РАЗДЕЛ I. БИОИНФОРМАТИКА
(научный редактор раздела Горшков В.В.)

Биоинформатика и математическая биология – предметы, преподаваемые на кафедре Прикладной математики и программирования Российского государственного университета имени А.Н. Косыгина (Технологии. Дизайн. Искусство). В этом направлении преподавателями кафедра ведутся активные исследования. В настоящий раздел вошли статьи преподавателей кафедры, посвященные биоинформатике.

УДК 519.23

АЛГОРИТМ ВЫРАВНИВАНИЯ КЛЕТОЧНЫХ ДЕРЕВЬЕВ

AN ALGORITHM FOR ALIGNMENT OF CELL TREES

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

Клеточным деревом (cell lineage tree, [1]) называется корневое дерево, каждая вершина которого помечена типом соответствующей клетки (например, мышечный, нервный, и т.п.). В рассматриваемой задаче даны два бинарных клеточных дерева и матрица сходств (произвольных рациональных чисел) между любыми двумя клеточными типами, а также штраф за сопоставление вершине делеции, который может зависеть от её пометки и от того, лист ли она (примем естественное предположение, что сопоставление делеция-делеция имеет нулевую цену). Требуется выровнять эти деревья (глобально или локально), максимизировав *качество выравнивания* – суммарное сходство между соответствующими парами типов минус штрафы за делеции. В [1] описан алгоритм DELTA решения этой задачи при дополнительном условии: множество всех делеций разбито на непересекающиеся подмножества, каждое из которых помечает целое поддерево и верхний конец его корневого ребра. В частности, при этом лист одного дерева не может ставиться в соответствие внутренней вершине другого, а только листу или делеции. Мы описываем алгоритм, строящий оптимальное выравнивания без этих ограничений (назовём такое

выравнивание *обобщённым*), работающий за такое же (квадратичное от размера данных) время.

ПОСТАНОВКА ЗАДАЧИ

Как известно, при выравнивании двух слов разрешается вставлять в них делеции, т.е. фактически выравниваются слова, полученные из исходных слов последовательностью операций – вставкой делеции. Аналогично, опишем операции вставки делеции в дерево, допустимые при построении из двух исходных деревьев двух топологически изоморфных, которые и выравниваются между собой по изоморфизму. Это следующие операции (символы *x* и *y* на рисунках – типы или делеции).

1) Добавление к листу дерева двух дочерних вершин-делеций, рис. 1.

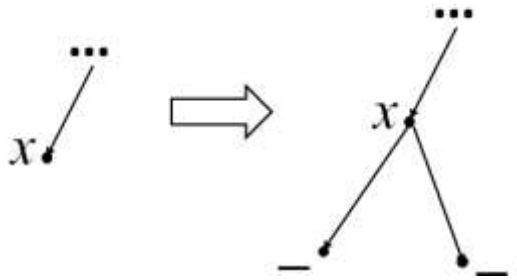


Рис. 1. Добавление к листу двух дочерних вершин-делеций

2) Добавление вершины-делеции на ребро вместе с листом-делецией, рис. 2.

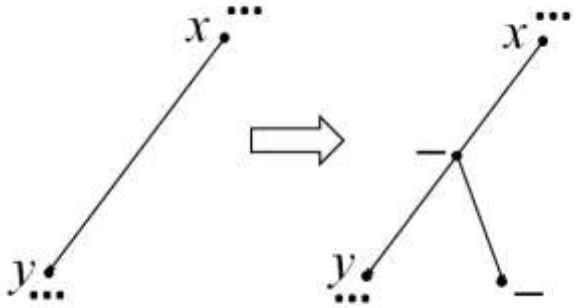


Рис. 2. Добавление вершины-делеции на ребро вместе с листом-делецией

3) Добавление нового корня-делеции вместе с листом-делецией, рис. 3.

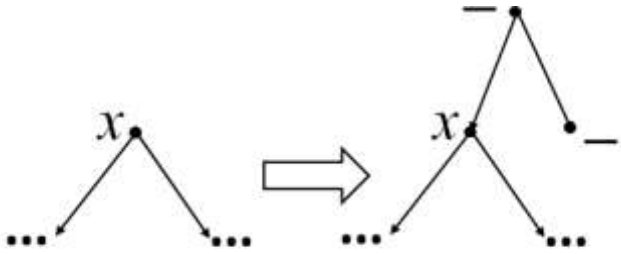


Рис. 3. Добавление нового корня-делеции вместе с листом-делецией

4) Вставка вершины-делеции в качестве корня поддерева и листа-делеции к прежнему корню, рис. 4.

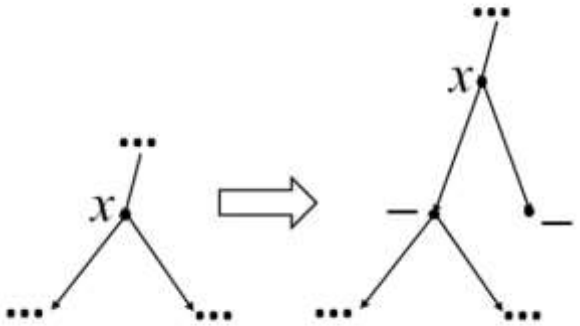


Рис. 4. Вставка вершины-делеции в качестве корня поддерева и листа-делеции к прежнему корню

Задача состоит в том, чтобы этими операциями преобразовать одно клеточное дерево так, чтобы оно стало топологически (т.е. без учёта всех пометок в деревьях) изоморфно другому клеточному дереву, и сумма (по парам соответствующих вершин) сходств типов и (отрицательных) штрафов за делеции была максимальной.

ОПИСАНИЕ АЛГОРИТМА

Пусть даны два бинарных дерева T_1 и T_2. Рассмотрим два множества D_1 и D_2 деревьев, где каждое D_i – объединение множества поддеревьев дерева T_i и множества тех же поддеревьев, в которых типовая пометка корня заменена на символ делеции. Обозначаем результат этой замены для дерева T как T^- и называем T^- *упрощением* дерева T, а T *– усложнением* дерева T^-. Скажем, что одно дерево *содержится* в другом, если первое дерево или его усложнение – нетривиальное поддерево второго.

Добавим к каждому из D_i специальный элемент – пустое дерево $T_\varnothing$, и будем считать, что каждый лист любого дерева имеет два дочерних пустых поддерева. На множестве P упорядоченных пар {непустое дерево T_1 из D_1, непустое дерево T_2 из D_2} введём частичный линейный порядок: пара T_1,T_2 меньше пары T'_1,T'_2, если T_1 содержится в T'_1 или $T_1=T'_1$ и T_2 содержится в T'_2. Алгоритм перебирает пары из P в порядке возрастания, т.е. пара рассматривается, лишь если все меньшие пары уже рассмотрены.

Базисные пары: два пустых дерева – качество нулевое, пустое дерево с непустым – качество равно минус сумме штрафов сопоставлений делеций всем вершинам непустого дерева.

Далее, для каждой пары из P алгоритм выбирает лучшее из следующих выравниваний (используется, что оптимальные выравнивания меньших пар уже известны, символы x и y на рисунках – типы или делеции).

1) Соответствие корней исходных деревьев и попарное соответствие дочерних поддеревьев (два случая), рис. 5.

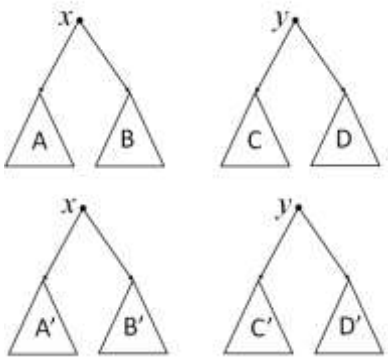


Рис. 5. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Символ со штрихом означает поддерево, в которое вставились делеции так, что либо A′ топологически изоморфно C′, а B′ – D′, либо A′ топологически изоморфно D′, а B′ – C′

2) Соответствие корней исходных деревьев и соответствие упрощённого дерева одного из исходных деревьев одному из дочерних поддеревьев другого исходного дерева и делеций – другому дочернему поддереву (4 случая), рис. 6.

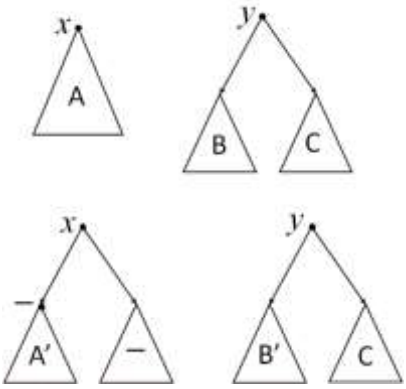


Рис. 6. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Поддерево A′ топологически изоморфно B′, делеционнее поддерево – C

3) Соответствие корней исходных деревьев и их упрощённых деревьев друг другу, рис. 7. Случай применим, если хотя бы один из символов *x*, *y* не делеция.

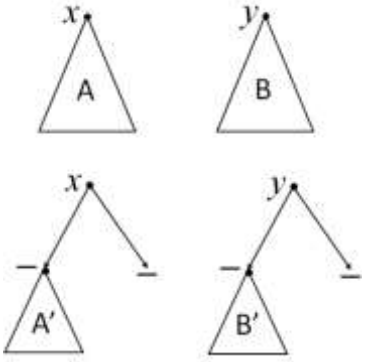


Рис. 7. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Поддеревья A′ и B′ топологически изоморфны друг другу

4) Соответствие корней исходных деревьев друг другу, всех остальных вершин – делециям, рис. 8.

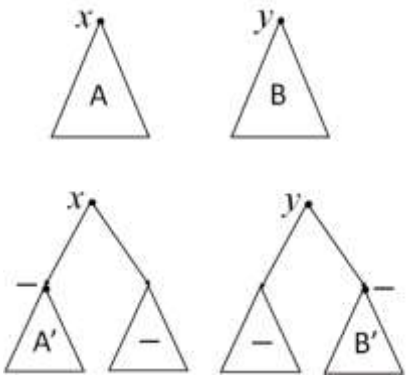


Рис. 8. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Корни соответствуют друг другу, остальные вершины – делециям

5) Соответствие корня одного из исходных деревьев делеции, одного его дочернего поддерева – другому исходному дереву, а другого дочернего поддерева – дереву из делеций (4 случая), рис. 9.

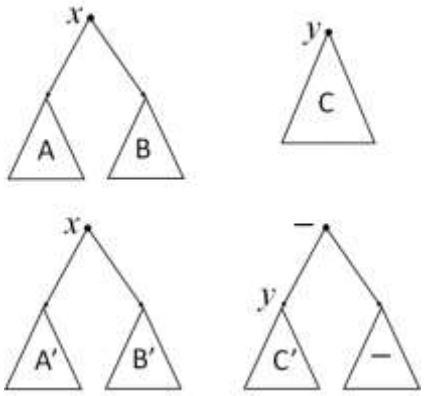


Рис. 9. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Поддерево, A′ топологически изоморфно C′, поддерево B′ – делеционному дереву

6) Соответствие корня одного из исходных деревьев делеции, его упрощённого дерева – второму исходному дереву (два случая), рис. 10. Случай применим, если упомянутый корень – не делеция.

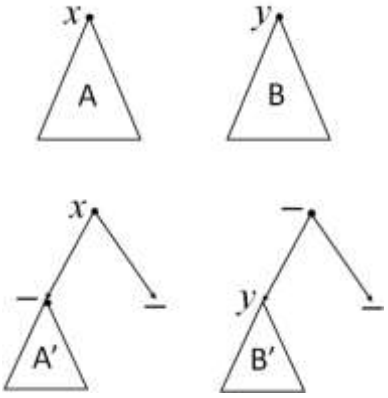


Рис. 10. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Поддеревья A′ и B′ топологически изоморфны друг другу

7) Вырожденный случай – все типовые вершины соответствуют делециям, рис. 11.

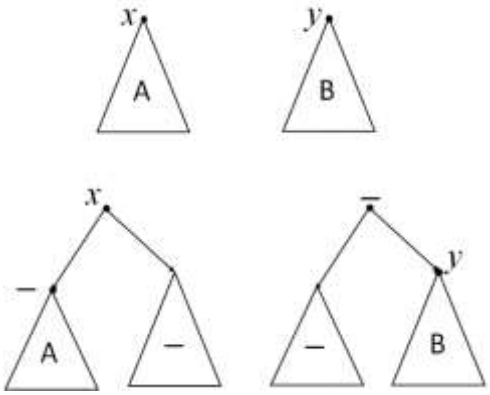


Рис. 11. Сверху: исходные деревья; снизу: те же деревья после вставки делеций. Каждая типовая вершина соответствует делеции

Из соответствия пары исходных деревьев обратным ходом получается решение задачи – глобальное обобщённое выравнивание. Как и в [1], локальные выравнивания – это соответствия поддеревьев, полученные в прямом ходе глобального выравнивания. Как и при локальном выравнивании символьных последовательностей, целесообразно отрицательные сходства поддеревьев полагать нулевыми.

ТОЧНОСТЬ АЛГОРИТМА И ПРИМЕР ЕГО РАБОТЫ

Теорема. *Алгоритм строит обобщённое выравнивание двух деревьев с максимальным качеством. Время его работы квадратично от размера исходных данных.*

Доказательство. Точность алгоритма вытекает из полного разбора случаев соответствия корней *x* и *y* исходных поддеревьев. Действительно, если *x* и *y* соответствуют друг другу, то либо обе эти вершины не подвергались операции вставки вместо них вершины-делеции (рассмотрено в пункте 1), либо ровно одна из них подвергалась ей (рассмотрено в пункте 2), либо обе (рассмотрено в пунктах 3 и 4). Пусть хотя бы одна из вершин *x*, *y* (пусть *x*) соответствует вершине-делеции, обозначим её *d*. Если *d* выше *y*, то либо *x* не подвергалась упомянутой операции (рассмотрено в пункте 5), либо подвергалась (рассмотрено в пунктах 6 и 7). Если *d* не сравнима с *y*, то все вершины ниже *x* соответствуют делеции, т.е. случай вырожденный. Если *d* ниже *y*, то *y* соответствует делеции, которая выше *x* или несравнима с ней, что уже рассмотрено. То же самое относится к случаю, когда одна из вершин *x*, *y* (пусть *x*) соответствует типовой вершине, но не *y*.

Время алгоритма квадратично от размера исходных деревьев, поскольку таково число пар в множестве *P*, а обработка каждой пары занимает константное время.

Пример. Даны два дерева *S* и *T*, изображённые на рис. 12:

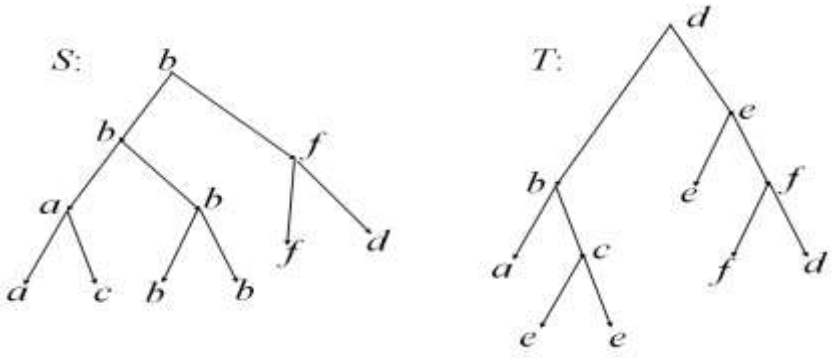


Рис. 12. Исходные деревья, к примеру работы алгоритма

Пусть цена совпадения букв равна 4, штраф за несовпадение букв равен 3, штраф за соответствие букве делеции равен 2. Обратный ход начинаем с соответствия S и T, лучший случай (пункт 1 в описании алгоритма) – соответствие корней друг другу, левых поддеревьев друг другу и, соответственно, правых. Для левых поддеревьев лучший случай (пункт 2 в описании алгоритма) – соответствие корней друг другу, поддерева в S с корнем a – упрощению поддерева в T с корнем b (там далее везде по пункту 1) и малого поддерева в S с корнем b – делециям. Для правых поддеревьев лучший случай (пункт 5 в описании алгоритма) – соответствие корня e и его левого поддерева-листа делеции, а правого поддерева – исходному правому поддереву. Итоговые топологически изоморфные деревья, полученные вставкой делеций в исходные, показаны на рис. 13:

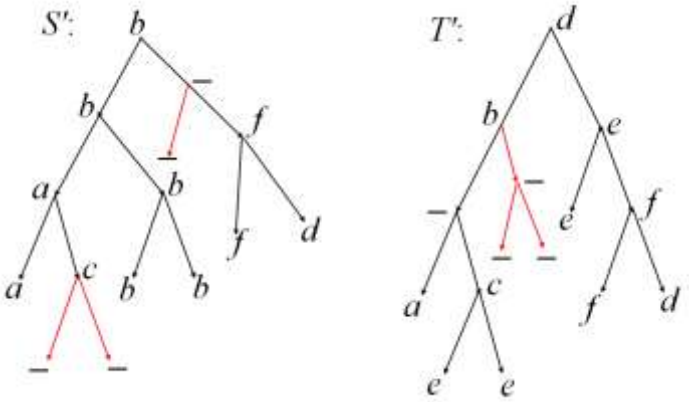


Рис. 13. Деревья с рис. 12 после вставки делеций. Красным выделены добавленные рёбра

Соответствие даёт нам 6 совпадений, одно несовпадение и 8 делеций. Таким образом, его качество равно 24–3–16=5.

Замечание. Нетрудно модифицировать алгоритм для случая политомических исходных деревьев так, что он будет выдавать их оптимальные бинарные разрешения и их выравнивание. Для пары вершин $\{x,y\}$ соответственно из S и T (пары перебираются как в бинарном случае) рассмотрим непустое множество X дочерних рёбер для x и, соответственно, Y для y. Если X или Y одноэлементно, то ограничения S и T от вершин x и y соответственно на X и Y имеют бинарные разрешения и их выравнивание с

минимальной ценой, которые известны от предыдущих пар вершин (шагов индукции). Иначе разобьём эти два множества на непустые M_1 и M_2, M_3 и M_4. Рассмотрим ограничения S и T от вершин x и y соответственно на M_1 и M_2, M_3 и M_4. Алгоритм перебирает пары множеств X, Y и их упрощений в порядке возрастания мощности X, а при одном и том же X – по возрастанию мощности Y; для фиксированных X, Y порядок (X^-,Y^-), (X^-,Y), (X,Y^-), (X,Y). Для каждой пары X и Y (остальные три пары рассматриваются аналогично) в любом порядке перебираются все пары их разбиений. Пусть X^* обозначает X или X^-. По индукции известны бинарные разрешения с максимальным качеством выравнивания для пар ограничений с корнями от x и y на, соответственно, (M_1,M_3), (M_1,M_4), (M_1,M_4), (M_2,M_3), (M_1,Y^*), (M_2,Y^*), (X^*,M_3), (X^*,M_4), (X^-,Y^*), (X^*,Y^-). Выбираем из всех разбиений и описанных в алгоритме способов выравнивания (7 пунктов) разбиение с максимальным качеством выравнивания и получим разрешение и оптимальное выравнивание для поддеревьев через X и Y. Взяв для $\{x,y\}$ наибольшие X и Y, получим бинарное разрешение с минимальной ценой выравнивания поддеревьев от x и от y, и так дойдём до пары корней исходных деревьев S и T.

Здесь перебираются четвёрки подмножеств, откуда возникает верхняя оценка 2^{4k} времени обработки одной пары вершин. Квадратичное время алгоритма для бинарного случая умножается на число порядка 2^{4k}. Таким образом, алгоритм применим при не очень больших k.

ЗАКЛЮЧЕНИЕ

Мы описали алгоритм построения оптимального выравнивания двух клеточных деревьев. Представляет интерес задача множественного выравнивания набора таких деревьев, в том числе политомических. Последняя задача связана с ещё одной сложной задачей – построить клеточное супердерево, т.е. дерево, согласующего данный набор клеточных деревьев. Дальнейшие исследования могут быть направлены на разработку эффективных алгоритмов для этих задач.

Л И Т Е Р А Т У Р А

1. Yuan M., Yang X., Lin J., Cao X., Chen F., Zhang X., Li Z., Zheng G., Wang X., Chen X. and Yang J-R. Alignment of Cell Lineage Trees Elucidates Genetic Programs for the Development and Evolution of Cell Types // iScience. – 2020. V.23(101273).