ПРОБЛЕМЫ ПЕРЕДАЧИ ИНФОРМАЦИИ

Том 47 2011 Вып. 3

УДК 621.391.1 : 514

© 2011 г. К.Ю. Горбунов, В.А. Любецкий

ДЕРЕВО, БЛИЖАЙШЕЕ В СРЕДНЕМ К ДАННОМУ НАБОРУ ДЕРЕВЬЕВ

Сформулирована задача построения дерева, ближайшего в среднем к данному набору деревьев. Понятие "ближайшее" сформулировано на основе представления о событиях, подсчет числа которых позволяет отличить каждое из данных деревьев от искомого дерева. Эти события называются дивергенцией, дупликацией, потерей, переносом; аналогично могут быть рассмотрены и другие списки событий. Предложен алгоритм, который решает эту задачу за кубическое время от размера исходных данных. Доказаны корректность алгоритма и кубическая оценка его сложности.

Хорошо известна и давно исследуется в связи с различными приложениями (например, в теории эволюции видов [1–5]) следующая задача. Дан набор деревьев G_i, где i меняется от 1 до некоторого n, нужно найти дерево S^*, *в среднем наиболее близкое* к каждому G_i. Обычно предполагается, что деревья G_i, а тогда и дерево S^*, – бинарные и корневые. Ниже (см. замечание 2) говорится, как можно перейти к этому более простому случаю от небинарных и неукорененных деревьев. После уточнения выделенных курсивом понятий задача состоит в поиске глобального минимума функционала на пространстве деревьев, ниже указан вид этих функционала и пространства. Известно, что задачи дискретной оптимизации редко решаются эффективно и математически строго. Для этой достаточно общей задачи нами предложен алгоритм решения в случае худших исходных данных кубической сложности и доказано, что он строго решает поставленную задачу за это время. На типичных данных он работает еще быстрее. Компьютерная программа, реализующая этот алгоритм вместе с примерами вычислений и инструкцией пользователя, свободно доступна на сайте `http://lab6.iitp.ru/ru/super3gl/`. В качестве одной из возможных интерпретаций задачи нами предложена модель эволюционных событий, которая описана на формальном уровне ниже, а также в [4], а на биологическом уровне – в [3, 5].

Итак, пусть листьям каждого дерева G_i ("дерева генов") приписаны пары $\langle k, l\rangle$ натуральных чисел; первое число назовем "геном", второе – "видом". Содержательно это – отношение "ген k обнаружен в виде l", которое далее будем называть отношением "ген-вид". В дереве генов какой-то вид l может сопровождаться несколькими генами $\langle k_1, l\rangle$, $\langle k_2, l\rangle, \ldots$ В разных деревьях генов G_i и G_j виды могут совпадать. Обозначим через V_0 множество всех видов, представленных в листьях всех G_i.

Условимся, что у всех деревьев корень располагается "сверху". Обозначим через e^- и e^+ верхний и нижний концы любого ребра e. Ребро понимается как пара вершин: начало e^- и конец e^+. Ребро, ведущее в вершину g, обозначим b_g. Каждое дерево рассматривается вместе с его "корневым ребром" – специально добавленным ребром, которое идет от корня вверх и соответствует времени, в котором жил общий предок всех представленных в дереве генов или видов; верхний конец корневого ребра назовем "суперкорнем". Ребра дерева видов S называются *трубами*, в частности, корневое ребро называется корневой трубой; этот термин вводится

только затем, чтобы различать ребра в S и ребра в G_i. На вершинах любого дерева определим отношение "ниже": $g_1 < g_2$, если $g_1 \neq g_2$ и в g_1 можно провести путь из суперкорня через g_2; везде "путь" понимается как *кратчайший* по числу ребер. Аналогично определяется отношение "ниже" между ребрами дерева. Различные ребра называются *несравнимыми*, если ни одно из них не ниже другого. Иначе ребра называются сравнимыми, они располагаются на общем пути из листа в суперкорень. Ребра, выходящие из одной вершины и дальние от корня, называются *смежными*, как и поддеревья с этими ребрами в качестве корневых; они образуют пару смежных ребер и соответственно пару смежных поддеревьев. На множестве всех вершин и труб в S определим единое отношение порядка $y < x$ таким образом: вершина или труба y "ниже" некоторой вершины или трубы x в S, если $y \neq x$ и в y можно провести путь из суперкорня через x; соответственно, "x выше y". Обозначим $y \leqslant x$, если $y < x$ или $y = x$. Каждое поддерево (все, что ниже некоторой вершины g) включает свое корневое ребро b_g, но не его верхний конец. Кладой M_s в дереве видов назовем множество видов, приписанных листьям, которые расположены ниже вершины s в дереве S. Кладой M_g (подразумевается: в одном из деревьев генов G_i) назовем множество видов, приписанных листьям, расположенным ниже вершины g в G_i. Вершины s и g назовем корнями соответствующих клад. Кладой M_e назовем множество видов, приписанных всем листьям ниже ребра/трубы e в дереве G или S.

Пусть P – фиксированный набор множеств видов, включающий V_0 и все его одноэлементные подмножества, но не включающий пустое множество. *Пространство деревьев* $\boldsymbol{P}$ состоит из всех деревьев видов S, у которых множество листьев взаимно-однозначно с видами из V_0 и все клады принадлежат P. В этом смысле будем называть P *набором клад*. *Стандартным набором* P назовем множество всех клад во всех исходных деревьях генов G_i, пополненное множеством V_0.

Вложением (без переносов) дерева G в дерево S называется отображение f всех вершин $V(G)$ дерева генов G в вершины $V(S)$ и трубы $E(S)$ дерева видов S, для которого выполнены условия:

1. Суперкорень в G отображается в корневую трубу в S; каждый лист g в G отображается в лист s в S согласно отношению ген-вид;
2. Пусть g_1 – сын g: если $f(g)$ – вершина, то $f(g_1) < f(g)$; а если $f(g)$ – труба, то $f(g_1) \leqslant f(g)$;
3. Пусть g_1 и g_2 – сыновья вершины g: если $f(g)$ – вершина, то путь в S из $f(g_1)$ в $f(g_2)$ проходит через $f(g)$.

Отметим, что вложение всюду определено, но не обязательно инъективно или сюръективно.

Для данного вложения f *дупликацией* называется несуперкорневая вершина g в G, для которой $f(g)$ – труба. *Дивергенцией* называется вершина g, для которой $f(g)$ – нелистовая вершина. *Потерей* называется пара $\langle e, s\rangle$, для которой e – ребро в G, s – вершина в S и выполняется $f(e^+) < s < f(e^-)$.

Замечание 1. Эти определения основаны на интуитивном представлении о процессе "эволюции первичного гена внутри первичного вида", расположенного в корневой трубе. Кратко такой процесс можно описать следующим образом. На рис. 1, 2 приведены иллюстрации понятий дупликации и потери гена, а также дивергенции. Дупликация гена – это появление двух его копий, не связанное с развилкой в дереве S; поскольку дупликация не соответствует никакой вершине в S, ее рисуют внутри трубы. Дивергенция гена – это появление двух его копий в развилке в S, одна копия идет ("линия жизни") в одну, другая – в другую из двух труб, выходящих из этой развилки (в смежные трубы), причем обе копии не теряются в них. Потеря гена происходит после появления двух копий в развилке в S, причем в одной из смежных труб копия теряется, а в другой не теряется; в этом случае линию жизни гена рисуют только в ту из этих смежных труб, в которой копия не теряется. Если обе копии гена потерялись в смежных трубах, то ген потерялся еще до развилки.

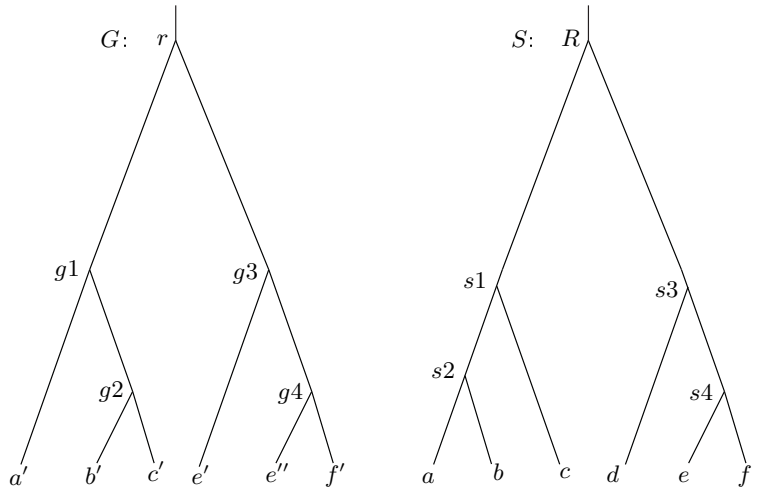

Рис. 1. Иллюстрация понятий дупликации, потери гена и дивергенции. Деревья генов G и видов S, у которых в листьях: ген a' взят из вида a и т.д., два гена e' и e'' взяты из одного вида e. Вид d не представлен в дереве G

Копия гена – также ген. В процессе эволюции ген как последовательность меняется, за большее время – больше.

Вложением f набора деревьев $\{G_i\}$ в дерево S назовем набор вложений $f = \{f_i\}$, в котором каждое f_i является вложением G_i в S.

Задача 1 – найти точку глобального минимума функционала

$$c(\{G_i\}, f, S) = \sum_i \big(c_l l(f_i, G_i, S) + c_d d(f_i, G_i, S)\big) \tag{1}$$

в указанном выше пространстве $\boldsymbol{P}$, где c_l и c_d – фиксированные неотрицательные числа, а S и все f_i – переменные, по которым выполняется глобальная минимизация. Напомним, что $\boldsymbol{P}$ состоит из деревьев с множеством V_0 листьев. В (1) использованы следующие обозначения: $l(f, G_i, S)$ – число потерь в дереве G_i при вложении f_i дерева G_i в S, а c_l – цена за одну потерю; таким образом, $c_l \sum\limits_i l(f_i, G_i, S)$ – суммарная цена за все потери во всех G_i. Аналогично, $d(f_i, G_i, S)$ – число дупликаций в дереве G_i при вложении f_i дерева G_i в S, а c_d – цена за одну дупликацию, и $c_d \sum\limits_i d(f_i, G_i, S)$ – суммарная цена за все дупликации во всех G_i. Таким образом, "близость" каждого G_i и S определяется через вложение f_i. Вложение $f^* = \{f_i^*\}$ назовем *сценарием* для набора деревьев $\{G_i\}$, если оно является решением задачи (1). Тогда значение c^* функционала (1) в точке глобального минимума $\langle f^*, S^* \rangle$ назовем *минимальной ценой* (сценария), а само S^* назовем *супердеревом* для набора $\{G_i\}$.

Содержательная интерпретация решения задачи (1) зависит от параметра P. Наши компьютерные эксперименты показали, что в задачах эволюции разумно выбирать сначала стандартный набор P, а затем расширять его разностями множеств, входящих в P, роль таких разностей поясняется в [6].

Парным сценарием h для одного дерева генов G и данного дерева видов S назовем вложение G в S, минимизирующее функционал (1), в котором i принимает ровно одно значение, т.е. $n = 1$ и $G = G_1$. Здесь S фиксировано, и единственной переменной является h. Если f^* – сценарий для какого-то набора $\{G_i\}$, а S^* – со-

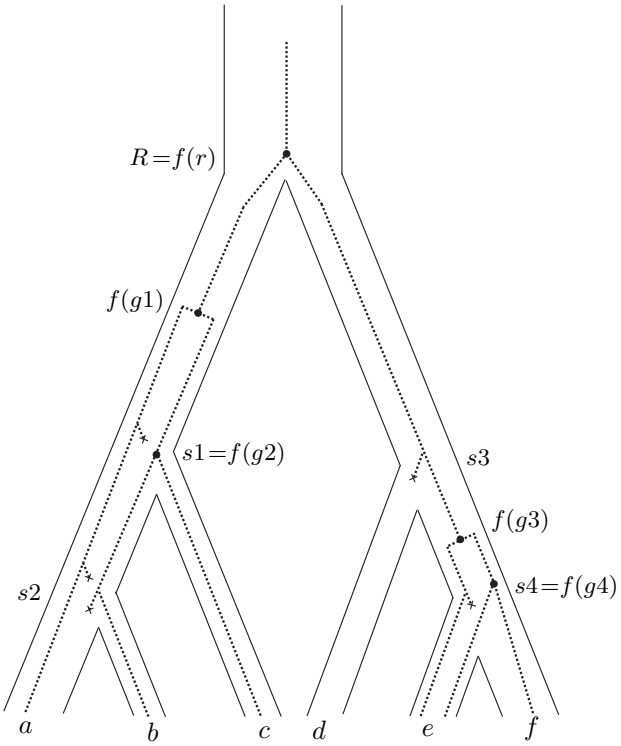

Рис. 2. Иллюстрация понятий дупликации, потери гена и дивергенции. Значения вложения f дерева G в дерево S показаны внутри труб дерева S жирными точками, кроме значений на листьях в G, которые совпадают с соответствующими листьями в S. Значение $f(g_1)$ показано в трубе (хотя формально оно равно этой трубе) и по определению вершина g_1 соответствует событию дупликации. Аналогично для вершины g_3. Значения f на всех других внутренних вершинах дерева G совпадают с соответствующими внутренними вершинами дерева S и по определению соответствуют событиям дивергенции. Для ребра $l = (g_1, a')$ вершины s_1 и s_2 лежат между значениями f на концах l и по определению пары $\langle l, s_1\rangle$ и $\langle l, s_2\rangle$ соответствуют событиям потери; потери показаны отростками с крестиком на конце. Аналогично, потерями являются пары $\langle (g_2, b'), s_2\rangle$, $\langle (g_3, e'), s_4\rangle$, $\langle (r, g_3), s_3\rangle$

ответствующее этому набору супердерево, то все f_i^* являются, очевидно, парными сценариями для каждой пары G_i и S^*.

Для любых G и S парный сценарий $h(G, S)$ единствен и даже не зависит от выбора фиксированных неотрицательных значений цен c_d и c_l за одну дупликацию и одну потерю. Этот сценарий h описан явно в лемме 1 (см. ниже). Поэтому при поиске супердерева S^* можно в каждое слагаемое в правой части в (1) подставить вместо f_i соответствующий единственный сценарий $h(G_i, S) = h_i$ как функцию от S, и тогда функционал (1) не будет зависеть от переменных f_i. Далее будем считать, что такая подстановка выполнена.

Множество V из P назовем *базисным*, если его можно разбить на какие-то две части из P, каждую часть, в свою очередь, можно разбить на две части из P, и так далее до достижения одноэлементных множеств, представляющих виды. Заранее не известно, какие множества из P являются базисными; в частности, не ясно, является ли само V_0 базисным. Однако, если решение S^* задачи (1) существует, то V_0 – базисное, так как требуемое измельчение V_0 задается кладами в S^*, которые в этом случае по определению входят в P.

Для описания нашего алгоритма понадобятся еще следующие понятия. Пусть f – вложение, тогда ребро e в G *входит* в трубу b в S, если $f(e^+) \leqslant b < f(e^-)$. Оказывается, что множество всех ребер, входящих в любую трубу b, можно эффективно находить для парного сценария $h(G,S)$. А именно, для любого множества видов M определим $\mathrm{Ed}(M,G)$ как множество ребер e в G, для которых $M_e \subseteq M$ и не существует ребра $e' > e$ с этим свойством. Таких e может быть несколько и все они несравнимы в G. Тогда для любой трубы b множество всех ребер, входящих в b, совпадает с $\mathrm{Ed}(M_b,G)$, в этом состоит лемма 2, а).

Напомним, что поддерево в G или S, определяемое вершиной g, имеет корень g, и b_g – его корневое ребро/корневая труба. Далее везде фиксированы набор P множеств и набор $\{G_i\}$ деревьев, поэтому они обычно не упоминаются явно.

Описание алгоритма. Для каждого дерева генов G_i и всех множеств V из P множества $\mathrm{Ed}(V,G_i)$ строятся перебором непосредственно по их определению.

Затем совместной индукцией по возрастанию числа элементов в V строятся определенные деревья $S(V)$. Точнее, вычисляется некоторое число – "цена" $c(V)$ множества V – и уже по ней тривиально строится $S(V)$, у которого все клады принадлежат P и листьям приписаны в точности виды из V. Любое дерево вида $S(V)$ будем называть *базисным*.

Итак, пусть V – базисное множество из P. Начальный шаг: в качестве V берутся одноэлементные множества из P, каждое из них состоит из одного вида; цена $c(V)$ по определению полагается равной нулю, а соответствующее дерево $S(V)$ по определению состоит из одного листа, которому приписан этот вид.

Индуктивный шаг: рассматриваются всевозможные разбиения множества V из P на два *базисных* множества V_1 и V_2. Интуитивно это означает проверку, не будет ли развилка при корне будущего дерева $S(V)$ задаваться разбиением V на V_1 и V_2. Если отсутствует хотя бы одно такое разбиение, то множество V помечается как "не базисное"; множества с меньшим числом элементов, чем в V, уже помечены как "базисные" или "не базисные". Если V_0 помечается как "не базисное", то алгоритм выдает сообщение "задача (1) не имеет решения". Далее рассматривается случай, когда V – базисное.

Для любого разбиения базисного множества V на базисные множества V_1 и V_2 по предположению индукции уже вычислены цены $c(V_1)$ и $c(V_2)$ и построены базисные деревья $S(V_1)$ и $S(V_2)$. Для каждого G_i положим $l(i) = |\mathrm{Ed}(V_1,G_i)| + |\mathrm{Ed}(V_2,G_i)|$ и $d(i) = l(i) - |\mathrm{Ed}(V,G_i)|$, здесь $|\cdot|$ обозначает мощность множества. В произвольном порядке переберем все деревья из набора $\{G_i\}$, и для каждого из них переберем все вершины в G_i. Для каждой такой вершины, если ребро одного ее сына принадлежит $\mathrm{Ed}(V_1,G_i)$, а ребро другого ее сына принадлежит $\mathrm{Ed}(V_2,G_i)$, уменьшим числа $l(i)$ на 2, а $d(i)$ – на 1. Полученные в результате числа обозначим $l(V,V_1,V_2,G_i)$ и $d(V,V_1,V_2,G_i)$.

Найдем разбиение V на V_1^* и V_2^*, для которого функционал

$$c(V,V_1,V_2) = \sum_i \bigl[c_l l(V,V_1,V_2,G_i) + c_d d(V,V_1,V_2,G_i)\bigr] + c(V_1) + c(V_2) \qquad (2)$$

достигает минимума по всем разбиениям фиксированного V на базисные множества V_1 и V_2. Пусть по определению $c(V)$ – значение функционала (2) на таком *минимальном разбиении* $\langle V_1^*, V_2^*\rangle$. Полученное $c(V)$ назовем *ценой множества* V. После этого базисное дерево $S(V)$ по определению получается добавлением корня к базисным деревьям $S(V_1^*)$ и $S(V_2^*)$; корень соответствует V, а его сыновья соответствуют V_1^* и V_2^*. Цены $c(V_1)$ и $c(V_2)$ множеств V_1 и V_2 индуктивно получаются как значения функционала (2) на каких-то своих минимальных разбиениях. Описание алгоритма закончено.

Может быть несколько минимальных разбиений; было бы интересно получить нетривиальную оценку их числа.

Чтобы охарактеризовать все базисные деревья, составляющие набор $\{S(V) : V -$ базисное множество$\}$, нужно несколько расширить задачу 1: найти глобальный минимум функционала (1) на том же пространстве $\boldsymbol{P}$ деревьев, в котором суммирование по деревьям G_i дополнено суммированием по всем их поддеревьям G', у которых корневыми являются ребра из $\mathrm{Ed}(V, G_i)$, а V_0 заменено на V. Получаем новый функционал

$$c(\{G_i\}, f, S) = \sum_i \sum_{G'} \big(c_l l(f_{G'}, G', S) + c_d d(f_{G'}, G', S)\big). \tag{3}$$

Это расширение задачи 1 назовем *задачей* 2. Если $V = V_0$, то все G' в G_i совпадают с самим G_i, и тем самым, функционал (3) совпадает с функционалом (1), а задача 2 – с задачей 1. Для любых деревьев G' и S единственный (как мы снова увидим из леммы 1) парный сценарий $h(G', S)$ можно подставить вместо $f_{G'}$, и тогда минимизация по переменным $f_{G'}$ в (3), как и в (1), не нужна. Решение S^* задачи 2 будем также называть *супердеревом* (для множества видов V). Далее будем считать, что в функционале (3) выполнена указанная подстановка $h(G', S)$ вместо $f_{G'}$.

Теорема. *Пусть P – набор клад.*

a) *Множество V_0 – базисное, если и только если найденное алгоритмом дерево $S(V_0)$ является решением задачи* 1.

b) *Для любого базисного множества V из P найденное алгоритмом дерево $S(V)$ – одно из решений задачи* 2. *И наоборот, любое решение задачи* 2 *имеет вид $S(V)$ при соответствующем выборе последовательности минимальных разбиений.*

c) *Если P – стандартный набор и среднее число листьев в наборе деревьев генов $\{G_i\}$ порядка $|V_0|$, то алгоритм определяет множество $\{S(V) : V -$ базисное множество$\}$ за число шагов порядка $|P|^3 + |P|^2|V_0|n \leqslant Cn^3|V_0|^3$. За это время алгоритм выдает решение задачи* 1 *или сообщает, что оно не существует.*

Доказательство теоремы использует нижеуказанные леммы 1–3 и будет приведено после их доказательств.

Вершины g и s назовем *согласованными*, если они не суперкорни и при этом 1) g и s – листья, находящиеся в отношении ген-вид, или 2) совпадают два разбиения множества M_g: то, которое определяется развилкой в g, и то, которое определяется развилкой в s. Последнее означает: ($M_{g1} \subseteq M_{s1}$ и $M_{g2} \subseteq M_{s2}$) или ($M_{g1} \subseteq M_{s2}$ и $M_{g2} \subseteq M_{s1}$), где g_1 и g_2 – сыновья g, а s_1 и s_2 – сыновья s. Обозначим через $\sup M$ вершину в дереве S, которая является точной верхней гранью множества M листьев (видов) в S. Напомним, что b_s обозначает трубу в S с концом в вершине s. Обозначим через $h(g) = h(G, S)$ следующее отображение вершин g дерева G в вершины и трубы дерева S:

$$\begin{cases} \text{Если } g - \text{суперкорень в } G, \text{ то } h(g) \text{ равно корневой трубе в } S; \\ \text{иначе: если вершины } g \text{ и } \sup M_g \text{ согласованы, то } h(g) = \sup M_g; \\ \text{иначе: } h(g) = b_s, \text{ где } s = \sup M_g. \end{cases} \tag{4}$$

Лемма 1. *Отображение $h(g)$ является вложением со следующим свойством: для любого вложения f дерева G в S, $f \neq h$, числа дупликаций и потерь для f не меньше, чем аналогичные числа для h, и хотя бы одно из этих чисел для f строго больше, чем для h.*

Отсюда сразу следует, что при любых неотрицательных ценах c_l и c_d вложение h является единственным парным сценарием для G и S.

Доказательство. Проверим, что h – вложение. Свойство 1 выполняется тривиально. Нестрогое неравенство в свойстве 2 следует из того, что $\sup M_{g1} \leqslant \sup M_g$.

Строгое неравенство в свойстве 2: вершины g и $\sup M_g$ согласованы, отсюда $h(g_1) \leqslant$ $\leqslant b_{s1}$ или $h(g_1) \leqslant b_{s2}$, т.е. $h(g_1) < s$. Свойство 3: аналогично предыдущему $h(g_1) \leqslant b_{s1}$ и $h(g_2) \leqslant b_{s2}$ (или симметрично), т.е. $h(g_1)$ и $h(g_2)$ лежат в разных смежных поддеревьях.

Пусть до конца доказательства вложение f отлично от h, т.е. $f \neq h$.

Для любой вершины g из G выполняется

$$f(g) \geqslant h(g). \tag{5}$$

Действительно, по свойству 2 имеем $f(g) \geqslant \sup M_g$. Если $f(g) = \sup M_g$, то из свойства 3 следует, что вершины g и $\sup M_g$ согласованы, откуда $f(g) = h(g)$. Если $f(g) > \sup M_g$, то $f(g) \geqslant h(g)$.

$$\text{Если } f(g) > \sup M_g, \text{ то } f(g) - \text{труба}. \tag{6}$$

Предположим, что $f(g)$ – некоторая вершина s. Существует такой сын s_1 вершины s, для которого не выполняется $s_1 \geqslant \sup M_g$. По свойству 3 для сыновей g_1 и g_2 вершины g выполняется $M_{g1} \subseteq M_{s1}$ или $M_{g2} \subseteq M_{s1}$, что противоречит условию $f(g) > \sup M_g$.

Из (5), (6) имеем:

$$\text{Для любой вершины } g, \text{ если } h(g) - \text{труба, то } f(g) - \text{труба}. \tag{7}$$

Отсюда число дупликаций для f не меньше, чем число дупликаций для h. Докажем такое же утверждение для потерь.

Пусть $\langle e, s\rangle$ – потеря для h. Пусть $f(e^+) < s$. Тогда $\langle e, s\rangle$ – потеря для f с учетом (5). Иначе, так как $f(e^+) \geqslant h(e^+) < s$, имеем $s = f(e^+)$ или $s < f(e^+)$. Первое невозможно, так как $f(e^+)$ – труба согласно (6). Пусть $s < f(e^+)$. Покажем, что путь из e^+ в любой лист l содержит такое ребро e', что $\langle e', s\rangle$ – потеря для f; и таким образом, исходной потере $\langle e, s\rangle$ для h сопоставляется множество, состоящее по крайней мере из двух разных потерь вида $\langle e', s\rangle$ для f. Действительно, $h(e^+) < s$ означает $\sup M_{e^+} < s$. Рассмотрим любую вершину g на пути из e^+ в l и покажем, что $f(g) \neq s$. Если это не так, то $\sup M_g \leqslant \sup M_{e^+} < f(g) = s$ и $f(g)$ – труба согласно (6), противоречие. По свойству 2 имеем $f(l) \leqslant f(g) \leqslant f(e^+)$; кроме того, выполняется $f(l) < s < f(e^+)$ – второе неравенство по условию, а первое в силу $f(l) = h(l) \leqslant h(e^+) < s$. Следовательно, на этом пути найдутся соседние вершины k^+ и k^-, для которых $f(k^+) < s < f(k^-)$, т.е. ребро $\langle k^+, k^-\rangle$ и вершина s образуют потерю для f. Итак, каждой потере для h сопоставлена та же самая потеря для f ("первый случай", рассмотрим ее как одноэлементное множество) или множество потерь для f мощности строго больше единицы ("второй случай"). Докажем, что эти множества не пересекаются. Пусть $\langle e_1, s_1\rangle$ и $\langle e_2, s_2\rangle$ – две потери для h. Соответствующие потери для f имеют вид $\langle e_1', s_1\rangle$ и $\langle e_2', s_2\rangle$. Если $s \neq s_1$, то эти пары разные. В противном случае $s_1 = s_2$ и $e_1 \neq e_2$. По определению потери ребра e_1 и e_2 несравнимы в G, поэтому e_1' и e_2' не равны. Поэтому число потерь для f не меньше, чем число потерь для h.

Предположим, что числа дупликаций для f и h совпадают. Покажем, что тогда хотя бы один раз встретится второй случай (т.е. $s < f(e^+)$) и поэтому число потерь для f будет строго больше числа потерь для h. Поскольку $f \neq h$, существует вершина g из G, для которой $f(g) > h(g)$, т.е. $f(g) > \sup M_g$ и $f(g)$ – труба согласно (6). В силу (7) и по предположению выполняется

$$\{k \mid h(k) - \text{труба}\} = \{k \mid f(k) - \text{труба}\}, \tag{8}$$

тогда $h(g)$ – труба. Рассмотрим вершину s, для которой $h(g) < s < f(g)$. Рассмотрим путь в G из g в суперкорень. Для любой вершины g' на нем выполняется $h(g') \neq s$.

Если это не так, то $h(g') = s$, и по свойству 2 имеем $f(g') \geqslant f(g) > s = h(g') = \sup M_{g'}$, $f(g')$ – труба по (6), и получаем противоречие с (8). Следовательно, на этом пути найдутся две соседние вершины g' (может быть, равная g) и g'' (может быть, равная суперкорню), для которых $h(g') < s < h(g'')$, т.е. ребро $e = (g'', g')$ и вершина s образуют потерю для h. По свойству 2 имеем $f(e^+) \geqslant f(g)$ и по выбору g и s имеем $f(g) > s$, отсюда $f(e^+) > s$, т.е. имеет место второй случай. ▲

Для любой трубы b множество всех ребер, входящих в b, и множество $\mathrm{Ed}(M_b, G)$ связаны следующим образом.

Лемма 2. *Если h – парный сценарий для G и S, то*:

a) *в трубу b в S входят в точности ребра из* $\mathrm{Ed}(M_b, G)$;

b) *если труба b_1 – сын трубы b, то для любого ребра e в G, входящего в b_1, существует ровно одно ребро $e' \geqslant e$, входящее в b.*

Доказательство. a) Пусть e – ребро из $\mathrm{Ed}(M_b, G)$. Тогда $\sup M_{e^+} \leqslant b^+$, а $\sup M_{e^-} \geqslant b^-$. По определению (4) парного сценария h получим, что e входит в b. Наоборот, если e входит в b, то $\sup M_{e^+} \leqslant b^+$, а $\sup M_{e^-} \geqslant b^-$, т.е. e принадлежит $\mathrm{Ed}(M_b, G)$.

b) Рассмотрим путь от e к корневому ребру. Пусть e_1 – первое ребро на этом пути, для которого $b^- \leqslant h(e_1^-)$. Тогда e_1^+ совпадает с верхним концом ребра, предшествующего e_1 на этом пути, или $e_1 = e$. В обоих случаях $h(e_1^+) \leqslant b$, поэтому e_1 входит в b. Поскольку из двух сравнимых ребер только одно может входить в любую трубу, то никакое другое ребро на этом пути не входит в b. ▲

Для любых деревьев генов G и видов S и соответствующего парного сценария $h(G, S)$ определим "место" каждого эволюционного события: *местом* дупликации g считаем трубу $h(g)$, а *местом* потери $\langle e, s\rangle$ – трубу b_s, что технически удобнее, чем было сказано в замечании 1. Напомним, что супердерево для V – это дерево, минимизирующее функционал $C(V, S)$, заданный формулой (3), а минимальное разбиение – разбиение, минимизирующее другой функционал $c(V, V_1, V_2)$, заданный формулой (2).

Нелистовую вершину g в дереве генов назовем *паралогичной*, если клада M_g состоит из одного вида. У каждого парного сценария паралогичная вершина является дупликацией в листовой трубе и обратно: любая дупликация в листовой трубе является паралогичной вершиной. Слагаемые в формуле (1), соответствующие паралогичным вершинам, можно опустить, так как они в сумме дают константу, не влияющую на минимизацию.

Лемма 3. a) *Пусть S_0 – дерево видов с множеством листьев V_0, а S – его поддерево с множеством листьев V. Суммарная цена событий $Z(S)$ (дупликаций и потерь), имеющих место в поддереве S при парных сценариях для всех G_i и данного S_0, равна $C(V, S)$.*

b) *Если V_1, V_2 – разбиение при корне любого дерева S с множеством листьев V, а поддеревья при корне имеют вид $S(V_1)$ и $S(V_2)$, то $c(V, V_1, V_2) = C(V, S)$.*

Если дерево S еще и минимальное, то разбиение V_1, V_2 – минимальное.

c) *Пусть S – любое дерево с множеством листьев V, а S_1 – любое его собственное поддерево с множеством листьев V_1. Если $[S, S_2/S_1]$ – результат замены в S поддерева S_1 на дерево S_2 с теми же листьями, что в S_1, и $C(S_2) \leqslant C(S_1)$, то $C([S, S_2/S_1]) \leqslant C(S)$.*

Доказательство. a) Покажем, что выполняется следующее: суммарная цена Z этих событий для одного G_i равна одному слагаемому $C(G_i, V, S)$ в сумме из $C(\{G_i\}, V, S)$, соответствующему этому G_i. Пусть G' – поддерево в G_i, для которого корневое ребро принадлежит $\mathrm{Ed}(V, G_i)$. Далее вместо G_i пишем G.

Сформулированное утверждение следует из более общего: каждое событие, имеющее место при парном сценарии для G' и S, является событием в S при парном сценарии для G и S_0, и наоборот, каждое событие, имеющее место в S при парном сценарии для G и S_0, является событием при парном сценарии для G' и S при единственном G'.

По определению (4) парного сценария h имеем следующее: если вершина g принадлежит поддереву G', то $h_{G'}(g)$ при парном сценарии для G' и S совпадает с $h_1(g)$ при парном сценарии для G и S_0. И наоборот: если $h_1(g)$ лежит в S, то найдется единственное G', содержащее g, и $h_{G'}(g) = h_1(g)$. Действительно, клада $h_1(g)$ в S содержится в V, тогда по упомянутому определению клада g содержится в V, и двигаясь от g вверх, найдем корневое ребро искомого единственного G'.

Проверим: если имеется дупликация или потеря в S при парном сценарии h_1, то она остается тем же событием для ровно одного парного сценария $h_{G'}$ для G' и S, и наоборот. Для дупликации это сразу вытекает из предыдущего абзаца.

Пусть $\langle e, s\rangle$ – потеря при парном сценарии h_1 для G и S_0, а b_s принадлежит S. Тогда e^+ принадлежит некоторому G'. Если e – не корневое ребро в G', то $\langle e, s\rangle$ – потеря при парном сценарии $h_{G'}$, поскольку образы вершин e^+ и e^- не изменились. Если e – корневое ребро в G', то $\langle e, s\rangle$ – также потеря при парном сценарии $h_{G'}$, поскольку образ вершины e^+ не изменился, а $h_{G'}(e^-)$ равно корневой трубе дерева S, т.е. $h_{G'}(e^+) < s < h_{G'}(e^-)$.

Пусть $\langle e, s\rangle$ – потеря для $h_{G'}$. Если e – не корневое ребро в G', то $\langle e, s\rangle$ – потеря и при парном сценарии h_1, поскольку образы вершин e^+ и e^- не изменились. Если e – корневое ребро в G', то $\langle e, s\rangle$ – также потеря для h_1, так как $h_{G'}(e^+) = h_1(e^+)$ и $h_{G'}(e^-) < h_1(e^-)$.

Суммируя по i слагаемые $C(G_i, V, S)$, получаем утверждение пункта a).

b) Второе утверждение этого пункта сразу следует из первого: если это разбиение – не минимальное, то перейдем к деревьям над минимальным разбиением и получим дерево над V со строго меньшей ценой C, что невозможно.

По индукции проверим первое утверждение. Если S состоит из одного листа, то V – из одного вида и $C(V, S) = 0$ по пункту a) (слагаемые, соответствующие паралогичным вершинам, опущены), и $c(V) = 0$ по определению.

Индуктивный шаг: по предположению индукции для деревьев $S(V_1)$ и $S(V_2)$ выполняется следующее: $c(V_1) = C(V_1, S(V_1))$, по пункту a) имеем $c(V_1, S(V_1)) = = Z(S(V_1))$ и аналогично $c(V_2) = C(V_2, S(V_2)) = Z(S(V_2))$. Тогда

$$c(V, V_1, V_2) = \sum_i \big[c_l l(V, V_1, V_2, G_i) + c_d d(V, V_1, V_2, G_i)\big] + Z(S(V_1)) + Z(S(V_2));$$

ниже показано, что первое слагаемое для каждого G_i – цена событий в корневой трубе дерева S при парном сценарии для G_i и произвольного S_0 с множеством листьев V_0, содержащего S в качестве поддерева. Поэтому правая часть – цена событий во всех трубах этого дерева, т.е. $Z(S(V))$. По утверждению пункта a) это равно $C(V, S)$. Обозначим через b корневую трубу дерева S, а через b_1 и b_2 – трубы, выходящие из b.

По лемме 2, a) для каждого дерева генов G в трубы b, b_1 и b_2 входят ребра дерева G, соответственно, из $\mathrm{Ed}(V, G)$, $\mathrm{Ed}(V_1, G)$ и $\mathrm{Ed}(V_2, G)$. По определению множества $\mathrm{Ed}(V_1, G)$ и $\mathrm{Ed}(V_2, G)$ не пересекаются, и любые два ребра из их объединения M несравнимы в G. По лемме 2, b) каждое ребро e из G, входящее в трубу b_1 или b_2, имеет единственного предка – ребро $e' \geqslant e$, входящее в трубу b. В вершинах из G, расположенных на пути от e к e', находятся дупликации в b, а в первой вершине пути или на ребре e находится, соответственно, дивергенция или потеря. Ребро e' из $\mathrm{Ed}(V, G)$ порождает поддерево в G с корнем в конце e' и листьями

в началах ребер из множества M. Для ребер из $\mathrm{Ed}(V,G)$ получается *лес* таких деревьев в количестве $|\mathrm{Ed}(V,G)|$; ребра из M взаимно-однозначно соответствуют листьям, а ребра из $\mathrm{Ed}(V,G)$ являются корневыми. В этих деревьях содержится $d(G) = |\mathrm{Ed}(V_1,G)| + |\mathrm{Ed}(V_2,G)| - |\mathrm{Ed}(V,G)|$ вершин дерева G. Каждая из них при парном сценарии для G_i и S_0 отображается в трубу b и тогда является дупликацией, или в вершину r (развилку при корне) и тогда является дивергенцией. Обратно: если образ вершины равен b или r, то она является одной из этих $d(G)$ вершин, так как двигаясь от этой вершины вниз по любому пути в G, мы обязательно придем в ребро из M. Из этих $d(G)$ вершин дивергенциями являются те, для которых ребро одного сына принадлежит $\mathrm{Ed}(V_1,G)$, а ребро другого принадлежит $\mathrm{Ed}(V_2,G)$. Остальные вершины являются дупликациями. Для любого ребра e из M пара $\langle e,r\rangle$ является потерей тогда и только тогда, когда e не является сыном дивергенции. Обратно: любая потеря вида $\langle e,r\rangle$ соответствует ребру e из M. Таким образом, мы показали, что $l(V,V_1,V_2,G_i)$ – число потерь на развилке, а $d(V,V_1,V_2,G_i)$ – число дупликаций в b.

c) Произвольно достроим S до некоторого дерева видов S_0 с множеством листьев V_0. Покажем, что верно следующее утверждение:

$$\begin{array}{l}\text{Если } C(S_2) \leqslant C(S_1), \text{ то } C([S_0,S_2/S_1]) \leqslant C(S_0),\\ \text{а если } C(S_2) < C(S_1), \text{ то } C([S_0,S_2/S_1]) < C(S_0).\end{array} \tag{9}$$

Рассмотрим парный сценарий для какого-то одного G_i и S_0 и сравним связанные с ним события в S_0, до замены в S_0 поддерева S_1 на S_2, и после этой замены, когда вместо S_0 получается дерево S_3. Согласно пункту a) суммарная цена событий, имеющих место в дереве S_2, не больше суммарной цены событий, имеющих место в дереве S_1. Теперь достаточно показать, что в той части S_0, которая не подверглась изменениям, события остались теми же.

Проверим, что каждое событие, которое происходит в дополнении S_2 до S_3, происходит в дополнении S_1 до S_0, и наоборот. Если $\langle g,h(g)\rangle$ – дупликация в дополнении S_2 до S_3 после замены, то по определению (4) до замены $h(g)$ – та же самая труба, поэтому $\langle g,h(g)\rangle$ – дупликация и до замены. Если $\langle e,s\rangle$ – потеря в дополнении S_2 до S_3 после замены, то значение $h(e^-)$ не лежит в S_2 (сейчас у нас поддерево не включает суперкорень) и по (4) не лежало в S_1 до замены. Если и значение $h(e^+)$ не лежит в S_2, то и оно не изменилось при замене, поэтому $\langle e,s\rangle$ и до замены была потерей. Если же значение $h(e^+)$ лежит в S_2, то оно по (4) до замены лежало в S_1, а так как s не лежит в S_2, то и до замены выполнялось $h(e^+) < s < h(e^-)$, а значит, $\langle e,s\rangle$ и до замены была потерей. Утверждение (9) доказано.

Из (9) сразу следует утверждение пункта c). Действительно, из условия имеем $C([S_0,S_2/S_1]) \leqslant C(S_0)$. Допустим, что утверждение пункта c) неверно, тогда $C(S) < C([S,S_2/S_1])$, и по второй части утверждения (9) получаем $C(S_0) < C([S_0,S_2/S_1])$ – противоречие. ▲

Доказательство теоремы. a) В одну сторону это утверждение следует из пункта b), поскольку множество $\mathrm{Ed}(V_0,G_i)$ состоит из корневого ребра дерева генов G_i. В другую сторону оно очевидно.

b) Воспользуемся индукцией по мощности V. Для любого базисного множества V рассмотрим минимальное дерево S^* с множеством листьев V и всеми кладами из P. Пусть V_1 и V_2 при корневой развилке в S^* соответствуют поддеревьям S_1 и S_2; по предположению индукции $S(V_1)$ и $S(V_2)$ – минимальные. Выберем S^* так, чтобы поддеревья S_1 и S_2 совпадали с $S(V_1)$ и $S(V_2)$. Для этого заменим S_1 на $S(V_1)$, и по лемме 3, c) значение функционала $C(V,S)$ не изменится; аналогично заменим S_2. По лемме 3, b) это разбиение V на V_1 и V_2 – минимальное, и обратно, любому минимальному разбиению соответствует минимальное дерево. Поэтому $S^* = S(V)$, что и утверждает пункт b) теоремы. Последнее утверждение пункта b) легко доказывается по индукции.

c) Для каждого элемента из P перебирается не более $|P|$ вариантов его разбиения, а для каждого варианта просматриваются все вершины во всех деревьях генов, что соответствует времени порядка $|P|^2|V_0|n$. Предварительное построение множеств $\mathrm{Ed}(M, G_i)$ для всех множеств M из P требует времени порядка $|P||V_0|n$. Предварительное построение отношений включения и пересечения множеств из P требует времени порядка $|P|^2|V_0|$. Предварительное построение всех вариантов разбиения множеств из P на два множества из P требует времени порядка $|P|^3$ (для каждой тройки множеств P_1, P_2, P_3 следует проверить, что P_2 и P_3 не пересекаются и $|P_2| + |P_3| = |P_1|$). Отсюда следует общая оценка времени порядка

$$|P|^3 + |P|^2|V_0|n \leqslant Cn^3|V_0|^3. \quad \blacktriangle$$

Замечание 2. 1) Если деревья генов небинарные, то описанный алгоритм нужно модифицировать следующим образом. Вместо числа ребер из $\mathrm{Ed}(V, G_i)$ рассматривать число вершин, таких что хотя бы одно их сыновнее ребро принадлежит $\mathrm{Ed}(V, G_i)$. Аналогично для $\mathrm{Ed}(V_1, G_i)$ и $\mathrm{Ed}(V_2, G_i)$. Вместо перебора дивергенций, т.е. вершин, у которых одно сыновнее ребро принадлежит $\mathrm{Ed}(V_1, G_i)$, а другое – $\mathrm{Ed}(V_2, G_i)$, перебирать вершины, такие что среди их сыновних ребер имеется хотя бы одно ребро из $\mathrm{Ed}(V_1, G_i)$ и хотя бы одно ребро из $\mathrm{Ed}(V_2, G_i)$.

2) Если деревья генов неукорененные, то для укоренения используется следующая процедура. Пусть каждому листу приписана метка – имя таксономической группы, к которой относится вид, представленный в этом листе. Эту метку будем называть *таксоном*. Из данного набора деревьев генов удаляются деревья с одним таксоном. Для оставшихся деревьев таксоны частично упорядочиваются по возрастанию древности; такая информация обычно доступна из биологических данных. Формально можно взять любую расстановку этих меток и любой порядок на них. Для каждого дерева G определим число k наиболее древних таксонов. Для примера опишем нашу процедуру для случаев $k = 1$ или $k = 2$, которые обычно имеют место в биологических данных. Общий случай рассматривается аналогично. Пусть $M(G)$ состоит из самого древнего таксона, если $k = 1$, и из двух самых древних таксонов, если $k = 2$ и число всех таксонов не меньше трех; в противном случае $M(G)$ состоит из одного из самых древних таксонов (не существенно какого). Вычислим $p(G)$ – степень "кучности" $M(G)$ в дереве G – следующим образом. Для этого сначала для каждого ребра $\{u, v\}$ (неупорядоченной пары) дерева G вычислим показатель d. Пусть b_u – число листьев с таксоном из $M(G)$ в той части U разбиения дерева G этим ребром, которая примыкает к вершине u, а b_v – в той части V, которая примыкает к вершине v. Пусть l_u – число всех листьев в части U, а l_v – в части V. Тогда $d_u = b_u/l_u$ – доля листьев с таксоном из $M(G)$ среди всех листьев в части U, а $d_v = b_v/l_v$ – в части V. Положим $d = (\sqrt{d_u} - \sqrt{d_v})^2$. Найдем ребро $e(G)$ с максимальным значением d (это значение обозначим $d_{\max}$). Если таких ребер несколько, то положим $d_{\mathrm{pmax}} = d_{\max}$. В противном случае пусть d_{pmax} равно второму по максимальности значению d. Положим $p = \sqrt{d_{\max}} - \sqrt{d_{\mathrm{pmax}}} + (d_{\max})^2$. В наборе деревьев генов оставим лишь деревья, у которых p больше заранее заданного порога. Для каждого из таких деревьев G поставим корень в середине ребра $e(G)$. К этому набору укорененных деревьев применим наш алгоритм.

Итоговый алгоритм показал хорошие результаты для наборов бинарных и небинарных, укорененных и неукорененных деревьев. Компьютерная программа построения супердерева вместе с примерами вычислений и инструкцией пользователя свободно доступна на сайте `http://lab6.iitp.ru/ru/super3gl/`.

В заключение сформулируем математическую проблему, которая представляется нам одной из центральных в математическом описании эволюции. Начнем с нескольких определений.

Во-первых, нужно ввести представление о времени протекания эволюционных событий. Может быть, для этого нужно перейти к непрерывному описанию дискретной картины, указанной ниже. Нами было предложено следующее описание дискретного времени [3, 4]. Будем различать исходное дерево видов S_0 и новое дерево видов S, которое получается из S_0 разбиением некоторых труб в S_0 на последовательные части ("новые трубы"). В результате в S появляются трубы с одним сыном. Некоторый алгоритм для перехода от S к S_0 предложен в [3]. В случае вложения без переносов $S = S_0$. "Временные слои" – это разбиение множества всех труб в S на непересекающиеся множества, которые пронумерованы от 1 до m; каждое множество – один "временной слой"; при этом должно выполняться следующее: для любой трубы b из i-го слоя ее сын b_1 принадлежит $(i+1)$-му слою. В первом слое лежит корневая труба из S, а последний (m-й по номеру) слой состоит из всех труб, ведущих в листья из S. Тогда i-й слой состоит из всех труб, в которые ведет путь из i труб, включая корневую трубу. Будем писать $b_1 \sim b_2$, если $b_1 \neq b_2$ и трубы b_1, b_2 принадлежат одному слою. Интуитивно в одном слое собираются трубы, относящиеся к одному периоду времени и между ними возможны одновременные события. Разбиением дерева G назовем дерево G', которое получается из G разбиением некоторых его ребер на последовательные части, в результате появляются "новые ребра" с одним сыном. В случае вложения без переносов $G' = G$.

Во-вторых, нужно описать эволюционное явление – горизонтальные переносы генов [3, 4]. Это было сделано нами следующим образом. *Вложением* (с переносами) называется отображение f всех вершин $V(G')$ некоторого разбиения G' дерева G в вершины $V(S)$ и трубы $E(S)$ дерева S, для которого выполнены следующие условия:

1. Суперкорень в G' отображается в корневую трубу в S; каждый лист g в G' отображается в лист s в S согласно отношению ген-вид.
   Далее g, g_1, g_2 – вершины в G';
2. Пусть g_1 – сын g: если $f(g)$ – вершина, то $f(g_1) < f(g)$, а если $f(g)$ – труба, то рассмотрим следующие два случая. Если g_2 – другой сын g, то для обоих сыновей $f(g_i) \leqslant f(g)$ или выполняется: для одного сына $f(g_i) \leqslant f(g)$ и для другого сына $f(g) \sim f(g_j)$; здесь $f(g_i)$ – вершина или труба и $f(g_j)$ – труба, $i, j = 1, 2$. Если g с родителем g' имеет только одного сына g_1, то $f(g_1) \leqslant f(g) \sim f(g')$ или $f(g) \sim$ $\sim f(g_1)$; здесь в первом выражении $f(g_1)$ – вершина или труба, $f(g')$ – труба, а во втором $f(g_1)$ – труба;
3. Пусть g_1 и g_2 – сыновья g: если $f(g)$ – вершина, то путь в S из $f(g_1)$ в $f(g_2)$ проходит через $f(g)$; если g имеет только одного сына, то $f(g)$ – труба.

Теперь *дупликация* гена – это вершина g в G' с двумя сыновьями g_1 и g_2, для которой $f(g)$ – труба из S и для обоих сыновей выполняется $f(g_i) \leqslant f(g)$, $i = 1, 2$. *Дивергенция* – это вершина g из G', для которой $f(g)$ – вершина в S и обе вершины g и $f(g)$ имеют по два сына. *Потеря* гена – это пара $\langle e, s\rangle$, для которой e – ребро в G', s – вершина в S, имеющая двух сыновей, и $f(e^+) < s < f(e^-)$. *Горизонтальный перенос* с сохранением – вершина g из G' с двумя сыновьями g_1 и g_2, для которой $f(g)$ – труба из S и ровно для одного из сыновей g_i выполняется $f(g) \sim f(g_i)$. *Горизонтальный перенос* без сохранения – вершина g из G' с единственным сыном g_1, для которой $f(g)$ – труба и $f(g) \sim f(g_1)$. Обычно перенос без сохранения рассматривается как последовательность двух событий: переноса с сохранением и потери копии гена в источнике. На рис. 3, 4 показано вложение с потерями и переносами.

Аналог задачи 1, сценария (с переносом), парного сценария (с переносом) и т.д. определяется аналогично предыдущему с использованием функционала, обобщающего функционал (1):

$$c_{\text{trans}}(\{G_i\}, f, S) = \sum_i \big(c_l l(f_i, G_i, S) + c_d d(f_i, G_i, S) + c_t^+ t^+(f_i, G_i, S) + \\ + c_t^- t^-(f_i, G_i, S)\big). \tag{10}$$

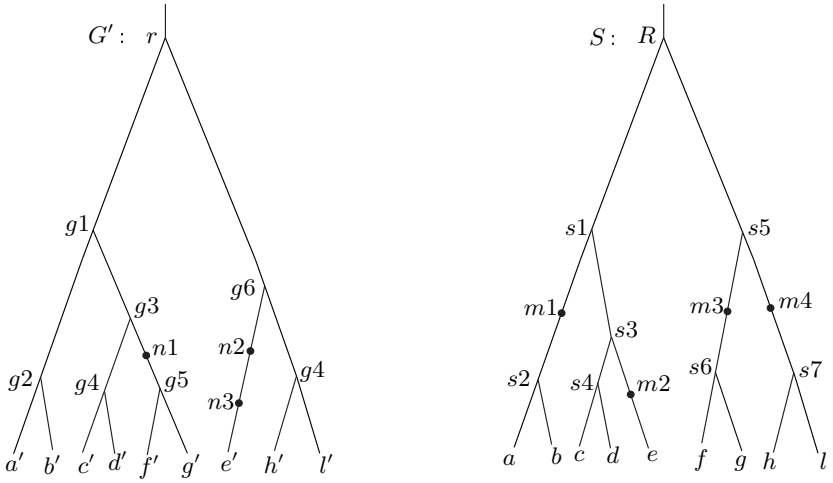

Рис. 3. Иллюстрация понятия горизонтального переноса: деревья генов G' и видов S, обозначения в листьях аналогичны тем, что на рис. 1, 2. Вершины с одним сыном, добавленные, соответственно, к G и S, обозначены жирными точками. В S в i-м временном слое находятся трубы, к которым ведет путь из суперкорня с i трубами

Здесь $f = \{f_i\}$, $t^+(f, G, S)$ – количество переносов с сохранением у вложения f, c_t^+ – цена одного переноса с сохранением, $t^-(f, G, S)$ – количество переносов без сохранения у f, и c_t^+ – цена одного переноса без сохранения.

Отметим, что в отличие от леммы 1 существуют деревья генов G и видов S и значения цен за одно событие, для которых парный сценарий (с переносами) не единствен.

Пример. Пусть $G = ((a, c), b)$ и $S = ((a, b), (c))$, видам a, b, c в G приписаны гены, имена которых не указаны. Обозначение (c) указывает, что в S труба, соединяющая корень с листом c, разбита на две последовательные части. Здесь имеются три временных слоя, в i-м слое находятся трубы, к которым ведет путь из суперкорня с i трубами. Вершины в G и S обозначим так же, как их клады; корневую трубу обозначим r, ребро/трубу в G или S, ведущую в лист, обозначим именем этого листа. Цены событий: $c_l = 1$, $c_d = 2$, $c_t^+ = 3$, $c_t^- = 4$. Тогда имеются два парных сценария: 1) сценарий f^* без переносов, в котором $f^*(\{a, c\}) = \{a, b, c\}$, $f^*(\{a, b, c\}) = r$, что соответствует одной дупликации $\{a, b, c\}$ и двум потерям $\langle b, \{a, b, c\}\rangle$ и $\langle b, \{a, b\}\rangle$; 2) сценарий f^* с переносами, где G' получается из G добавлением новой вершины g' на ребро a и $f(\{a, c\}) = c$, $f(\{a, b, c\}) = \{a, b, c\}$, $f(g') = a$, что соответствует одному переносу с сохранением $\{a, c\}$ и одной потере $\langle b, \{a, b\}\rangle$. Если увеличить цену за один перенос с сохранением, то останется только один сценарий – первое вложение, а если уменьшить, то снова только один сценарий – второе вложение.

Проблема. Доказать утверждение, аналогичное теореме, для более сложного функционала (10).

Замечание 3. Устраним одно недоразумение, касающееся алгоритма из [3, 4], который тесно связан с алгоритмом, изложенным в данной статье. Первая фраза в [4, пункт 3] гласит: “Время работы алгоритма пропорционально произведению числа ребер в дереве генов на число труб в дереве видов, уже разбитом на временные слои”. В [7] первое из этих чисел обозначено $|G|$, а второе – $|S'|$. Тогда время работы алгоритма из [3, 4] равно $O(|S'||G|)$. Эта оценка доказана в [3], а чуть более формально – в [4]. В точности эта оценка утверждается как основной результат в [7] (конец второго абзаца на с. 94), хотя в [7] цитируются обе работы [3, 4].

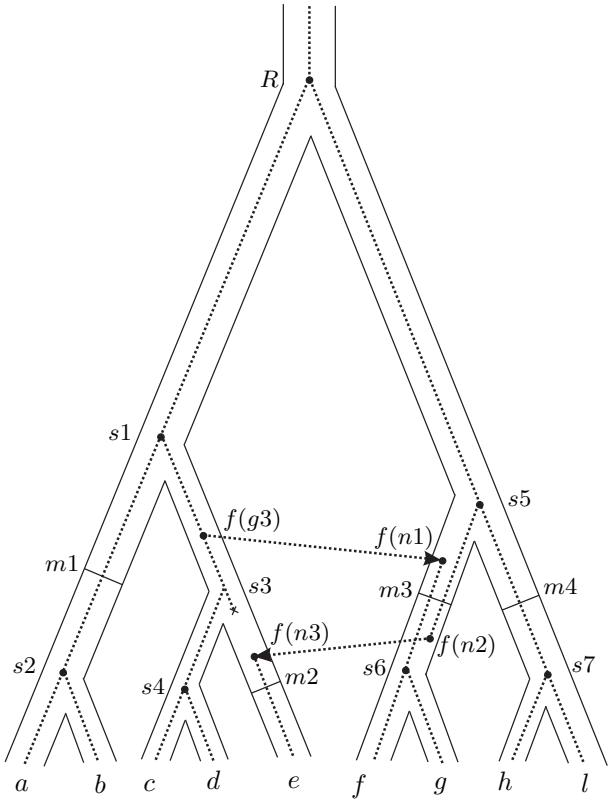

Рис. 4. Иллюстрация понятия горизонтального переноса: вложение f дерева G' в дерево S (для деревьев, показанных на рис. 3). Значения вложения f дерева G' в дерево S показаны внутри труб дерева S жирными точками, кроме значений на листьях в G', которые совпадают с соответствующими листьями в S. Значение $f(g_3)$ показано в трубе (хотя формально оно равно этой трубе) и по определению вершина g_3 соответствует событию переноса с сохранением. Стрелка проводится из $f(g_3)$ в $f(n_1)$, где n_1 – соответствующий сын вершины g_3. Значение $f(n_2)$ показано в трубе (хотя формально оно равно этой трубе) и по определению вершина n_2 соответствует событию переноса без сохранения. Стрелка проводится по тому же правилу. Потеря показана отростком с крестиком на конце. Дивергенции: $R = f(r)$, $s_1 = f(g_1)$, $s_2 = f(g_2)$, $s_4 = f(g_4)$, $s_5 = f(g_6)$, $s_6 = f(g_5)$, $s_7 = f(g_7)$. Вершина s_3 не является значением вложения f

Несмотря на некоторые эволюционистские термины, биологическое содержание которых не существенно для этой статьи, теорема и эта проблема имеют чисто математическое содержание.

ПРИЛОЖЕНИЕ

Пример работы алгоритма. Проиллюстрируем работу алгоритма на искусственном примере, в котором даны десять деревьев генов G_i, показанных на рис. 5. Эти деревья подобраны так, чтобы было легко найти для них супердерево S^*, которое показано на том же рисунке. Пусть P – стандартный набор, цена потери равна 2, а цена дупликации – 3. Здесь $V_0 = \{a, b, c, d, e\}$. Вычислим цены всех десяти двухэлементных множеств V, используя набор $\{G_i\}$. Разбиения $V = \{x\} \cup \{y\}$ и их цены, вычисленные по формуле (2), приведены в таблице, в ней также указано число t

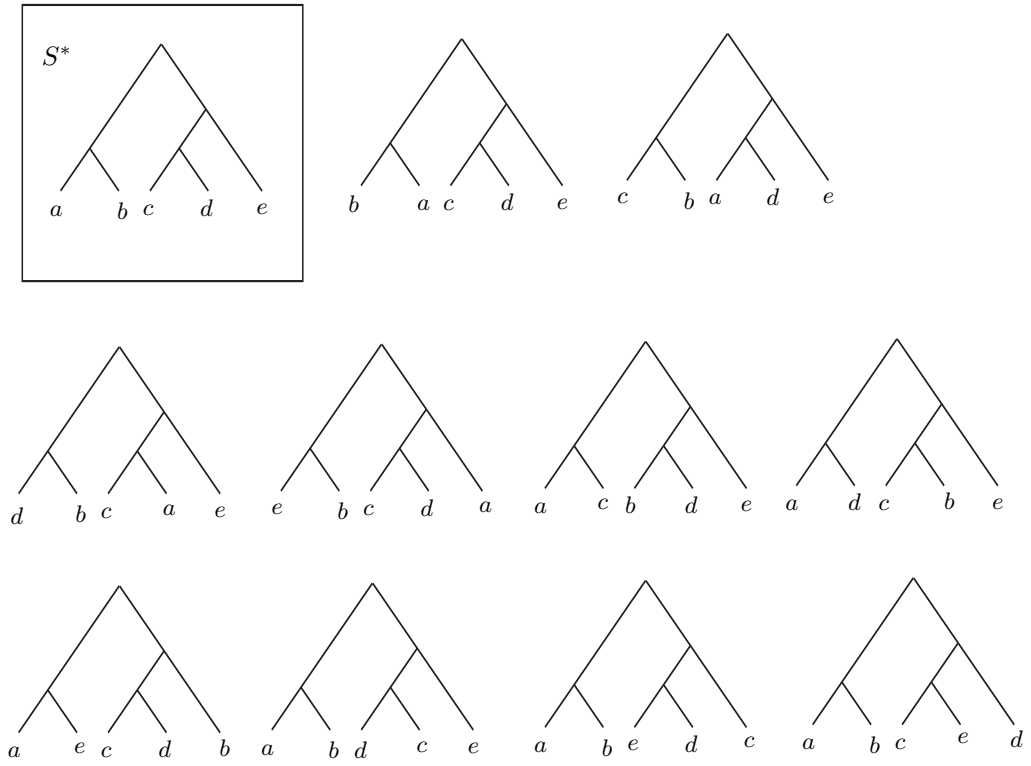

Рис. 5. Супердерево видов S^* и десять деревьев генов, которые подаются на вход алгоритма

Таблица

$V=\{x,y\}$	t	$c(V)$	$V=\{x,y\}$	t	$c(V)$
a,b	6	24	b,d	8	32
c,d	6	24	a,e	9	36
a,c	8	32	b,e	9	36
a,d	8	32	c,e	9	36
b,c	8	32	d,e	9	36

деревьев генов G_i, в которых V не является кладой. Эти деревья порождают по две потери и по 0 дупликаций, поэтому их вклад в $c(V,V_1,V_2)$ равен числу таких деревьев, умноженному на четыре. Остальные деревья дают нулевые слагаемые в $c(V,V_1,V_2)$. В результате получаем, что множества $\{a,b\}$ и $\{c,d\}$ имеют минимальную цену, т.е. для них $c(V)=24$.

Теперь рассмотрим трехэлементное множество $V=\{c,d,e\}$. С методической целью назовем то его разбиение, которое совпадает с разбиением в S^*, *стандартным*, а остальные разбиения – *нестандартными*. Конечно, наш алгоритм не использует S^*, а перебирает все разбиения. Здесь стандартным является одно разбиение $V=\{c,d,e\}$ на $V_1=\{c,d\}$ и $V_2=\{e\}$. Вычислим на нем значение $c(V,V_1,V_2)$ функционала (2). Для этого рассмотрим три случая: 1) $\{c,d\}$ – клада, а $\{c,d,e\}$ – не клада у двух деревьев генов; 2) $\{c,d,e\}$ – клада, а $\{c,d\}$ – не клада у двух деревьев генов; 3) $\{c,d\}$ и $\{c,d,e\}$ – не клады у четырех деревьев генов. Окончательно для стандартного разбиения получаем $c(V,V_1,V_2)=8+10+24+c(\{c,d\})+c(\{e\})=42+24+0=66$,

так как по индукции $c(\{c,d\}) = 24$ и $c(\{e\}) = 0$. Теперь рассмотрим нестандартное разбиение того же множества V на $V_1 = \{c,e\}$ и $V_2 = \{d\}$ – одно из двух симметричных разбиений. Вычислим значение $c(V, V_1, V_2)$. Снова рассмотрим три случая: 1) $\{c,d\}$ – клада, а $\{c,d,e\}$ – не клада у двух деревьев; 2) $\{c,d,e\}$ – клада, а $\{c,e\}$ – не клада у трех деревьев; 3) никакое из трех множеств $\{c,d\}$, $\{c,e\}$, $\{d,e\}$ не является кладой у четырех деревьев. Окончательно для нестандартного разбиения получаем $c(V, V_1, V_2) = 4 + 15 + 24 + c(\{c,e\}) + c(\{d\}) = 43 + 36 + 0 = 79$, так как по индукции имели $c(\{c,e\}) = 36$ и $c(\{d\}) = 0$. Разобраны все случаи разбиения V на две части из P, выбирается разбиение с наименьшим значением (равным 66) функционала (2), в данном случае это – стандартное разбиение. Поэтому дерево $S(\{c,d,e\})$ совпадает с поддеревом в S^*. Аналогично вычисляется значение $c(V_0, V_1, V_2)$ для множества $V_0 = \{a,b,c,d,e\}$ всех видов и его стандартного разбиения на $V_1 = \{a,b\}$ и $V_2 = \{c,d,e\}$ (оно равно 128) и для его нестандартных разбиений (среди них наименьшее значение равно 143). В итоге алгоритм выдает дерево $S(V_0)$ с ценой 128, которое совпадет с деревом S^*.

СПИСОК ЛИТЕРАТУРЫ

1. Phylogenetic Supertrees. Combining Information to Reveal the Tree of Life. Dordrecht–Boston–London: Kluwer Acad. Publ., 2004.
2. *Guigo R., Muchnik I., Smith T.F.* Reconstruction of Ancient Molecular Phylogeny // Mol. Phylogenet. Evol. 1996. V. 6. № 2. P. 189–213.
3. *Горбунов К.Ю., Любецкий В.А.* Реконструкция эволюции генов вдоль дерева видов // Молекулярная биология. 2009. V. 43. № 5. С. 946–958.
4. *Горбунов К.Ю., Любецкий В.А.* Об одном алгоритме согласования деревьев генов и видов с учетом дупликаций, потерь и горизонтальных переносов генов // Информационные процессы. 2010. V. 10. № 2. С. 140–144.
5. *Горбунов К.Ю., Любецкий В.А.* Быстрый алгоритм построения супердерева видов по набору белковых деревьев // Молекулярная биология. 2011 (в печати).
6. *Горбунов К.Ю., Любецкий В.А.* Поиск предковых генов, нарушающих согласованность деревьев белков и видов // Молекулярная биология. 2005. V. 39. № 5. С. 847–858.
7. *Doyon J.P., Scornavacca C., Gorbunov K.Yu., Szeollosi G.J., Ranwez V., Berry V.* An Efficient Algorithm for Gene/Species Trees Parsimonious Reconciliation with Losses, Duplications and Transfers // Comparative Genomics (Proc. Int. Workshop RECOMB-CG 2010. Ottawa, Canada. October 9–11, 2010). Lecture Notes Comp. Sci. V. 6398. Lecture Notes in Bioinformatics. Berlin–Heidelberg: Springer-Verlag, 2010. P. 93–108.

Горбунов Константин Юрьевич
Любецкий Василий Александрович
Институт проблем передачи информации
им. А.А. Харкевича РАН
gorbunov@iitp.ru
lyubetsk@iitp.ru

Поступила в редакцию
13.10.2010
После переработки
26.05.2011