УДК 519.23

РЕКОНСТРУКЦИЯ ЦИКЛИЧЕСКИХ ГЕНОМНЫХ СТРУКТУР С ПОМОЩЬЮ SAT-РЕШАТЕЛЯ

RECONSTRUCTION OF CYCLIC GENOME STRUCTURES BY A SAT-SOLVER

Горбунов К.Ю. [1, 2]
Gorbunov K.Yu. [1, 2]

[1] РГУ им. А.Н. Косыгина (Технологии. Дизайн. Искусство), Москва
[1] The Kosygin State University of Russia
[2] Институт проблем передачи информации им. А.А. Харкевича РАН
[2] Institute for Information Transmission Problems (Kharkevich Institute)

ВВЕДЕНИЕ

Циклической геномной структурой называется любой ориентированный граф, компоненты которого – циклы (петель нет), а рёбрам приписаны имена. Такой граф отражает строение генома: рёбра соответствуют генам, а компоненты – кольцевым хромосомам. Операция *двойной переклейки* (DM-операция), применяемая к таким структурам, состоит в разрезании двух вершин в графе и отождествлении (склеивании) четырёх образовавшихся краёв по-новому. Подробности, касающиеся геномных структур и операций над ними, приведены ниже, а также в [1].

Задача кратчайшей *реконструкции* циклических структур, заданных в листьях дерева состоит в поиске такого продолжения структур с листьев на внутренние вершины дерева, чтобы суммарное число DM-операций, требующихся для преобразования структур по всем рёбрам дерева, было минимальным. Эта задача является NP-трудной даже для простейшего дерева типа «звезда». В [2] описан алгоритм решения задачи реконструкции, основанный на её сведении к задаче целочисленного линейного программирования. Здесь опишем другой путь решения для дерева-звезды, основанный на сведении задачи к последовательности задач выполнимости конъюнктивной нормальной формы (КНФ).

ПОСТАНОВКА ЗАДАЧИ

Геномная структура – ориентированный граф, состоящий из циклов (включая петли) и цепей. Его рёбра занумерованы, мы рассматриваем случай, когда номера не повторяются их множество одно и то же для всех структур. Известны четыре основные операции, применяемые для преобразования таких структур:

1) *Двойная переклейка* (DM). Разрезать две вершины, имеющиеся в структуре, и по-новому отождествить (склеить) четыре образовавшихся края.

2) *Полуторная переклейка* (SM). Разрезать вершину и по-новому отождествить (склеить) один образовавшийся край с каким-то свободным краем структуры.

3) и 4) *Одинарные переклейки* (OM): *разрез* и *склейка*. Разрезать вершину или склеить два свободных края. Эти операции взаимно обратны друг другу.

Например, на рис. 1 показано преобразование структуры *a* (вверху слева) в структуру *b* (внизу справа). В нём семь операций: четыре DM и три SM.

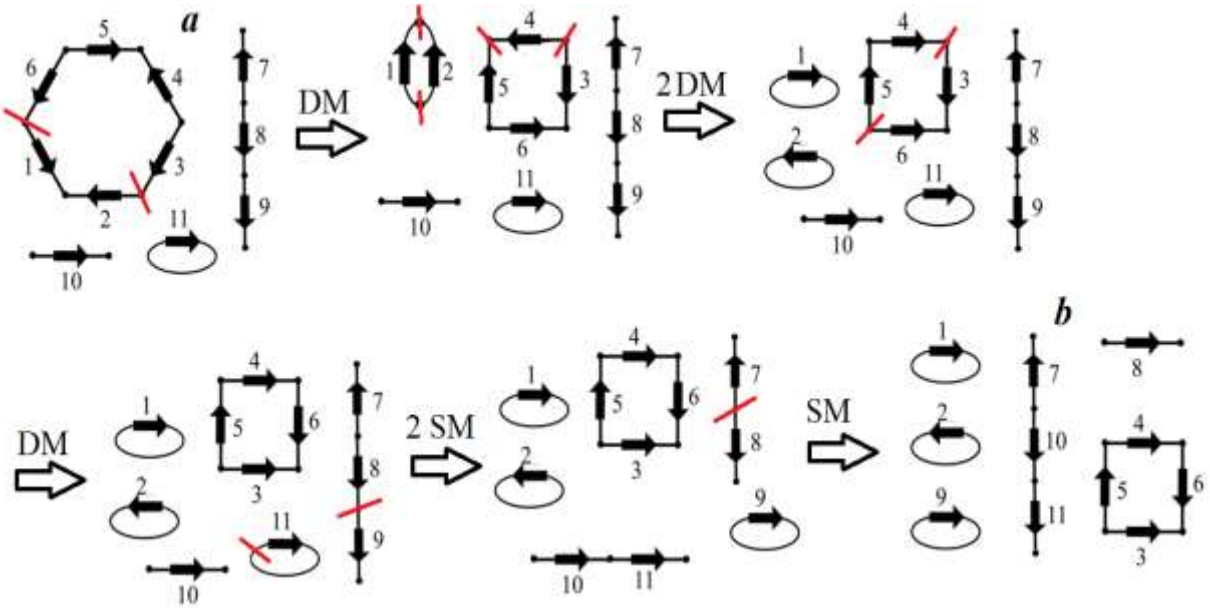


Рис. 1. Пример преобразования структуры *a* в структуру *b*

Циклическая геномная структура (далее просто ЦС) – ориентированный граф, состоящий из циклов, без петель. Как и ранее, его рёбра занумерованы, номера не повторяются и их множество одно и то же для всех рассматриваемых ЦС. Преобразование таких структур друг в друга осуществляется только DM-операциями: из четырёх упомянутых операций только DM применима к ЦС и сохраняет цикличность структуры. Таким образом, DM-*расстояние* от ЦС *a* до ЦС *b* равно минимальному числу DM-операций, которыми *a* преобразуется в *b*. Далее, рёбра орграфа будем называть генами.

Задача реконструкции циклических структур состоит в следующем. Дано корневое дерево, которое может быть небинарным. В каждом его листе задана ЦС, все структуры содержат одно и то же множество номеров генов. Требуется так расставить ЦС по внутренним вершинам дерева, чтобы сумма по всем его рёбрам DM-расстояний между ЦС на концах ребра была минимальной.

Эта задача NP-трудна даже для дерева, состоящего из корня и трёх листьев [3]. Для таких задач часто применяется либо сведение к задаче целочисленного линейного программирования, либо к задаче выполнимости конъюнктивной нормальной формы (SAT-проблема). Последний способ

предпочтительнее, поскольку в настоящее время существуют программы – SAT-решатели, которые за несколько минут проверяют выполнимость формулы с десятками тысяч переменных и сотнями тысяч клауз (т.е. дизъюнкций) и, если она выполнима, выдают выполняющий набор переменных – см., например, [4], а также интернет-сайт http://fmv.jku.at/lingeling, где указаны ссылки на соответствующие программы и публикации. Мы опишем сведение к последовательности SAT-проблем задачи реконструкции ЦС для дерева, типа «звезда», состоящего из корня r и произвольного числа листьев.

ОПИСАНИЕ АЛГОРИТМА

Для каждого фиксированного k опишем сведение к SAT-проблеме задачи о существовании такой расстановки (в данном случае, ЦС в r), что суммарное (по всем рёбрам дерева) число операций, преобразующих корневую структуру в листовые не превосходит k. Далее, взяв нижнюю и верхнюю границы на искомое (минимальное) k, мы на каждом шаге вдвое сужаем текущий отрезок его расположения, запуская SAT-решатель на SAT-задаче, соответствующей середине отрезка. Таким образом, за логарифмическое от длины начального отрезка число шагов найдём минимальное k, а затем по выполняющему набору переменных найдём саму корневую структуру.

Итак, пусть k фиксировано, n – число номеров генов, l – число листьев в дереве. Любой ген имеет два *края*: начало и конец. Заметим, что ЦС взаимно-однозначно соответствует полному паросочетанию между краями всех генов: два края составляют пару, если они склеены в одной вершине структуры. Легко видеть (см. также [1]), что минимальное число DM-операций, необходимых для преобразования одной структуры в другую, равно n минус число циклов, образованных двумя паросочетаниями этих структур на множестве краёв генов. Поэтому вместо минимизации k будем максимизировать суммарное число таких циклов по всем рёбрам дерева, равное $nl–k$. Таким образом, описываем КНФ, которая выполнима тогда и только тогда, когда существует такое паросочетание на множестве краёв в r, что упомянутое число циклов не меньше заданного c. Будем представлять циклы занумерованными, начиная от 1, их номера назовём *цикловыми*. Для каждой неупорядоченной пары p различных краёв введём переменную, которую будем также обозначать как p (смысл: p=1, если p склеена в r и 0 иначе). Для каждой склеенной пары p в листе и натурального числа $m \le c$ введём переменную p_m (смысл: p_m=1, если p лежит в цикле с номером m и 0 иначе). Выразим сначала, что каждый край i_j в r склеен ровно с одним краем. Это, во-первых, дизъюнкция всех p, содержащих i_j, а во-вторых для каждых двух различных краёв, отличных от i_j, дизъюнкция отрицаний склеек края i_j с этими краями. Аналогичным образом выразим, что каждая склеенная пара в листе имеет ровно один цикловой номер. Так же выразим, что для любого циклового номера m в каком-нибудь листе существует склеенная пара с этим номером – дизъюнкция переменных p_m по всем p.

Теперь обеспечим совпадение цикловых номеров всех пар p в каждом цикле. Достаточно обеспечить совпадение номеров для соседних пар, т.е. отделённых ровно одной склейкой в r. Иначе говоря, обеспечить импликацию: если две склеенных пары в одном листе соединены склейкой в r (любой из четырёх возможных), то они имеют одинаковый цикловой номер. Соответствующее множество дизъюнкций: для любых склеенных пар s и t в одном листе, любой из четырёх пар p между ними и любого циклового номера m выполняется $(\neg p \vee \neg s_m \vee t_m)(\neg p \vee \neg t_m \vee s_m)$.

Может показаться, что следует также обеспечить, чтобы в различных циклах склеенные пары имели различные цикловые номера. Однако этого не требуется: если разные циклы будут иметь одинаковые номера, количество циклов будет больше числа использованных номеров, что соответствует требованию на расстановку.

Из описанного построения и тривиальных границ на k (0 и nl) сразу вытекает следующая теорема.

Теорема. *Задача реконструкции на дереве типа «звезда» в циклическом случае при равном генном составе структур сводится к последовательности из* $\log(nl)$ *SAT-проблем. Каждая из них содержит порядка* n^2l^2 *переменных и порядка* n^3l^2 *клауз.*□

Уточнить нижнюю границу на k можно, построив *общий граф* всех листовых структур. Это граф, в котором вершины – края всех генов, а ребро соединяет два края, если они склеены в некоторой листовой структуре, это ребро помечено номером соответствующего листа. Таким образом, две вершины могут соединяться несколькими (вплоть до l) рёбрами, а степень любой вершины графа равна l. Общий граф имеет финальный вид, если каждая связная компонента – две вершины, соединённых l рёбрами. Легко видеть, что задача реконструкции эквивалентна задаче приведения общего графа к финальному виду двойными переклейками, т.е. заменами двух одноимённых рёбер двумя другими рёбрами с теми же четырьмя краями и именем. Чередующимся циклом назовём цикл, каждое ребро которого помечено одним из двух номеров и эти пометки чередуются. В графе финального вида $nl(l–1)/2$ чередующихся циклов, в произвольном общем графе их не больше. Двойная переклейка может увеличить число чередующихся циклов не более, чем на $l–1$. Отсюда получаем нижнюю оценку на k, равную $\frac{nl}{2} - \frac{c}{l-1}$, где c – число чередующихся циклов в общем графе. Верхнюю границу на k можно получить, перебрав расстановки, где в r повторяется структура из любого листа, и вычислив минимальное по этим расстановкам число DM-операций.

Часто разумно ограничиться сокращённым вариантом алгоритма, где корневая структура ищется среди таких структур, в которых каждая склеенная пара краёв присутствует хотя бы в одном листе. В большинстве реальных примеров оптимальная структура обладает этим свойством, тогда как число переменных и клауз при сокращённом варианте существенно уменьшается.

ПРИМЕР РАБОТЫ АЛГОРИТМА

В качестве примера описанного алгоритма рассмотрим задачу реконструкции, исходные данные которой изображены на рис. 2.

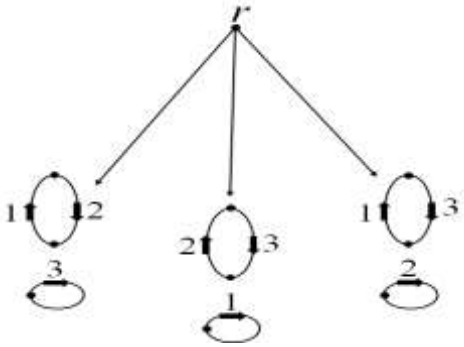


Рис. 2. Исходные данные задачи реконструкции ЦС для сведения её к последовательности SAT-проблем

Применим сокращённый вариант алгоритма. Переменные p будем индексировать тремя индексами – двумя нижними и одним верхним. Первый нижний индекс – символ l, m или r указывает, соответственно, на левый, средний или правый лист. Второй нижний индекс – символ u, m или l указывает, соответственно, на верхнюю, среднюю или нижнюю склейку в этом листе. Верхний индекс – символ r или число указывает, соответственно, на корень дерева или номер цикла.

Часть формулы, не зависящая от k:

$$(p^r_{lm} \vee p^r_{ml} \vee p^r_{rm})(p^r_{lu} \vee p^r_{mr} \vee p^r_{ru})(p^r_{lu} \vee p^r_{mm} \vee p^r_{rl})(p^r_{lm} \vee p^r_{mu} \vee p^r_{rl})(p^r_{ll} \vee p^r_{mu} \vee p^r_{ru})(p^r_{ll} \vee p^r_{mm} \vee p^r_{rm})$$
$$(\neg p^r_{lm} \vee \neg p^r_{ml})(\neg p^r_{lm} \vee \neg p^r_{rm})(\neg p^r_{ml} \vee \neg p^r_{rm})(\neg p^r_{lu} \vee \neg p^r_{mr})(\neg p^r_{lu} \vee \neg p^r_{ru})(\neg p^r_{mr} \vee \neg p^r_{ru})$$
$$(\neg p^r_{lu} \vee \neg p^r_{mm})(\neg p^r_{lu} \vee \neg p^r_{rl})(\neg p^r_{mm} \vee \neg p^r_{rl})(\neg p^r_{lm} \vee \neg p^r_{mu})(\neg p^r_{lm} \vee \neg p^r_{rl})(\neg p^r_{mu} \vee \neg p^r_{rl})$$
$$(\neg p^r_{ll} \vee \neg p^r_{mu})(\neg p^r_{ll} \vee \neg p^r_{ru})(\neg p^r_{mu} \vee \neg p^r_{ru})(\neg p^r_{ll} \vee \neg p^r_{mm})(\neg p^r_{ll} \vee \neg p^r_{rm})(\neg p^r_{mm} \vee \neg p^r_{rm}).$$

Пусть k=2, тогда c=7. Добавим цикловой номер для пары lu:

$$(p^1_{lu} \vee p^2_{lu} \vee p^3_{lu} \vee p^4_{lu} \vee p^5_{lu} \vee p^6_{lu} \vee p^7_{lu})$$
$$(\neg p^1_{lu} \vee \neg p^2_{lu})(\neg p^1_{lu} \vee \neg p^3_{lu})(\neg p^1_{lu} \vee \neg p^4_{lu})(\neg p^1_{lu} \vee \neg p^5_{lu})(\neg p^1_{lu} \vee \neg p^6_{lu})(\neg p^1_{lu} \vee \neg p^7_{lu})$$
$$(\neg p^2_{lu} \vee \neg p^3_{lu})(\neg p^2_{lu} \vee \neg p^4_{lu})(\neg p^2_{lu} \vee \neg p^5_{lu})(\neg p^2_{lu} \vee \neg p^6_{lu})(\neg p^2_{lu} \vee \neg p^7_{lu})$$
$$(\neg p^3_{lu} \vee \neg p^4_{lu})(\neg p^3_{lu} \vee \neg p^5_{lu})(\neg p^3_{lu} \vee \neg p^6_{lu})(\neg p^3_{lu} \vee \neg p^7_{lu})$$
$$(\neg p^4_{lu} \vee \neg p^5_{lu})(\neg p^4_{lu} \vee \neg p^6_{lu})(\neg p^4_{lu} \vee \neg p^7_{lu})(\neg p^5_{lu} \vee \neg p^6_{lu})(\neg p^5_{lu} \vee \neg p^7_{lu})(\neg p^6_{lu} \vee \neg p^7_{lu})$$

и аналогично для остальных восьми пар.

Обеспечим наличие цифрового номера 1:

$$(p^1_{lu} \vee p^1_{lm} \vee p^1_{ll} \vee p^1_{mu} \vee p^1_{mm} \vee p^1_{ml} \vee p^1_{ru} \vee p^1_{rm} \vee p^1_{rl})$$

и аналогично для других шести номеров.

Наконец, обеспечим совпадение цикловых номеров всех склеенных пар в одном цикле. По номеру 1:

$$(\neg p^r_{ml} \vee \neg p^1_{lu} \vee p^1_{lm})(\neg p^r_{ml} \vee \neg p^1_{lm} \vee p^1_{lu})(\neg p^r_{rl} \vee \neg p^1_{lu} \vee p^1_{lm})(\neg p^r_{rl} \vee \neg p^1_{lm} \vee p^1_{lu})$$
$$(\neg p^r_{ll} \vee \neg p^1_{mu} \vee p^1_{mm})(\neg p^r_{ll} \vee \neg p^1_{mm} \vee p^1_{mu})(\neg p^r_{rl} \vee \neg p^1_{mu} \vee p^1_{mm})(\neg p^r_{rl} \vee \neg p^1_{mm} \vee p^1_{mu})$$
$$(\neg p^r_{ll} \vee \neg p^1_{ru} \vee p^1_{rm})(\neg p^r_{ll} \vee \neg p^1_{rm} \vee p^1_{ru})(\neg p^r_{ml} \vee \neg p^1_{ru} \vee p^1_{rm})(\neg p^r_{ml} \vee \neg p^1_{rm} \vee p^1_{ru})$$

$$(\neg p^r_{mm} \vee \neg p^1_{lu} \vee p^1_{ll})(\neg p^r_{mm} \vee \neg p^1_{ll} \vee p^1_{lu})(\neg p^r_{ru} \vee \neg p^1_{lu} \vee p^1_{ll})(\neg p^r_{ru} \vee \neg p^1_{ll} \vee p^1_{lu})$$
$$(\neg p^r_{lm} \vee \neg p^1_{mu} \vee p^1_{ml})(\neg p^r_{lm} \vee \neg p^1_{ml} \vee p^1_{mu})(\neg p^r_{ru} \vee \neg p^1_{mu} \vee p^1_{ml})(\neg p^r_{ru} \vee \neg p^1_{ml} \vee p^1_{mu})$$
$$(\neg p^r_{lu} \vee \neg p^1_{ru} \vee p^1_{rl})(\neg p^r_{lu} \vee \neg p^1_{rl} \vee p^1_{ru})(\neg p^r_{mu} \vee \neg p^1_{ru} \vee p^1_{rl})(\neg p^r_{mu} \vee \neg p^1_{rl} \vee p^1_{ru})$$
$$(\neg p^r_{mu} \vee \neg p^1_{lm} \vee p^1_{ll})(\neg p^r_{mu} \vee \neg p^1_{ll} \vee p^1_{lm})(\neg p^r_{rm} \vee \neg p^1_{lm} \vee p^1_{ll})(\neg p^r_{rm} \vee \neg p^1_{ll} \vee p^1_{lm})$$
$$(\neg p^r_{lu} \vee \neg p^1_{mm} \vee p^1_{ml})(\neg p^r_{lu} \vee \neg p^1_{ml} \vee p^1_{mm})(\neg p^r_{rm} \vee \neg p^1_{mm} \vee p^1_{ml})(\neg p^r_{rm} \vee \neg p^1_{ml} \vee p^1_{mm})$$
$$(\neg p^r_{lm} \vee \neg p^1_{rm} \vee p^1_{rl})(\neg p^r_{lm} \vee \neg p^1_{rl} \vee p^1_{rm})(\neg p^r_{mm} \vee \neg p^1_{rm} \vee p^1_{rl})(\neg p^r_{mm} \vee \neg p^1_{rl} \vee p^1_{rm})$$

и аналогично по другим шести номерам.

Полученная КНФ невыполнима, поэтому положим k=3, тогда c=6. КНФ строится точно так же, за исключением того, что цикловых номеров теперь 6. Она выполнима, выполняющий набор значений переменных (указываем только равные 1):

$$p^r_{lu},\ p^r_{mu},\ p^r_{rm},\ p^1_{lu},\ p^2_{lm},\ p^2_{ll},\ p^3_{mu},\ p^4_{mm},\ p^4_{ml},\ p^5_{rm},\ p^6_{ru},\ p^6_{rl}.$$

Первые три переменные этого набора определяют итоговую расстановку, рис. 3:

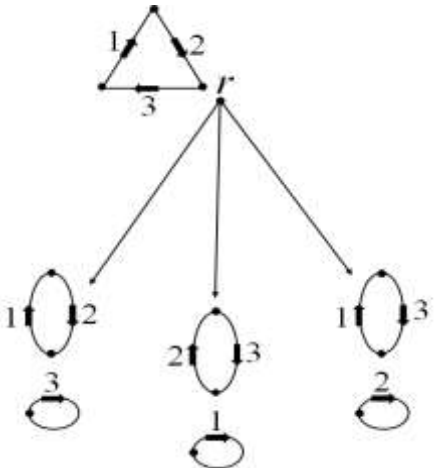


Рис. 3. Итоговая расстановка задачи реконструкции, полученная сведением её к последовательности SAT-проблем

Очевидно, корневая структура преобразуется в каждую листовую одной DM-операцией, разбивающей цикл на два. Таким образом, суммарное число DM-операций равно 3, как и ожидалось.

ЗАКЛЮЧЕНИЕ

Мы описали алгоритм решения задачи реконструкции циклических структур на дереве типа «звезда», основанный на её сведении к последовательности SAT-проблем. Интересен вопрос о возможности уменьшения размера КНФ в этих проблемах. Кроме того, дальнейшие исследования могут быть направлены на обобщение предложенного подхода на случай произвольного дерева и произвольных геномных структур.

Л И Т Е Р А Т У Р А

1. Горбунов К.Ю., Любецкий В.А. Линейный алгоритм минимальной перестройки структур // *Проблемы передачи информации.* – 2017. – Т.53(1). – С. 60-78.

2. Lyubetsky V.A., Gershgorin R.A., Gorbunov K.Yu. Chromosome structures: reduction of certain problems with unequal gene content and gene paralogs to integer linear programming // *BMC Bioinformatics.* – 2017. – V.18(537). – P.1-18.

3. Tannier E., Zheng C., Sankoff D. Multichromosomal median and halving problems under different genomic distances // BMC Bioinformatics. – 2009. – V.10(120). – P. 1-15.

4. Sohanghpurwala A.A., Hassan M.W., Athanas P. Hardware accelerated SAT solvers – a survey // J. Parallel Distrib. Comput. – 2017. – V.106. – P. 170-184.

УДК 004.32

КОНЦЕПТУАЛЬНЫЕ ПОДХОДЫ К ПРОЦЕССНО-ОРИЕНТИРОВАННЫМ ИССЛЕДОВАНИЯМ МОЗГА: ФОРМИРОВАНИЕ КОГНИТИВНЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

CONCEPTUAL APPROACHES TO PROCESS-ORIENTED BRAIN RESEARCH: FORMATION OF COGNITIVE MODELS OF MACHINE LEARNING

Смирнов О.А.
Smirnov O.A.

РГУ им. А.Н. Косыгина (Технологии. Дизайн. Искусство), Москва
The Kosygin State University of Russia, Moscow

ВВЕДЕНИЕ

Несмотря на более чем столетние исследования интеллекта, остается один из самых фундаментальных вопросов об интеллекте является проблема отличия людей по интеллекту. В исследованиях можно найти объяснение индивидуальных различий в когнитивных способностях на разных уровнях анализа.

На первом уровне мы можем идентифицировать когнитивные процессы, которые порождают мышление и поведение, и механизмы, которые обеспечивают эти процессы, и определить, как они различаются у разных людей при формировании рационального поведения.

На втором уровне можно идентифицировать нейронные механизмы, которые реализуют когнитивные механизмы и определяют, какие из их особенностей различаются у разных людей и тем самым вызывают различия в качестве обработки информации.