УДК 577.323:519.852

РЕСТРИКЦИОННЫЕ ГРАФЫ И ФИЗИЧЕСКОЕ КАРТИРОВАНИЕ МОЛЕКУЛ ДНК *

П. А. Певзнер

Введение. Физическое картирование — один из самых распространенных методов анализа ДНК, применяющийся при клонировании, определении первичной последовательности ДНК, выявлении эволюционной близости геномов, определении специфичности рестриктаз и т. п. Математические аспекты задачи физического картирования интенсивно изучаются с середины 70-х годов [1, 2]. В последнее время появились методы и программы для ЭВМ [3—7], позволяющие проводить картирование при числе сайтов порядка 10 (по каждой рестриктазе), однако задачи построения более подробных физических карт и оптимального планирования биохимических экспериментов при картировании еще ждут своего решения.

Большинство алгоритмов физического картирования основано на последовательном выдвижении, анализе и отбраковке гипотез о том или ином расположении сайтов рестрикции. В связи с огромным числом взаимного расположения сайтов (и, как следствие, необходимостью проверки большого числа гипотез) при реализации таких алгоритмов возникают значительные трудности, и на первый план выдвигаются следующие задачи: 1) оптимальная организация перебора гипотез о взаимном расположении сайтов; 2) эффективная проверка и отбраковка гипотез о взаимном расположении сайтов. Первая задача для случая парной физической карты уже перешла в разряд классических, и для ее решения предложен целый ряд алгоритмов, в частности метод ветвей и границ [8]. Значительно более сложной становится первая задача при построении множественных физических карт (этот случай гораздо чаще встречается на практике); для ее решения предложен метод потенциалов [9].

В настоящей работе исследуется вторая задача [10]; другое ее название — уточнение физических карт, так как проверка гипотез сводится, как правило, к уточнению положения сайтов рестрикции относительно некоторой точки на карте — начала отсчета. Для ее решения вводится понятие рестрикционного графа и строится взаимно-однозначное соответствие между физическими картами, удовлетворяющими принятой гипотезе о взаимном расположении сайтов, и циркуляциями в рестрикционном графе (в работе используется терминология, принятая в теории графов, определения даны в [11]). В результате удается свести задачу уточнения физической карты к задаче поиска циркуляции в сети с двухсторонними ограничениями. Это позволяет перенести основные проблемы в область построения потоковых алгоритмов и, используя результаты теории графов, получить как исчерпывающее теоретическое (теорема Гофмана), так и алгоритмическое (алгоритм построения максимального потока в сети) решение проблемы уточнения физических карт.

Постановка задачи. Исходной информацией для построения физических карт служили данные электрофореграмм о длине фрагментов одиночных и совместных рестрикций (в соответствии с общепринятой терминологией фрагменты одиночных рестрикций называются SD-фрагментами, а совместных — DD-фрагментами). В дальнейшем мы считаем, что зафиксирована гипотеза о порядке расположения сайтов рестрикций и следовании фрагментов на молекуле ДНК. Каждой такой гипотезе соответствует диаграмма D (схема 1), на которой указаны порядок следования сайтов и соответствующие длины SD- и DD-фрагментов. Физическая карта M (схема 1) дает информацию о порядке

* Представлена членом редколлегии М. Д. Франк-Каменецким.

следования сайтов и расстоянии каждого сайта от начальной точки. Всякая физическая карта M порождает некоторую диаграмму D_M, в качестве длины фрагментов на этой диаграмме используются расстояния между соответствующими сайтами. Однако обратное утверждение неверно: не по всякой диаграмме можно построить физическую карту (например, по диаграмме D нельзя построить физической карты, так как уже для первых двух фрагментов рестрикции B: $8+15\neq24$). Поэтому возникает необходимость в построении карты, наилучшим образом приближающей диаграмму,— это и есть задача уточнения физических карт.

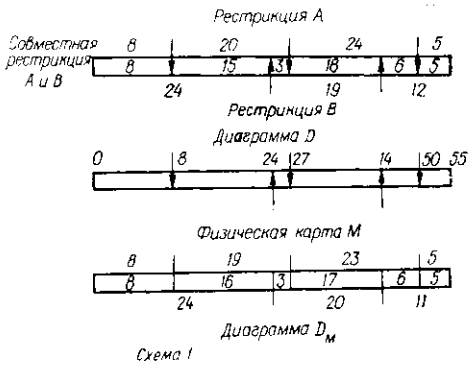


Схема 1

В качестве меры расхождения между диаграммой и физической картой рассматривается равномерная норма, т. е. если длины фрагментов диаграммы представлены вектором $D=(d_1, \ldots, d_t)$, а длины фрагментов физической карты — вектором $M=(m_1, \ldots, m_t)$, то под отклонением физической карты M от диаграммы D понимается

$$\max_{1\leqslant i\leqslant t} |d_i - m_i|$$

(в векторах D и M представлены в некотором порядке длины фрагментов SD- и DD-рестрикций). Таким образом, задача уточнения физической карты состоит в отыскании по диаграмме D физической карты M^*, дающей решение следующей минимальной задачи:

$$\max_{1\leqslant i\leqslant t} |d_i - m_i^*| = \min_{M} \max_{1\leqslant i\leqslant t} |d_i - m_i|.$$

Физическая карта M^* называется оптимальным представлением диаграммы D (можно показать, что на схеме 1 карта M — оптимальное представление диаграммы D).

Нам представляется, что выбор равномерной нормы в большей степени соответствует реальной задаче картирования, чем, скажем, выбор евклидовой нормы l^2:

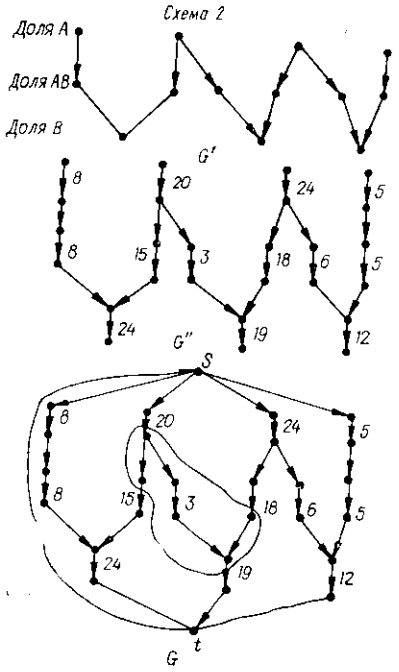
$$\left(\sum_{i=1}^{t} (d_i - m_i)^2\right)^{1/2}.$$

Дело в том, что незначительные отклонения от данных электрофореграмм вполне допустимы (они постоянно встречаются при картировании), а вот значительные отклонения хотя бы на одном фрагменте недопустимы и должны приводить к отбраковке гипотезы о порядке расположения сайтов.

Рестрикционные графы. Всякой диаграмме можно поставить в соответствие рестрикционный граф с пропускными способностями дуг (сеть). Построение графа проводили в три этапа (на схеме 2 показан процесс построения рестрикционного графа, соответствующего диаграмме на схеме 1).

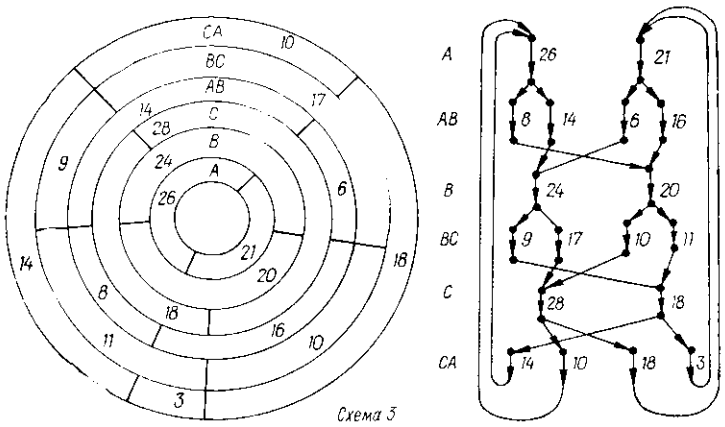
1. Строили трехдольный ориентированный граф G' (схема 2) с долями: A — множество SD-фрагментов рестрикции A; B — множество

SD-фрагментов рестрикции B; AB — множество DD-фрагментов. Из вершины v верхней доли A проводили дугу (v, w) в вершину w средней доли AB, если SD-фрагмент v содержит DD-фрагмент w. Аналогично, в вершину v нижней доли B проводили дугу (w, v) из вершины w средней доли AB, если SD-фрагмент v содержит DD-фрагмент w.



2. Граф G' преобразовывали в граф G'' (схема 2), заменив каждую вершину дугой, пропускная способность которой равна длине соответствующего фрагмента.

3. Для завершения формирования рестрикционного графа G добавляли к графу G'' (схема 2) две вершины — источник s и сток t — и соединяли источник s с началами всех дуг, соответствующих SD-фрагментам A, а все концы дуг, соответствующих SD-фрагментам B, соединяли со стоком t. После этого добавляли в граф последнюю дугу (t, s). Описанный процесс позволяет строить рестрикционные графы по диаграммам как линейных, так и кольцевых молекул; исходя из него, можно обобщить понятие рестрикционного графа для случая трех рестриктаз, учета информации о недорестрикциях и т. д. (на схеме 3 приведен пример построения рестрикционного графа кольцевой молекулы для трех рестриктаз).



Физические карты и циркуляция в рестрикционных графах. Каждому DD-фрагменту физической карты, согласующейся с нашей гипотезой о порядке следования сайтов рестрикции, соответствует некоторая дуга в графе G. Можно показать, что через эту дугу в графе G проходит ровно один ориентированный цикл. Взвешенная сумма таких циклов (в качестве веса цикла используется длина соответствующего фрагмента) дает циркуляцию в G. Таким образом, каждой физической карте, согласующейся с принятой гипотезой о порядке следования сайтов, можно поставить в соответствие циркуляцию в G. Наоборот, если дана циркуляция в G, то она однозначно определяется значениями на дугах, соответствующих DD-фрагментам (это следует из того, что всякий ориентированный цикл в G проходит через некоторую DD-дугу и через каждую такую DD-дугу проходит ровно один цикл). По значениям на DD-дугах, в свою очередь, однозначно строится соответствующая физическая карта (координата i-го сайта рестрикции на ней равна $\sum_{j=1}^{i} f_j$, где f_j — поток по j-й DD-дуге графа G). Таким образом, установлено взаимно-однозначное соответствие между физическими картами и циркуляциями в G (заметьте, что при таком подходе поток по дуге (t, s) определяет общую длину молекулы ДНК). Так как физическим картам соответствуют циркуляции, для отыскания оптимального представления диаграммы физической картой необходимо найти циркуляцию в графе G, наилучшим образом приближающую пропускные способности дуг, т. е. циркуляцию, минимизирующую

$$\max_{(v,w)} |d(v, w) - f(v, w)|,$$

где $d(v, w)$ — пропускная способность дуги (v, w), а $f(v, w)$ — поток по дуге (v, w) (максимум в формуле берется по всем дугам, на которых определена пропускная способность).

Для решения этой задачи введем нижние и верхние пропускные способности на дугах по правилу

$$d^-(v, w) = d(v, w) - \varepsilon; \qquad d^+(v, w) = d(v, w) + \varepsilon$$

(для дуг (v, w), на которых ранее пропускная способность не определялась, положим $d^-(v, w) = 0$; $d^+(v, w) = \infty$). В этом случае существование физической карты, отклонение которой от диаграммы не превышает ε, эквивалентно существованию циркуляции в сети с нижними и верхними пропускными способностями d^- и d^+.

Для циркуляций в сети с двухсторонними ограничениями пропускной способности верна теорема Гофмана.

Теорема Гофмана. Циркуляция в графе G с двухсторонними ограничениями пропускной способности существует тогда и только тогда, когда для любого подмножества X вершин графа выполняется

$$d^+(X, \bar{X}) \geqslant d^-(\bar{X}, X)$$

(здесь $d^+(X, \bar{X})$ — суммарная верхняя пропускная способность прямых ребер разреза $(X, \bar{X})$, а $d^-(\bar{X}, X)$ — суммарная нижняя пропускная способность обратных ребер разреза $(\bar{X}, X)$).

Определим дефицит сети, как $DF = \max_X (d^-(\bar{X}, X) - d^+(X, \bar{X}))$. Если в сети с двухсторонними ограничениями пропускной способности не существует циркуляции, то по теореме Гофмана $DF > 0$. Как уже отмечалось, в реальной ситуации, как правило, не существует физической карты, в точности соответствующей диаграмме, поэтому при $\varepsilon = 0$ (в этом случае функции d^- и d^+ совпадают и равны d) найдется множество X, такое что

$$d(\bar{X}, X) > d(X, \bar{X})$$

и дефицит сети $DF>0$. Значение DF может быть определено с помощью алгоритма вычисления максимального потока в некоторой сети специального вида [11]. Множество X, при котором достигается дефицит сети, позволяет локализовать фрагменты, на которых принятая гипотеза дает «максимальные» противоречия (на схеме 2 для графа G выделено множество X из шести вершин, дающее максимальный дефицит: сумма пропускных способностей по его обратным дугам равна 20+18, а по прямым — 15+19. Следовательно, $DF=38-34=4$). При возрастании ε $d^-(\bar{X}, X)$ уменьшается, а $d^+(X, \bar{X})$ увеличивается для любого множества X. Таким образом, дефицит сети $d^-(X, \bar{X}) - d^+(X, \bar{X})$ — убывающая функция $DF(\varepsilon)$. Можно показать, кроме того, что $DF(\varepsilon)$ — кусочно-линейная выпуклая вверх функция, число звеньев которой оценивается как $0(r)$, где r — число сайтов рестрикции. Как уже отмечалось, $DF(0)>0$ для реальных диаграмм. Обозначим через ε_0 точку, в которой $DF(\varepsilon)$ обращается в 0. При таком ε_0, по теореме Гофмана, впервые появляется возможность построить циркуляцию в сети с двухсторонними ограничениями, и, следовательно, ε_0 дает минимальное отклонение исходной диаграммы от физической карты, а соответствующая циркуляция порождает физическую карту, являющуюся оптимальным представлением диаграммы.

В результате задача уточнения физической карты сводится к нахождению ε_0 и построению циркуляции в соответствующей сети. Заметим, что функция $DF(\varepsilon)$ неизвестна, однако ее значение в любой точке можно вычислить решением задачи о максимальном потоке в сети. Проблема нахождения точки пересечения таких функций с прямой (нас интересует пересечение графика функции $DF(\varepsilon)$ с прямой $DF=0$) рассматривалась в работе [12], и для ее решения был предложен алгоритм, число этапов которого не превышает числа звеньев кусочно-линейной функции.

Таким образом, решение задачи уточнения физических карт свелось к серии итераций, на каждой из которых находится максимальный поток в сети. Так как для поиска максимального потока в сети известны эффективные полиномиальные алгоритмы [13], предложенный метод позволяет проводить уточнение физических карт практически при любом числе фрагментов.

GRAPHS OF RESTRICTIONS AND DNA PHYSICAL MAPPING

P. A. Pevzner

Research Institute for Genetics and Breeding of Commercial Microorganisms, Moscow

Summary

DNA physical mapping concluded from the single and double restrictions analysis leads to a great variety of hypotheses about order of the sites. The concept of graph of restrictions was introduced for examination and selection of such hypotheses. It allows applying methods of discrete optimization for physical mapping and solving the major problems by maximal flow-minimum cut algorithms. This approach throws away maps with significant deviations from experimental data (such deviations on individual fragments are allowed in the Schroeder-Blattner method).

1. *Stefic M.* Inferring DNA structure from segmentation data // Artif. Intel.— 1978.— **11**, N 1.— P. 85—115.
2. *Экспертные* системы / М. Стефик, Я. Эйкинс, Р. Балзер и др. // Киберн. сб.— М.: Мир, 1985.- Вып. 22.— С. 170—220.
3. *Pearson W.* Automatic construction of restriction site maps // Nucl. Acids Res.— 1982.—**10**, N 1.— P. 217—227.
4. *Polner G., Dorgai L., Orosz L.* PMAP, PMAPS: DNA physical map constructing programs // Ibid.— 1984.—**12**, N 1.— P. 227--236.
5. *Durand R., Bregegere F.* An efficient program to construct restrictions maps from experimental data with realistic error level // Ibid.— P. 703—716.

6. *Nolan G., Maina C., Szalay A.* Plasmid mapping computer program // Ibid.— P. 717—729.
7. *Fitch W., Smith T., Ralph W.* Mapping the order of DNA restriction fragments // Gene.— 1983.—22, N 1.— P. 19—29.
8. *Певзнер П. А., Миронов А. А.* Применение метода ветвей и границ для решения задач физического (рестрикционного) картирования // Генетика и биохимия микроорганизмов-биотехнологии: Тез. сообщ. конф.— М., 1986.— С. 80.
9. *Певзнер П. А., Миронов А. А.* Эффективный метод физического картирования молекул ДНК // Молекуляр. биология.— 1987.—**21**, № 3.— С. 788—796.
10. *Schroeder J. L., Blattner F. R.* Least-squares method for restriction mapping // Gene.— 1978.—**4**, N 2.— P. 167—174.
11. *Форд Л., Фалкерсон Д.* Потоки в сетях.— М. : Мир, 1966.—266 с.
12. *Певзнер П. А.* Эффективный алгоритм упаковки ветвлений во взвешенном графе // Комбинаторные методы в потоковых задачах.— М., 1979.— С. 91—104.
13. *Адельсон-Вельский Г. М., Диниц Е. А., Карзанов А. В.* Потоковые алгоритмы.— М. : Наука, 1975.—119 с.

ВНИИ генетики и селекции пром. микроорганизмов, Москва

Получено 04.02.87

УДК 577.150.6

ВОЗМОЖНОЕ КОДИРОВАНИЕ ЖЕЛЕЗО-СЕРНЫХ БЕЛКОВ В МИТОХОНДРИАЛЬНОМ ГЕНОМЕ МЛЕКОПИТАЮЩИХ *

Н. Н. Береговская, А. В. Савич

Железо-серные белки (ЖСБ), согласно существующей номенклатуре [1], подразделяются на простые и сложные. Среди сложных известны железо-серные флавопротеины, молибдено-флавопротеины и др. Простые содержат только железо-серные функциональные группы; к ним относятся: рубредоксины, принадлежащие анаэробным и сульфатредуцирующим бактериям, у которых атом Fe координационно связан с четырьмя атомами S от цистеиновых остатков белка и не содержится свободных атомов серы; ферредоксины (ФДК), содержащие железо-серные кластеры с одинаковым количеством атомов железа и свободной серы типа 2Fe-2S, 3Fe-3S, 4Fe-4S, в которых лигандами железа служат еще аминокислотные остатки белка — чаще всего цистеиновые; ФДК типа 8Fe-8S (кластеры 4+4) имеются у анаэробных и фотосинтезирующих бактерий; типа 7Fe-7S (3+4) найдены у анаэробной азотфиксирующей бактерии; 4Fe-4S — у анаэробных, сульфатредуцирующих и фотосинтезирующих бактерий; 2Fe-2S — у бактерий и в хлоропластах растений и водорослей [2].

Высокопотенциальные ЖСБ типа 4Fe-4S найдены у фотосинтезирующих пурпурных бактерий и в митохондриях высших животных [3, 4]. ЖСБ (простые и сложные) являются обязательными и наиболее многочисленными компонентами систем электронного транспорта. Они входят в комплексы НАДН-дегидрогеназы, сукцинатдегидрогеназы, комплекс цитохромов b — c_1, а также в цепь β-окисления ненасыщенных жирных кислот. Этим белкам приписывается непосредственное участие в сопряжении дыхания и фосфорилирования [3—5].

Для многих ЖСБ из бактерий и хлоропластов определена первичная структура, но ни одна из аминокислотных последовательностей митохондриальных ЖСБ неизвестна, что связано с трудностью их выделения.

Нуклеотидная последовательность митохондриального генома человека расшифрована полностью [6]. В нем имеются участки, кодирующие рибосомальные и транспортные РНК, а также информацион-

* Представлена членом редколлегии В. И. Ивановым.