос даже не ставится. Как и многие другие исследователи, Э. Шапиро полагает, что основное назначение ДНК вычислителей - это тонкий химический синтез, сборка нужных молекул и конструкций. В самом деле, как мы видим, собственно вычисление - обработка входной последовательности, занимает очень малое время. Значительное время тратится на то, чтобы понять, какой собственно результат получен.

3.3. Эксперимент Винфри

В лаборатории молекулярных вычислений в Калифорнийском технологическом институте под руководством Э. Винфри успешно разрабатываются методы синтеза различных поверхностей при помощи ДНК. В этих экспериментах переосмысливается само понятие вычисления. Оказывается, можно использовать двумерные плитки различной формы, которые могут взаимодействовать по локальным правилам (соединяться друг с другом), для того, чтобы получить в результате взаимодействия множества плиток желаемую глобальную структуру. При этом под вычислением понимается процесс создания такой структуры.
Разберем простейший пример вычисления, который приводится в работах сотрудников лаборатории Э. Винфри. Пусть необходимо реализовать простейший алгоритм - счетчик. Для этого нам понадобятся рабочие элементы четырех типов , и элементы, задающие граничные условия - трех типов (рис. 24).

Правило создания структуры чрезвычайно простое: во главу угла ставится плитка S, две оставшиеся граничные плитки выкладываются в направлении вверх и влево, затем, справа налево ряд за рядом укладываются рабочие плитки. При этом укладывать плитку можно лишь в том случае, если уже уложены ее соседи снизу и справа. Результат показан на рис. 25 и соответствует счетчику.

Еще в 60-х годах доказано, что при помощи “плиточных вычислений"можно реализовать машину Тьюринга. Однако обратное утверждение неверно - проблема замощения плоскости плитками различной формы не разрешима в парадигме машины Тьюринга.

В работах Э. Винфри отработана методика перехода от двумерных плиток к молекулам ДНК. Например описывается эксперимент синтеза известной фрактальной структуры - ковра Серпинского. В опыте используются всего 4 плитки, которые соответствуют правилам таблицы истинности для оператора XOR (рис. 26).Начальный слой укладывается из плиток типа T-00. Затем плитки укладываются по направлению снизу вверх (рис.27).

Далее, каждой плитке ставится в соответствие молекула ДНК. В реальном опыте используются несколько иные плитки, чем показанные на рис. 26. Схема опыта Винфри на порядок сложнее рассмотренных опытов Эдлмана и Шапиро.

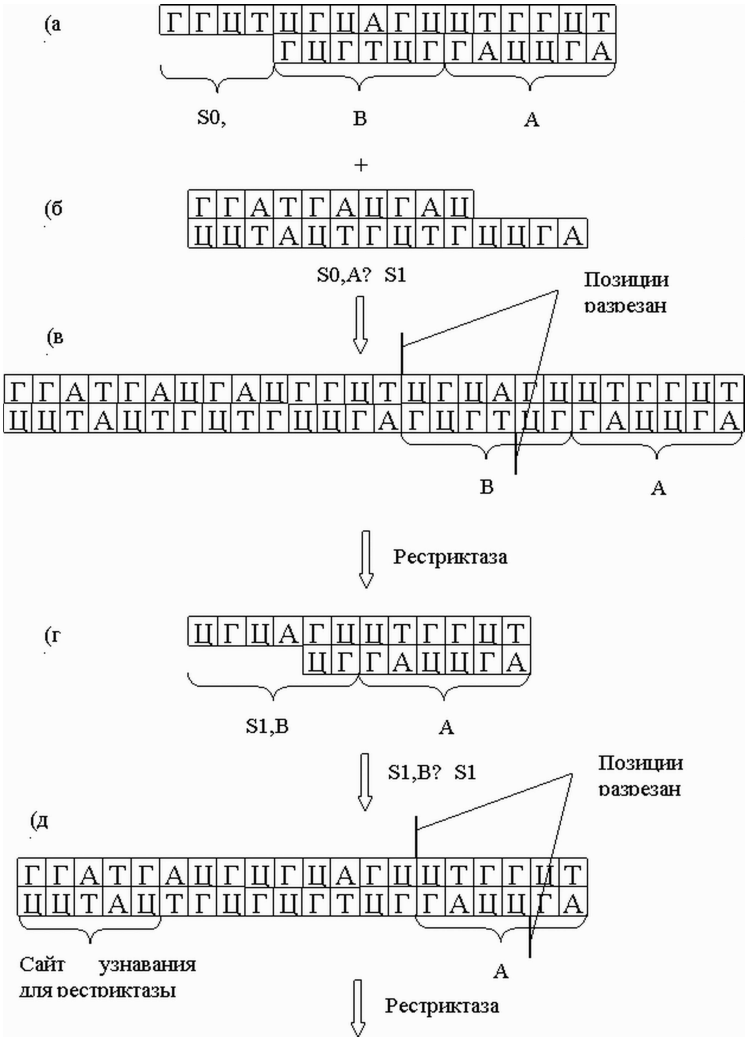

Рис. 22.

В результате опыта под атомно-силовым микроскопом можно видеть следующую структуру (рис. 28). На рисунке видно, что в результате опыта получаются достаточно большие (порядка десятков слоев) структуры, в которых количество ошибок не слишком велико (ошибки отмечены крестиками).

4. Модели и попытки формализации

После проведения первых простых опытов возникает потребность в общих моделях молекулярных вычислений, которые бы позволяли проектировать новые эксперименты и обобщать существующие.

4.1. Модель параллельной фильтрации (Parallel Filtering Model)

Происхождение данной модели уходит корнями в эксперимент Элдмана.В модели основной упор делается на фильтрацию потому, что множество всевозможных решений задачи получается уже на первом шаге за счет того, что взаимодействующие молекулы ДНК спроектированы нужным образом. А основная

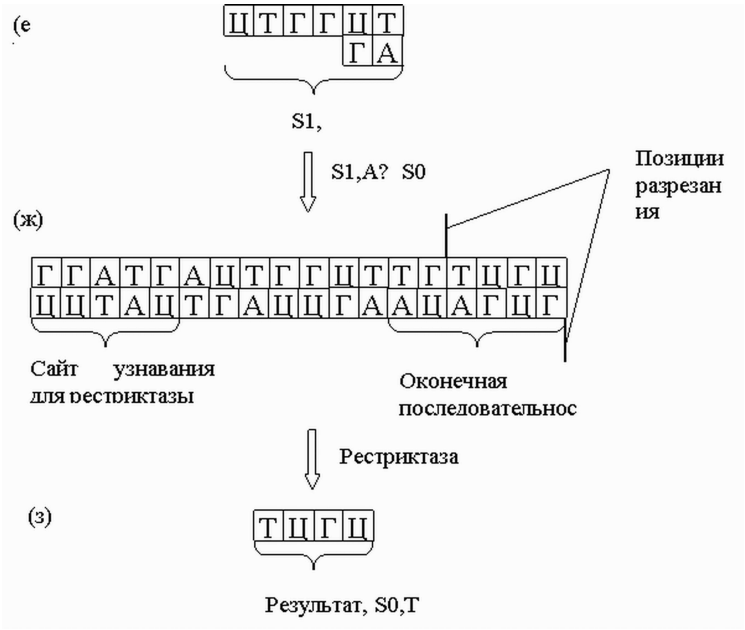

Рис. 23.

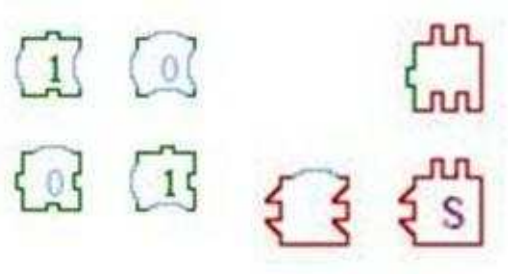

Рис. 24.

часть алгоритма - это извлечение нужного результата из множества всевозможных результатов.

Основные определения:

Пробирка - это мультимножество слов (конечных строк) над алфавитом **{А, Ц, Г, Т}**.

Мультимножество - это, по сути, объединение множеств, каждое из которых содержит элементы только одного типа, или же о мультимножестве можно думать как о множестве, которое определяется множеством неповторяющихся элементов, каждому из которых приписано натуральное число, означающее количество элементов данного типа в мультимножестве. Следующие основные операции были первоначально определены для пробирок, т.е. мультимножеств одинарных цепочек ДНК. Однако их подходящие модификации будут применяться и к двойным цепочкам.

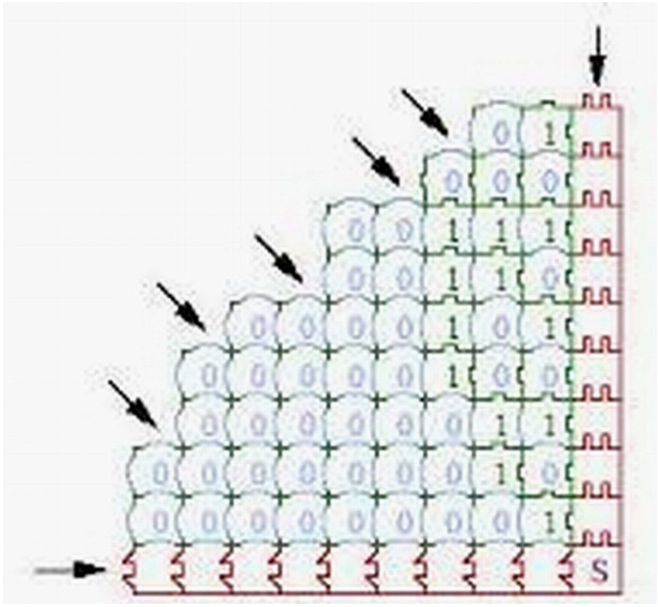

Рис. 25.

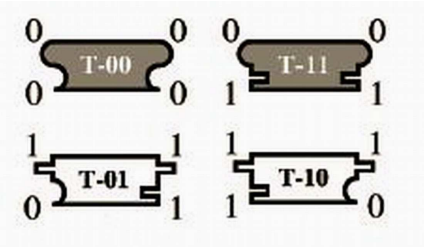

Рис. 26.

Слить - образовать объединение $N_1 \cup N_2$(в смысле мультимножеств) двух заданных пробирок N_1 и N_2.

Размножить - изготовить две копии данной пробирки N.

Обнаружить - возвратить значение *истина*, если данная пробирка N содержит по крайней мере одну цепочку ДНК, в противном случае возвратить значение *ложь*.

Разделить (или Извлечь) - по данным пробирке N и слову w над алфавитом **{A,Ц,Г,Т}** изготовить две пробирки $+(N, w)$ и $-(N, w)$ такие, что $+(N, w)$ состоит из всех цепочек в N, содержащих w в качестве (последовательной) подстроки, а $-(N, w)$ состоит из всех цепочек в N, которые не содержат w в качестве подстроки.

Разделить по длине - по данным пробирке N и целому числу n, изготовить пробирку $L(N, \leq n)$, состоящую из всех цепочек из N длины не больше n .

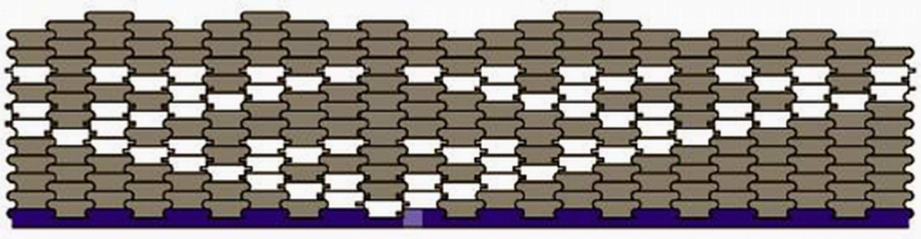

Рис. 27.

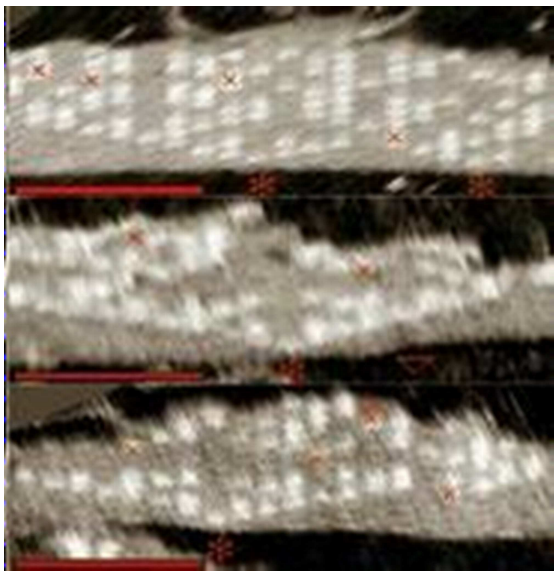

Рис. 28.

Разделить по префиксу (суффиксу) - по данным пробирке N и слову w, изготовить пробирку $B(N,w)$ (соответственно $E(N,w)$), состоящую из всех цепочек в N, начало (соответственно конец) которых совпадает со словом w.

В приведенных терминах стадия фильтрации в опыте Эдлмана может быть описана следующей программой, которая начинает свою работу после того, как произошло сшивание всех нужных молекул и в пробирке N образовалось множество всех возможных путей в графе G(Каждый из олигонуклеотидов $s_i, 0 \leq i \leq 6$, имеет длину 20).

Алгоритм Эдлмана :

1. Ввести (N)
2. $B(N, s_0) \rightarrow N$ - выделить все цепочки, которые начинаются с вершины s_0)
3. $E(N, s_6) \rightarrow N$ - выделить все цепочки, которые заканчиваются на s_6)
4. $L(N, \leq 140) \rightarrow N$ - выделить все цепочки длиной не больше 140)

5. Для i от 1 до 5 выполнить $+(N, s_i) \rightarrow N$(для каждой из вершин от s1 до s5 выделить только те цепочки, которые содержат данную вершину)

6. Обнаружить (N) - *true* если осталась хоть одна цепочка, *false* - в противном случае).

Как мы видим, Модель параллельной фильтрации соответствует классической парадигме вычислений и реализуется в три стадии: генерация всех вариантов, параллельный отсев всех неудовлетворительных вариантов и расшифровка решения.

4.2. Плиточная модель

Существует задача об отыскании набора геометрических фигур на плоскости (плиток), которыми Евклидова плоскость может быть покрыта только непериодическим образом. В 1961 г. было показано, что невозможно создать алгоритм, который определяет, можно ли покрыть плоскость при помощи заданного набора плиток, или нет. Позже был предъявлен набор из 20426 плиток, которыми можно покрыть плоскость только непериодически. В дальнейшем количество плиток было сокращено сначала до 104, а затем и до 6(набор Робинсона), и, наконец, до двух(набор Пенроуза на рис.29).

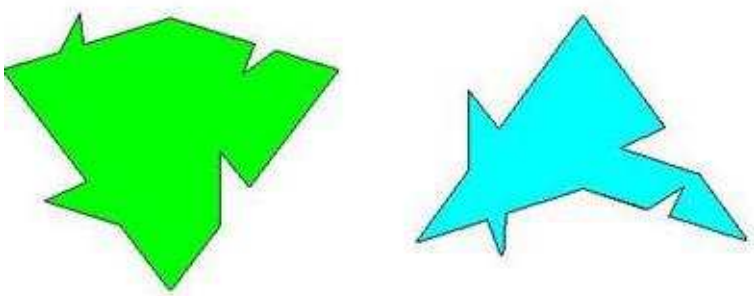

Рис. 29.

В свете мысли о задаче покрытия и мысли об экспериментах Э.Винфри, в которых исходным материалом служат наборы плиток, которые затем преобразуются в молекулы ДНК, рождается идея о разработке парадигмы ДНК-вычислений именно в “плиточных терминах". При этом ДНК-вычислитель будет представлять собой клеточный автомат из клеток произвольной формы, а локальные правила взаимодействия клеток будут определяться их формой. С одной стороны, такой автомат будет дискретным, т.к. будет состоять из отдельных взаимодействующих плиток, и к нему будет применимо понятие шага. А с другой стороны, локальные правила задаются за счет непрерывной формы границы взаимодействующих плиток. Данный подход сразу же обеспечивает возможность описания параллельных процессов, которые изначально присущи ДНК-вычислителю. При всей фантастичности данного подхода, нельзя не признать, что он несет значительный эвристический потенциал.

Теоретическим базисом “плиточной"модели могут быть, с одной стороны, все работы, относящиеся к проблеме покрытия (Ванга, Бергера, Робинсона, Пенроуза), с другой стороны - работы Э. Винфри, направленные на получение нужных структур на практике, а с третьей - работы по теории клеточных автоматов с “квадратными клетками".

5. Текущие результаты

5.1. Практические результаты

Теперь попробуем оценить практическую пользу поставленных экспериментов.

В реализации эксперимента Эдлмана оптимальный маршрут обхода отыскивался всего для 7 вершин графа, а сами вычисления длились семь (!) дней. Человеку на решение подобной задачи понадобилось бы не более пяти минут. Компьютер на основе кремниевого чипа решит миллион подобных задач за одну секунду.

Конечно, с двумя сотнями вершин обычный компьютер уже не справится, слишком уж много времени потребуют вычисления. Но и ДНК-компьютер такую задачу не осилит - для ее решения потребуется количество ДНК, по массе превышающее вес всей нашей планеты.

Эксперимент Шапиро предлагает замечательный конечный автомат, реализующий однобитный счетчик. Но этот счетчик просто распадается после 756 вопросов о четности количества символов “a"во входном потоке. Кроме того, результат вычислений в данном случае неоднозначен. Где-то 0.02% счетчиков выдавали неверные ответы на поставленный вопрос. Конечно, процент ошибки очень низок, но сама возможность ошибки вынуждает создавать контрольные схемы, проверяющие результат, схемы куда более сложные, чем сам автомат. Эксперимент Винфри, синтезирующий структуру, напоминающую ковер Серпинского, не может похвастаться абсолютной точностью создания этой структуры. В практической реализации эксперимента можно было наблюдать не менее 5% ошибок.

Не так давно, в начале 2006 года был построен конечный автомат на ДНК, реализующий игрока в крестики нолики на поле 3x3. Получая на вход в специальном формате ходы противника, автомат способен свести любую партию к ничье, или даже выиграть, если противник ошибется. Но у этого автомата есть один значительный недостаток - чтобы считать его ход требуется в среднем 30 минут и вычислительная работа обычного кремниевого компьютера.

5.2. Решенные задачи

Поиск гамильтонова пути в графе	1994
Достижимость пропозициональных формул	1994
3-раскраска графа	1995
Quantified Boolean formulae	1995
Indendent Set	1996
Задача о рюкзаке	1996
Задача изоморфизма с подграфом	1996
Задача о клике	1996
MAX-CNF SAT	1996
Задача о выполнимости для схем	1996
(3-2) System	1997
Shortest common superstring	1998
Bounded Post correspondence	2000

5.3. Программные средства

Xgrow. Симулятор Xgrow разработан в лаборатории молекулярных вычислений Калифорнийского технологического института Э.Винфри. Он использует в своей работе модели aTAM (abstract Tile Assembly Model) и kTAM (kinetic Tile Assembly Model). соответственно. Попросту говоря, симулятор Xgrow позволяет имитировать процесс синтеза различных структур, получая на входе набор плиток, а также позволяет

оценить возможные ошибки при создании структуры. Например, на рис. 30 представлен процесс моделирования синтеза структуры “ковер Серпинского".

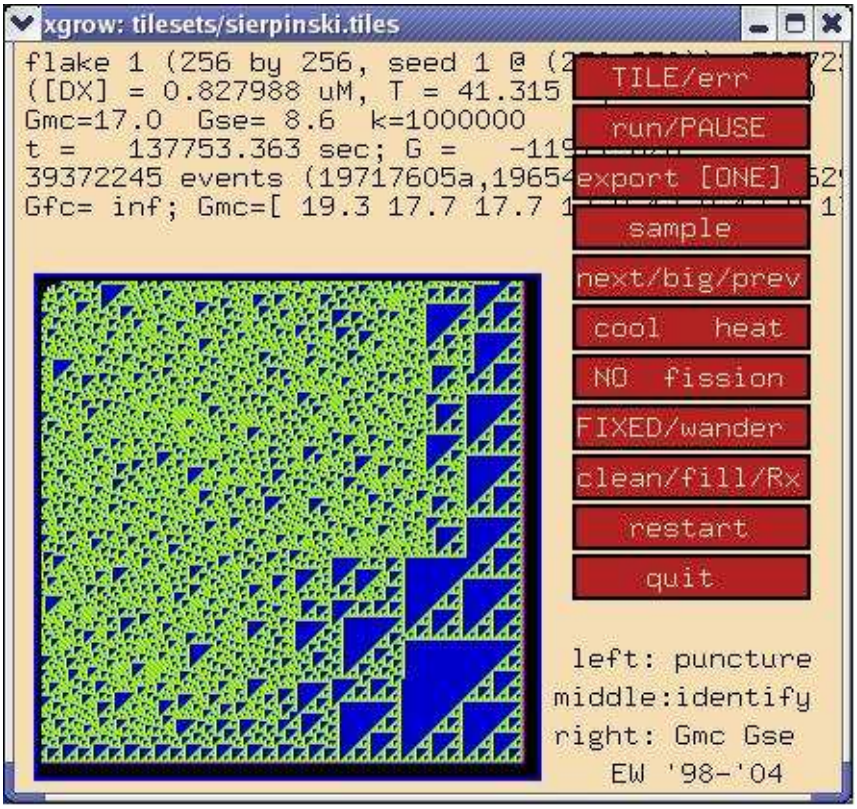

Рис. 30.

Namot. Система Namot была разработана в 1994-1995 годах в Лос-Аламосской лаборатории США. Namot расшифровывается как Nucleic Acid MOdeling Tool. Она представляет собой графическое средство работы с молекулярными структурами. С ее помощью можно составлять структуры из атомов, задавать связи в трехмерном пространстве, строить последовательности молекулярных операций. Внешний вид программы с собранной молекулярной структурой показан на рис. 31.

6. Заключение

Мы рассмотрели общую схему ДНК-вычислений, основные используемые объекты, их свойства, и операции, которые мы умеем производить. Подобной детализации вполне достаточно для моделирования несложных экспериментов.

Но для того чтобы научиться решать более сложные практические задачи на ДНК-вычислителе, необходимо ответить на многочисленные вопросы. Во-первых, пока не понятно, какой класс задач вообще удастся решить. Во-вторых, даже если мы определим этот класс точно, нам необходимо построить общие методы преобразования задачи в термины ДНК-операций, иначе к каждой задаче придется применять эвристический подход. Именно поэтому значительные усилия прилагаются к созданию общей формализованной модели ДНК-вычислений, пригодной как для реализации, так и для использования.

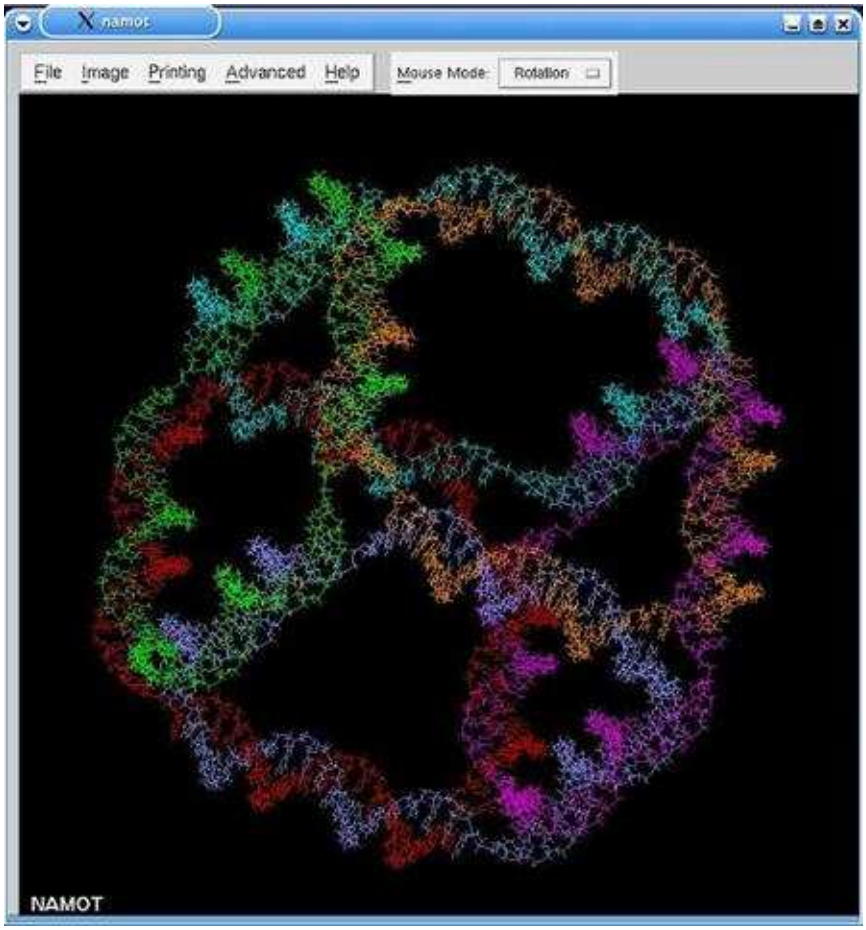

Рис. 31.

В частности, хотелось бы научиться использовать локальные взаимодействия для получения глобальной структуры. И по этому аспекту уже есть многочисленные теоретические наработки. Но, хотя и доказано, что клеточный автомат является Тьюринг-полной моделью вычислений, общей схемы перевода программ для машины Тьюринга в программу для клеточного автомата пока не имеется.

Дальнейшее развитие области ДНК-вычислений требует значительных междисциплинарных усилий. Наработки специалистов по теории вычислений и математическому моделированию позволят рассматривать более сложную модель молекулярных взаимодействий, приближенную к реальности. Необходимо также участие специалистов по молекулярной биологии, которые смогут ответить на вопросы принципиальной осуществимости тех или иных идей сборки. А специалисты - нанотехнологи помогут ответить на вопрос, какие структуры и объекты нужно синтезировать, и какие структуры могут быть синтезированы при текущем уровне развитии технологий.

Состоится ли область ДНК-вычислений? На настоящий момент ответить на этот вопрос с уверенностью нельзя. Эксперименты подтверждают, что некоторых полезных результатов, при дальнейших исследованиях, достичь мы все-таки сможем. В ближайшем будущем, скорее всего, удастся использовать ДНК-вычислители для синтеза определенных типов лекарств. Возможно, мы даже научимся решать некоторые из тех вычислительных проблем, с которыми обычный компьютер справиться не может, в частности, задачи криптоанализа. Но все же, с большой вероятностью ДНК-вычислители никогда не смогут вытеснить

обыкновенные компьютеры на основе кремниевых чипов.

Список литературы

[1] Малинецкий Г.Г., Науменко С.А. Вычисления на ДНК.

[2] Adleman L.M., Molecular Computation of Solutions to Combinatorial Problems.

[3] Istvan Katsanyi. Molecular Computing Solutions of some Classical Problems.

[4] Robin Varghese. Implementing models of DNA computing.