

УДК 575.852

РЕКОНСТРУКЦИЯ ЭВОЛЮЦИИ ГЕНОВ ВДОЛЬ ДЕРЕВА ВИДОВ

© 2009 г. К. Ю. Горбунов*, В. А. Любецкий

Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, Москва, 127994

Поступила в редакцию 09.12.2008 г.

Принята к печати 20.01.2009 г.

Предложена модель и алгоритм восстановления эволюции семейства генов, описанного данным деревом генов, относительно эволюции набора видов, описанного данным деревом соответствующих видов. Согласно модели, эволюция описывается с помощью нового понятия – внутреннего дерева. Работа алгоритма иллюстрирована на примере нескольких комплексов ортологических групп белков. В качестве эволюционных событий рассмотрены видообразование, дупликация и потеря гена, горизонтальный перенос гена с сохранением копии в источнике. Горизонтальный перенос гена без сохранения копии рассматривается как сочетание переноса и потери. Модель позволяет привязать каждое эволюционное событие к определенному моменту в развитии вида.

Ключевые слова: эволюция вдоль дерева видов, вложение дерева генов в дерево видов, дупликация гена, потеря гена, горизонтальный перенос гена.

RECONSTRUCTING GENES EVOLUTION ALONG A SPECIES TREE, by K. Yu. Gorbunov, V. A. Lyubetsky (Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 127994 Russia; *e-mail: gorbunov@iitp.ru). A model and algorithm are proposed to infer evolution of a gene family, given the gene tree, with respect to species evolution, given the species tree. The model describes such evolution using the “nested tree” concept. Performance of the algorithm is verified on several orthologous protein groups. The evolutionary events inferred are: speciation, gene duplication and loss, horizontal gene transfer retaining the original gene copy. A transfer event with the loss of original gene copy is considered a combination of gene transfer and loss. The model maps gene evolution events onto the species phylogeny.

Keywords: evolution along a species tree, mapping of a gene tree into a species tree, gene duplication, gene loss, horizontal gene transfer.

ВВЕДЕНИЕ

Тема молекулярной эволюции видов и таксонов, генов и белков имеет давнюю историю и отражена в большом числе публикаций и сайтов интернета, например, в [1–3]. Построению филогенетического дерева белкового семейства посвящена обширная литература (см. [4–8] и ссылки к ним). Дерево видов строится одним из двух способов: либо по объединенным (конкатенированным) в одно выравнивание последовательностям многих генов, либо как супердерево. Последнее означает, что по заданному множеству деревьев генов ищется обобщенное дерево, в том или ином смысле наилучшим образом соответствующее каждому из этих деревьев, и оно принимается за дерево видов. Например, дерево видов ищется как минимизирующее число дупликаций генов при некотором соотношении каждого дерева генов и искомого дерева видов или как минимизирующее суммарное число дупликаций и потерь генов [9]. В [10] дерево видов ищется на основе минимизации

только числа потерь, а в [11] – как в некотором смысле “содержащее” в себе все данные деревья генов. Обычно представление о том, что одно дерево содержится в другом или соотносено с другим основано на *вложении* первого из них во второе: первое – это дерево генов G , а второе – дерево видов S . Классическим является определение одного из таких вложений дерева G в дерево S , традиционно обозначаемого α [12], которое дает информацию о числе и месте дупликаций, пропусков и потерь, а также содержит определение *цены* вложения α . Доказана связь чисел односторонних дупликаций, пропусков и потерь, которые определяются с помощью вложения α [13].

Вложение α также используется для нахождения горизонтальных переносов генов, [3]. Для поиска горизонтальных переносов, современных и предковых, – одной из наиболее популярных тем исследования, используют статистические критерии [3], дифференциально-стохастические уравнения [14], нечеткие множества [15]. Определение генов или профилей сигналов во внутренних вершинах дерева

* Эл. почта: gorbunov@iitp.ru

видов дает информацию о его эволюционно значимых ребрах [16]. Рассматривается расширение задачи поиска генов, горизонтально перенесенных в листья дерева видов, за счет рассмотрения и поиска генов, горизонтально перенесенных в предковые вершины дерева видов.

Отметим, что в свое время горизонтальные переносы учитывали на биологическом уровне, без всякой модели [17]. Идея вложения дерева генов в дерево видов предложена позднее [18], там же поставлена задача учета горизонтальных переносов, но не приведены определения и алгоритмы для построения такого вложения и, тем более, для учета переносов.

Другой популярный круг задач – реконструкция событий в эволюции нуклеотидной последовательности: замены букв, вставки и делеции [19]. К отдельному кругу задач следует отнести реконструкцию событий в эволюции целого семейства генов (точнее, КОГа – комплекса ортологичных групп белков): дубликации, потери и приобретения генов, горизонтальные переносы генов [3, 15] и, наконец, реконструкции предковых регуляторных сайтов или их характеристик по набору современных регуляторных сайтов или их характеристик вдоль данного эволюционного дерева [16, 21].

“Статистические” характеристики, такие как значения чисел дубликаций или потерь, или горизонтальных переносов и т.д. в одном геноме или в одном поддереве, наследуемом из некоторой вершины эволюционного дерева, также представляют интерес [3].

Другая тема – сравнение различных сценариев, например, допущение горизонтальных переносов против их запрещения (тогда их роль берут на себя дубликации и потери). Так можно определить среднее число дубликаций и потерь, приходящихся на один горизонтальный перенос [3]. Изучают и другие постановки задач, связанные со сравнением сценариев [20].

В последние годы особенный интерес вызывает молекулярная эволюция самих регуляторных систем, в рамках которой рассматривают различные типы регуляции: регуляцию, основанную на ДНК-белковых взаимодействиях [21–22]; регуляцию, основанную на вторичной структуре мРНК [19, 23–25]; регуляцию, основанную на конкуренции РНК-полимераз, транскрибирующих общих локусов с комплементарных цепей ДНК, [26–27, неопубликованные результаты авторов]; различные типы посттранскрипционной и посттрансляционной регуляции [23, 28].

Ниже рассмотрена, насколько нам известно, новая постановка задачи, в которой даны *дерево генов G* и *дерево видов S* , и ищется “привязка” событий, имевших место в семействе генов, представленных в G , к эволюционным событиям, представленным в S .

Эта привязка выполняется определяемым ниже *внутренним* деревом G' .

ПОСТАНОВКА ЗАДАЧИ

Эволюционный процесс состоит в видообразовании, а также в дубликации иногда многих генов в геноме вида, что приводит к образованию большого числа паралогов, каждый из которых претерпевает в дальнейшем индивидуальную эволюцию. Эта эволюция приводит к расхождению нуклеотидных последовательностей генов между собой и с последовательностью их общего предка, к смене положения гена в геноме, его специфичности, к смене регуляторного механизма, к неортологичной замене гена, к его потере и горизонтальному переносу. Ниже мы сосредоточимся на учете именно горизонтальных переносов. Алгоритм, который мы применим, относится к классу методов наибольшей взвешенной экономии. Перейдем к точным формулировкам.

Мы хотим описать модель и алгоритм, которые позволили бы изучать некоторые аспекты развития семейства генов, описанного деревом G , внутри семейства видов, описанного деревом S . Модель будет основана на рассмотрении вспомогательного дерева G' , которое мы назовем *внутренним* деревом.

Все рассматриваемые деревья являются *корневыми* и представляются *растущими вниз*, причем к деревьям G и S добавлено выходящее из корня вверх ребро. Это ребро назовем *корневым*, новую самую высокую вершину назовем *суперкорнем* (далее используется только суперкорень у G). Множество ребер в дереве видов S разбито на части, называемые *временными слоями*. Предполагается, что горизонтальный перенос гена может происходить с ребра a дерева S на *несравнимое* ребро b этого дерева. “Несравнимое” означает, что оба ребра, a и b , расположены в одном временном слое, и из a в b нельзя попасть вдоль дерева S , двигаясь только в одном направлении: от корня или к корню. Если слои заранее не заданы, то предлагаемые ниже модель и алгоритм также применимы и работают в предположении, что имеется только один временной слой, охватывающий сразу все дерево. В разделе “Алгоритмы ...”, пункт c , предложен алгоритм разбиения ребер на временные слои; он носит предварительный характер, так как основан на допущениях, биологический статус которых не ясен. Вопрос о выделении временных слоев в дереве видов S в этой работе в биологическом аспекте не исследуется.

Пусть каждый лист дерева генов G помечен именем некоторого гена, а каждый лист дерева S помечен именем некоторого вида и множеством имен генов, взятых из этого вида, это множество может быть и пустым. Деревья G и S бинарные, вспомогательное также бинарное дерево G' , которое мы сейчас рассмотрим, имеет у некоторых из его листьев

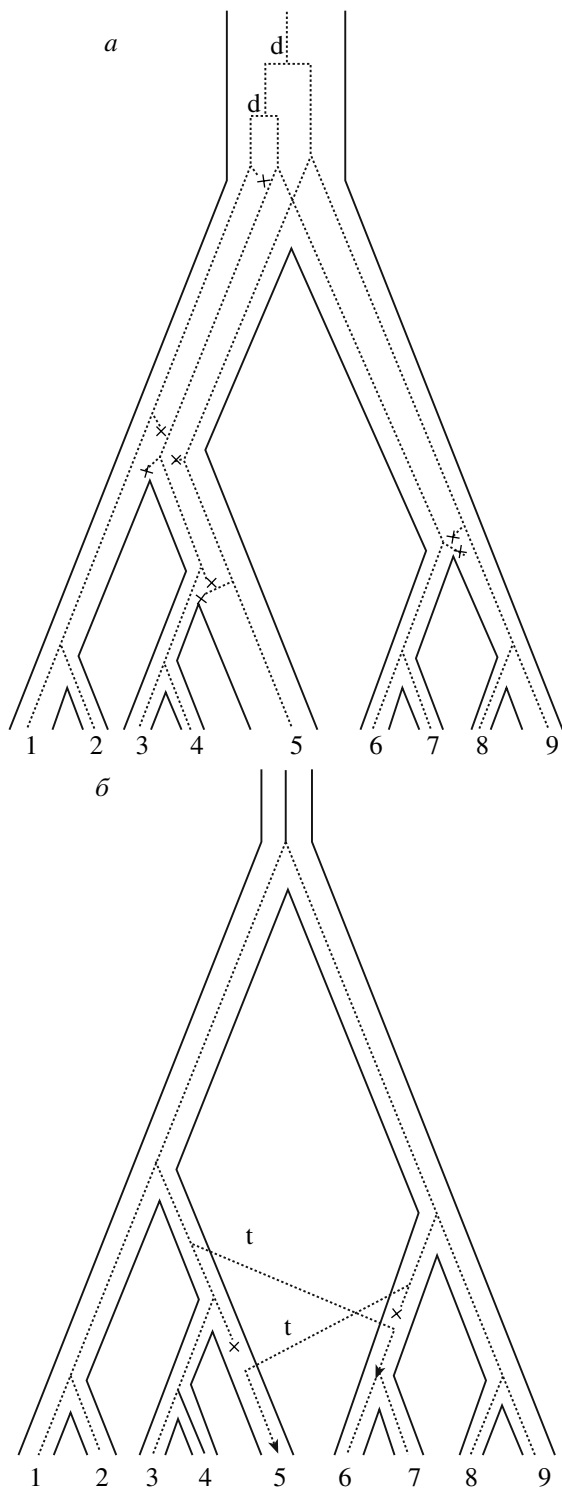


Рис. 1. Внутреннее дерево без горизонтальных переносов (а) и с горизонтальными переносами (б). Само дерево показано пунктирной линией внутри дерева видов. Его вершины, соответствующие дубликации гена, помечены буквой *d*, соответствующие видообразованию, без пометки, а соответствующие потери отмечены крестиком. Числа нумеруют листья внутреннего дерева и одновременно листья дерева видов. Ребра (в виде ломаных) указывают на горизонтальные переносы с сохранением и без сохранения копии в источнике переноса, они дополнительно помечены буквой *t*.

крестик, который означает потерю гена, приписанного этому листу.

Сейчас будет определено это дерево G' (назовем его *внутренним*), которое в случае, когда **не рассматриваются горизонтальные переносы**, по другому описывает упомянутое выше вложение α . Затем переносы будут включены в рассмотрение за счет усложнения в определении внутреннего дерева G' , и тогда оно не будет сводиться к вложению α (чуть ниже дается определение вложения α , эквивалентное оригинальному, которое мы здесь не будем напоминать). Примеры искусственных внутренних деревьев (которые служат для иллюстрации определения) без и с горизонтальными переносами показаны на рис. 1.

Представим себе ребра дерева S в виде полых труб и будем называть их трубами, считая, что труба *не включает ее начало и конец*. Вершины дерева S будем называть началом (или концом) соответствующей трубы. Рассмотрим некоторое дерево G' , расположенное внутри этих труб следующим образом. Дерево G' имеет суперкорень в корневой трубе и имеет ребра, идущие вниз целиком внутри труб. В трубе могут проходить несколько ребер, ребро должно разветвиться в развилке труб (что означает: ген удвоился в связи с видообразованием), но может разветвиться и внутри самой трубы (что означает дубликацию гена); листу этого дерева может быть приписан крестик (потеря гена); все разветвления в дереве G' , которые не относятся к видообразованию, считаются дубликациями. Рассматриваются только те G' , которые имеют минимальную величину суммы значений чисел дубликаций и потерь, взятых каждое с некоторым положительным весом. Свойства G' и алгоритмы, описанные далее, не зависят от значений этих весов (конечно, для биологически содержательного счета их нужно разумно выбрать). Задача состоит в том, чтобы найти внутреннее дерево G' в данном дереве видов S .

Такое G' после удаления в нем листьев, помеченных крестиками, изоморфно исходному дереву генов G (при этом имена генов, указанных в листьях, должны сохраняться). Удаление листьев, помеченных крестиком, означает и удаление их последнего общего предка, включая перестройку дерева G' , необходимую для сохранения его бинарности. Такое G' единственно в том смысле, что два внутренних дерева изоморфны между собой (а без крестиков – и с G), а две изоморфных вершины располагаются в одной и той же трубе или вершине дерева S . Говоря об изоморфизме G' и G , всегда подразумевается, что в G' удалены крестики, как указано выше. Таким образом, дерево G' существует и единственно; его существование вытекает из очевидной корректности описанного в разделе “Алгоритмы ...”, пункт *a*, алгоритма его построения.

Внутреннее дерево G' определяет упомянутое выше вложение α следующим образом: вершине x

из G сопоставляется соответствующая ей по изоморфизму вершина u в G' , а последней – вершина $\alpha(x)$ в S , которая является концом трубы, содержащей u . Если u совпадает с самим этим концом, то также $\alpha(x) = u$. Последняя оговорка связана с тем, что труба рассматривается без ее начала и конца.

Теперь рассмотрим возможные **горизонтальные переносы**. Для этого в бинарном *внутреннем дереве* G' разрешается провести ребро (“стрелку”) от одной его вершины a , лежащей внутри некоторой трубы, к другой его вершине b , лежащей уже в несравнимой трубе или в ее конце; такое ребро обозначает *горизонтальный перенос гена*. *Переносом гена* (точнее, переносом гена с сохранением копии) называется перенос, при котором в трубе, содержащей вершину a , сохраняется копия гена из этой вершины, в то время как другая копия перенеслась по стрелке в несравнимую трубу. Иногда рассматривают событие *переноса гена без сохранения* его копии. Это – последовательность двух событий: переноса гена с сохранением копии с последующей потерей этой копии. На рис. 1б и рис. 2 показаны искусственные случаи переноса гена без сохранения и с сохранением его копии.

По-прежнему рассматриваются только те деревья G' , на которых достигает минимума функция, равная сумме с некоторыми положительными весами значений чисел дупликаций, потерь и горизонтальных переносов с сохранением исходного гена. Это дерево по-прежнему называется *внутренним*. Веса называются *ценой* каждого из трех событий. В результатах счета, которые приведены в разделе “Результаты тестирования ...”, они принимают следующие значения: за каждую потерю полагаем цену 2, за каждую дупликацию – цену 3, за каждый перенос с сохранением – цену 11 (переносу без сохранения соответствует цена 13). Эти численные значения выбраны нами из наблюдений частот соответствующих событий ([29] и неопубликованные данные авторов), хотя вопрос о выборе значений цен далек от своего обоснования.

Такое дерево G' существует, но уже не единственно. Его существование следует из очевидной корректности описанного в разделе “Алгоритмы ...”, пункт b алгоритма его построения.

АЛГОРИТМЫ ПОСТРОЕНИЯ ВНУТРЕННЕГО ДЕРЕВА И ВРЕМЕННЫХ СЛОЕВ

a. Случай, когда не учитываются горизонтальные переносы

Будем строить внутреннее дерево G' как множество значений изоморфизма из G на G' . По аналогии с α этот изоморфизм будем называть *вложением*. Начнем с корня $r(G)$ дерева G и будем двигаться вдоль дерева G от его корня к листьям. Из построе-

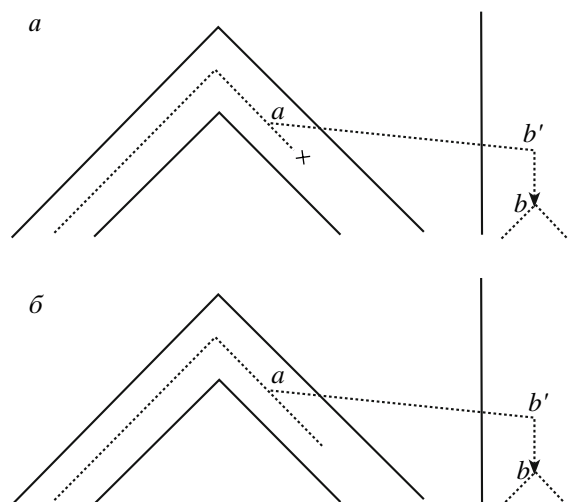


Рис. 2. Горизонтальные переносы без сохранения (а) и с сохранением (б) копии гена в источнике переноса. Внутренние деревья показаны пунктирной линией. Ребра, соответствующие переносу, показаны составленными каждое из двух отрезков.

ния очевидно, что существует ровно одно G' , изоморфное G .

Пусть r – вершина внутреннего дерева G' , которая сопоставлена корню $r(G)$ дерева G . Для r возможны следующие три расположения: (1) r совпадает с $r(S)$ – корнем дерева S ; (2) r лежит внутри корневой трубы; (3) r лежит ниже конца корневой трубы (с указанием одного из двух направлений). Возможны следующие три соотношения двух пар (A, B) и (C, D) , где A и B – множества генов, указанных в листьях двух поддеревьев в G с корнями – сыновьями $r(G)$, а C и D – множества генов, указанных в листьях двух поддеревьев в S с корнями – сыновьями $r(S)$. Тогда:

- 1) если $(A = C, B = D)$ или $(A = D, B = C)$, то принимаем (1);
- 2) если $((A \neq C \ \& \ A \neq D)$ или $(B \neq C \ \& \ B \neq D))$ и $(C \neq \emptyset \ \& \ D \neq \emptyset)$, то принимаем (2);
- 3) если $((A \neq C \ \& \ A \neq D)$ или $(B \neq C \ \& \ B \neq D))$ и $(C = \emptyset$ или $D = \emptyset)$, то принимаем (3) и продолжаем корневое ребро внутреннего дерева G' в непустую часть, а ребро, идущее в пустую часть, оканчиваем крестиком (потерей).

Легко видеть, что в каждом из трех случаев выбор иного варианта из (1–3) породит лишние дупликации и потери или не позволит построить вложение.

Далее изоморфизм и внутреннее дерево G' строятся по индукции: рассмотрим ребро e' дерева G' , которое по построению уже изоморфно ребру e дерева G и находится в трубе d дерева S . Пусть r – вершина дерева G' , которая сопоставлена концу a ребра e . Возможны следующие три расположения r : (1) вершина r совпадает с a (видообразование); (2) вершина r лежит в трубе d (дупликация); (3) вершина r лежит

ниже трубы d (потеря). Рассуждаем, как сказано выше для начального шага индукции с заменой $r(G)$ на a , $r(S)$ на d , C на $C \cap (A \cup B)$ и D на $D \cap (A \cup B)$. Используя индуктивное предположение $(A \cup B) \subseteq (C \cup D)$, продолжаем построение до достижения листьев в G .

в. Случай, когда учитываются возможные горизонтальные переносы

Для ребра e из G обозначим $T(e)$ поддерево в G , которое начинается с e , так что e для $T(e)$ служит корневым ребром. Составляется список вложений дерева $T(e)$ в S для всех пар (ребро e из G , труба d из S), для каждой пары вложение ищется среди вложений дерева $T(e)$ в S , у которых начало e помещается в d . Искомое вложение соответствует паре (корневое ребро в G , корневая труба в S). При составлении списка сначала рассматривают ребра e , расположенные дальше от корня, а при фиксированном e сначала рассматривают трубы d , лежащие в более поздних временных слоях. Обозначим x конец ребра e в G и y – его образ при вложении G в S .

Начальным шагом индукции является случай, когда e и d ведут соответственно в листья $l_e = x$ и l_d . Тогда цена вложения равна 0, если среди генов в l_d имеется ген l_e , иначе цена вложения равна цене горизонтального переноса без сохранения. Действительно, единственный способ дотянуть ветвь до гена l_e в S состоит в том, чтобы провести ее в трубу, идущую в лист, содержащий l_e . В любом случае y совпадает с l_d .

Индуктивный шаг для очередной пары $\langle e, d \rangle$. Обозначим e_1 и e_2 ребра, на которые разветвляется ребро e в вершине x (если она не лист), и, соответственно, обозначим d_1 и d_2 две трубы, выходящие из d . Выберем одну из следующих возможностей.

(1) Пусть ребро e не оканчивается листом. Вершина x соответствует дубликации внутри d . По предположению индукции вложения, соответствующие парам $\langle e_1, d \rangle$ и $\langle e_2, d \rangle$, уже построены. Их соединения дает вложение для $\langle e, d \rangle$, цена которого равна сумме цен для $\langle e_1, d \rangle$, $\langle e_2, d \rangle$ и цены дубликации. Тогда вершина y располагается внутри трубы d .

(2) Пусть ребра e и d не оканчиваются листьями. Вершина x соответствует разветвлению трубы d . По предположению индукции вложения, соответствующие парам $\langle e_1, d_1 \rangle$ и $\langle e_2, d_2 \rangle$ и парам $\langle e_1, d_2 \rangle$ и $\langle e_2, d_1 \rangle$, уже построены. Их соединения дает два возможных случая для $\langle e, d \rangle$, цены которых равны соответственно сумме цен для $\langle e_1, d_1 \rangle$, $\langle e_2, d_2 \rangle$ и для $\langle e_1, d_2 \rangle$, $\langle e_2, d_1 \rangle$. Тогда вершина y совпадает с концом трубы d .

(3) Пусть труба d не оканчивается листом. Вершина x не соответствует событию внутри трубы d или ее разветвлению, а ближайшее событие на ребре e – потеря. В этом случае ребро e проходит разветвление трубы d в S , сворачивая в одну из труб – d_1 или d_2 (рассматриваются оба варианта, пусть для опреде-

ленности ребро e продолжается в трубе d_1), а в другой трубе происходит потеря (в G' ребру e соответствует цепь из нескольких последовательных ребер). По предположению индукции вложение, соответствующее паре $\langle e, d_1 \rangle$, уже построено. Цена соответствующего вложения для $\langle e, d \rangle$ равна сумме цены для $\langle e, d_1 \rangle$ и цены потери. Расположение вершины y дано в индуктивном предположении.

(4) Пусть ребро e не оканчивается листом. Вершина x соответствует горизонтальному переносу с сохранением внутри d . Рассматривают два варианта: перенос задается ребром e_1 или ребром e_2 (пусть для определенности e_1). Перебираем все трубы d_1 в S , в которые возможен перенос с трубы d . По предположению индукции вложения, соответствующие парам $\langle e_1, d_1 \rangle$ и $\langle e_2, d \rangle$, уже построены. Их соединение дает вложение для $\langle e, d \rangle$, цена которого равна сумме цен для $\langle e_1, d_1 \rangle$, $\langle e_2, d \rangle$ и цены переноса. Тогда вершина y располагается внутри трубы d .

(5) Вершина x не соответствует ни событию внутри трубы d , ни ее разветвлению, а ближайшее событие на ребре e – горизонтальный перенос. В этом случае с трубы d осуществляется горизонтальный перенос без сохранения. Перебираем все трубы d_1 в S , в которые возможен перенос с трубы d . Заметим, что из трубы d_1 ребро e не может еще раз перенестись без сохранения в некоторую трубу d_2 , поскольку “дешевле” сразу перенестись в d_2 (случай, когда d_2 сравнима с d , сейчас запрещен, см. замечание ниже). Таким образом, в трубе d_1 для e остаются лишь возможности 1, 2, 3 и 4. Случаи 1, 2 и 4 связаны с разветвлением e на e_1 и e_2 , что уже позволяет применить предположение индукции и установить положение вершины y внутри d_1 (случаи 1 и 4) или в конце d_1 (случай 2). Отметим, что случай 4 можно не рассматривать, так как, если произошел перенос с сохранением из d_1 в d_2 , то такую же цену будет иметь эквивалентный ход событий, когда сначала на e произошел перенос в d_1 с сохранением, а затем сохранившаяся копия перенеслась с e в d_2 без сохранения. В случае 3 логически возможны два варианта: если прошло разветвление трубы d_1 и свернуло в трубу d_3 , лежащую в том же временном слое, что и d_1 , и несравнимую с d , то дешевле сразу перенестись в d_3 (тем самым, этот вариант запрещен), а если труба d_3 лежит в более позднем временном слое (либо ниже d), то применяем предположение индукции (в том числе и относительно расположения вершины y).

Легко видеть, что не более чем за кубическое от числа генов время алгоритм строит внутреннее дерево G' и изоморфизм G и G' . Этот изоморфизм называется *вложением*.

На рис. 3 показана схема алгоритма из этого пункта b для случая 4. Алгоритм из пункта a может рассматриваться как частный случай алгоритма из

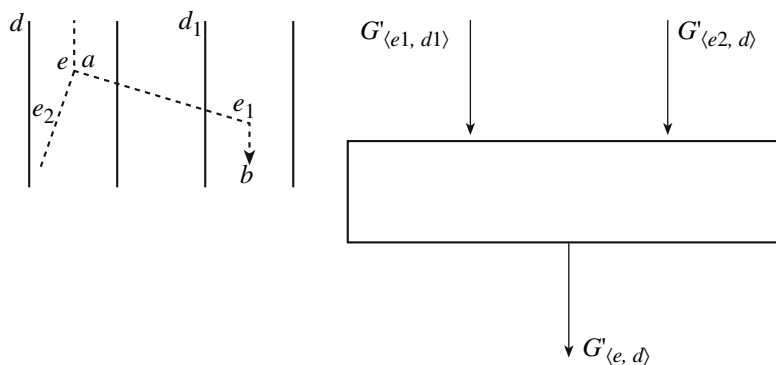


Рис. 3. Схема алгоритма построения внутреннего дерева G' . Показан случай 4 из пункта b раздела “Алгоритмы ...”. Содержание прямоугольника (одного шага алгоритма) таково: из развилки ребер e_1 и e_2 вверх выпускается ребро e , выполняется соединение двух деревьев, при этом их вершины становятся прикорневыми и появляется новая вершина – корень нового дерева. Алгоритм ведет индукцию по паре $\langle e, d \rangle$: пусть алгоритм уже построил деревья $G'_{\langle e_1, d_1 \rangle}$ и $G'_{\langle e_2, d \rangle}$. Дерево $G'_{\langle e, d \rangle}$ искомое. Показан случай горизонтального переноса с сохранением из трубы d в трубу d_1 по стрелке $e_1 = \overrightarrow{a, b}$ (случай 4 в том же разделе).

пункта b : если положить цену горизонтального переноса достаточно большой, то алгоритм из пункта b выдаст (единственное) вложение без переносов.

Замечания. При определении цены горизонтального переноса считаем, что разбиение труб на временные слои таково, что суммарная цена перехода из трубы a в ее трубу-потомок b , заключающегося в переносе из a в некоторую несравнимую трубу c с последующим переносом из c в b , велика по сравнению с суммарной ценой перехода из трубы a в ее трубу-потомок b по нисходящей в дереве S вместе с ценой всех потерь при этом переходе. Это выполняется, если разбиение на временные слои не включает в один слой двух сравнимых труб. В частности, описанный ниже алгоритм выделения временных слоев обладает этим свойством. Если сказанное не выполняется, то наш алгоритм требует дополнительных пунктов к приведенному выше описанию.

Иногда возникает необходимость допускать горизонтальные переносы извне рассматриваемого белкового семейства (КОГа). Для этого в дерево видов удобно добавить аутгруппу, т.е. трубу, ведущую от корня в лист, помеченный пустым множеством генов. Тогда переносы из этой трубы будут восприниматься как переносы извне. Наш алгоритм обобщается на этот случай.

Часто исходное дерево генов G бывает бескорневым. В этом случае алгоритм перебирает все его ребра, ставит на каждом из них корень и строит описанное вложение. По качеству положения вложения алгоритм определяет возможное положение корня в исходно бескорневом дереве.

При разбиении труб на временные слои иногда естественно разбить длинную трубу на несколько коротких труб, отсюда в дереве видов возникают вершины, имеющие лишь одного сына. Описанные

выше модель и алгоритм легко обобщаются на этот случай.

с. Вычисление временных слоев, если они не заданы

Идя от листьев к корню, вычислим в условных единицах время $r(v)$ от каждой вершины v дерева видов до соответствующего листа. После этого к одному временному слою относим трубы, лежащие в одном временном интервале. При этом могут встретиться столь длинные трубы, что они не умещаются в один временной интервал, тогда такая труба разбивается на новые участки – трубы, каждая из которых попадает в свой интервал; при этом образуются ребра в S с одним сыном. Согласно замечанию в конце предыдущего пункта, все сказанное выше применимо к такому дереву S .

Итак, нужно построить функцию $r(v)$. Если v – лист, то положим $r(v) = 0$. Пусть по предположению индукции вычислены расстояния $r(v_1)$ и $r(v_2)$ для двух сыновей v_1 и v_2 некоторой вершины v . Возможно, что они вычислены в разных масштабах: для v_1 расстояние вычислено в единицах t_1 и соответственно равно $a_1 t_1$, а для v_2 – в единицах t_2 и равно $a_2 t_2$.

Нужно вычислить $r(v)$. Пусть длина ребра e_1 от v до v_1 равна $x t_1$, а длина ребра e_2 от v до v_2 равна $y t_2$, где x и y – искомые величины. Имеем очевидное равенство

$$(a_1 + x)t_1 = (a_2 + y)t_2, \quad (1)$$

и будем исходить из следующего принципа: отношение длин ребер e_1 и e_2 обратно пропорционально отношению средних количеств f_1 и f_2 разветвлений, встретившихся, если идти из v по e_1 до какого-то ли-

ста (усреднение проводится по листьям), и аналогично по e_2 . Имеем:

$$xt_1f_1 = yt_2f_2. \quad (2)$$

Рассмотрим следующие случаи.

1) Обе вершины v_1 и v_2 являются листьями. Тогда полагаем $r(v) = 1$.

2) Вершина v_1 – лист, а v_2 – не лист, т.е. $a_1 = 0, a_2 > 0$. Обозначим $k = f_2/f_1$. В этом случае равенства (1) и (2) позволяют сразу выразить длину z ребра (v, v_1) в единицах t_2 : если $xt_1 = zt_2$, то $zt_2 = a_2t_2 + yt_2$; $z = a_2 + y$; $ky = a_2 + y$; $y = a_2/(k - 1)$.

3) Вершина v_2 – лист, а v_1 – не лист, т.е. $a_1 > 0, a_2 = 0$. Этот случай симметричен случаю 2.

4) Обе вершины v_1 и v_2 не листья. В этом случае из соотношений (1) и (2) имеем $(a_1 + x)yf_2 = (a_2 + y)xf_1$. Это равенство с неизвестными (переменными) x и y на плоскости определяет кривую (гиперболу или прямую). Чтобы найти на ней решение, найдем сначала точку (x^*, y^*) вне кривой, а затем спроектируем ее на кривую. Для этого применим сформулированный выше принцип обратной пропорциональности, чтобы найти отношения длин ребер e_1 к e_{11} и e_1 к e_{12} , где e_{11} и e_{12} – ребра, соединяющие v_1 со своими сыновьями v_{11} и v_{12} . Пусть f_{11} и f_{12} – средние числа разветвлений на пути из вершины v в лист по ребрам e_1 , e_{11} и соответственно по e_1 и e_{12} . Тогда две приближенные оценки числа x – это k_1l_1 и k_2l_2 , где $k_1 = f_{11}/f_1$, $k_2 = f_{12}/f_1$, $l_1 = a_1 - r(v_{11})$, $l_2 = a_1 - r(v_{12})$. Взяв среднее геометрическое этих двух оценок, получим x^* и аналогично получим y^* :

$$x^* = \frac{\sqrt{f_{11}f_{12}(a_1 - r(v_{11}))(a_1 - r(v_{12}))}}{f_1},$$

$$y^* = \frac{\sqrt{f_{21}f_{22}(a_2 - r(v_{21}))(a_2 - r(v_{22}))}}{f_2}.$$

где обозначения в формуле для y^* аналогичны обозначениям для x^* .

В качестве решения (x, y) возьмем проекцию точки (x^*, y^*) на упомянутую кривую, которая находится стандартным алгоритмом.

Используя равенство (1), найдем отношение $t_1/t_2 = (a_2 + y)/(a_1 + x)$. Теперь в одном из поддеревьев, скажем с корнем v_1 , перейдем от масштаба, связанного с единицами измерения t_1 к масштабу, связанному с единицами t_2 , и положим $r(v) = a_2 + y$. В равной мере можно в поддереве с корнем v_2 перейти от масштаба единиц t_2 к масштабу единиц t_1 и положить $r(v) = a_1 + x$ (в алгоритме каждый раз выбирается тот из двух вариантов, который позволяет избежать слишком малых и слишком больших величин).

РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ АЛГОРИТМА И ИХ ОБСУЖДЕНИЕ

Приведем результаты, полученные с использованием этого алгоритма, согласно нашей модели, и сравним их с результатами работы *совсем других* алгоритмов: алгоритма нахождения горизонтальных переносов генов в листья [30] и алгоритма для определения переносов в предковые вершины дерева видов [15]. Напомним, что первый метод [30] основан на двух критериях выделения горизонтально перенесенного гена – листа в дереве генов G : один состоял в том, что удаление этого листа приводило к наибольшему уменьшению цены вложения, а другой – в том, что при отображении α окрестность этого листа располагалась в дереве видов S далеко от самого листа. Результаты предложенного здесь метода и алгоритмов из [15, 30] в одних случаях подтверждают друг друга, а в другом расходятся, этот последний случай обсуждается ниже.

Использованы следующие виды.

Археи: *Archaeoglobus fulgidus* (Afu), *Halobacterium sp. NRC-1* (Hbs), *Methanococcus jannaschii* (Mja), *Methanobacterium thermoautotrophicum* (Mth), *Thermoplasma acidophilum* (Tac), *T. volcanium* (Tvo), *Pyrococcus horikoshii* (Pho), *P. abyssi* (Pab), *Aeropyrum pernix* (Ape), *Sulfolobus solfataricus* (Sso). **Грамположительные бактерии:** *Streptococcus pyogenes* (Spy), *Bacillus subtilis* (Bsu), *B. halodurans* (Bha), *Lactococcus lastis* (Lla), *Staphylococcus aureus* (Sau), *Ureaplasma urealyticum* (Uur), *Mycoplasma pneumoniae* (Mpn), *M. genitalium* (Mge). **α -Протеобактерии:** *Mesorhizobium loti* (Mlo), *Caulobacter crescentus* (Ccr), *Rickettsia prowazekii* (Rpr). **β -Протеобактерии:** *Neisseria meningitidis* MC58 (Nme). **γ -Протеобактерии:** *Escherichia coli* K12 (Eco), *Buchnera sp. APS* (Buc), *Pseudomonas aeruginosa* (Pae), *Vibrio cholerae* (Vch), *Haemophilus influenzae* (Hin), *Pasteurella multocida* (Pmu), *Xylella fastidiosa* (Xfa). **ϵ -Протеобактерии:** *Helicobacter pylori* (Hpy), *Campylobacter jejuni* (Cje). **Хламидии:** *Chlamydia trachomatis* (Ctr), *C. pneumoniae* (Cpn). **Спирохеты:** *Treponema pallidum* (Tra), *Borrelia burgdorferi* (Bbu). **Другие:** *Deinococcus radiodurans* (Dra), *Mycobacterium tuberculosis* (Mtu), *Synechocystis* (Syn), *Aquifex aeolicus* (Aae), *Thermotoga maritime* (Tma).

Используется дерево видов, показанное на рис. 4 и построенное с помощью алгоритма TIQMAX [31]. На рис. 5а и 5б приведены деревья двух КОГов, у каждого листа этих деревьев в квадратных скобках указано название вида, а затем сокращенное название гена, который взят из этого вида; на всех рисунках числа нумеруют вершины дерева, корень в дереве генов помечен нулем. Сами деревья получены на основе стандартных алгоритмов PhyloBayes и PhyML [2].

На рис. 5б и 6б для этих двух КОГов показаны фрагменты внутреннего дерева G' , найденного нашим алгоритмом; вершины внутреннего дерева также нумеруются числами, но шрифтом меньшего

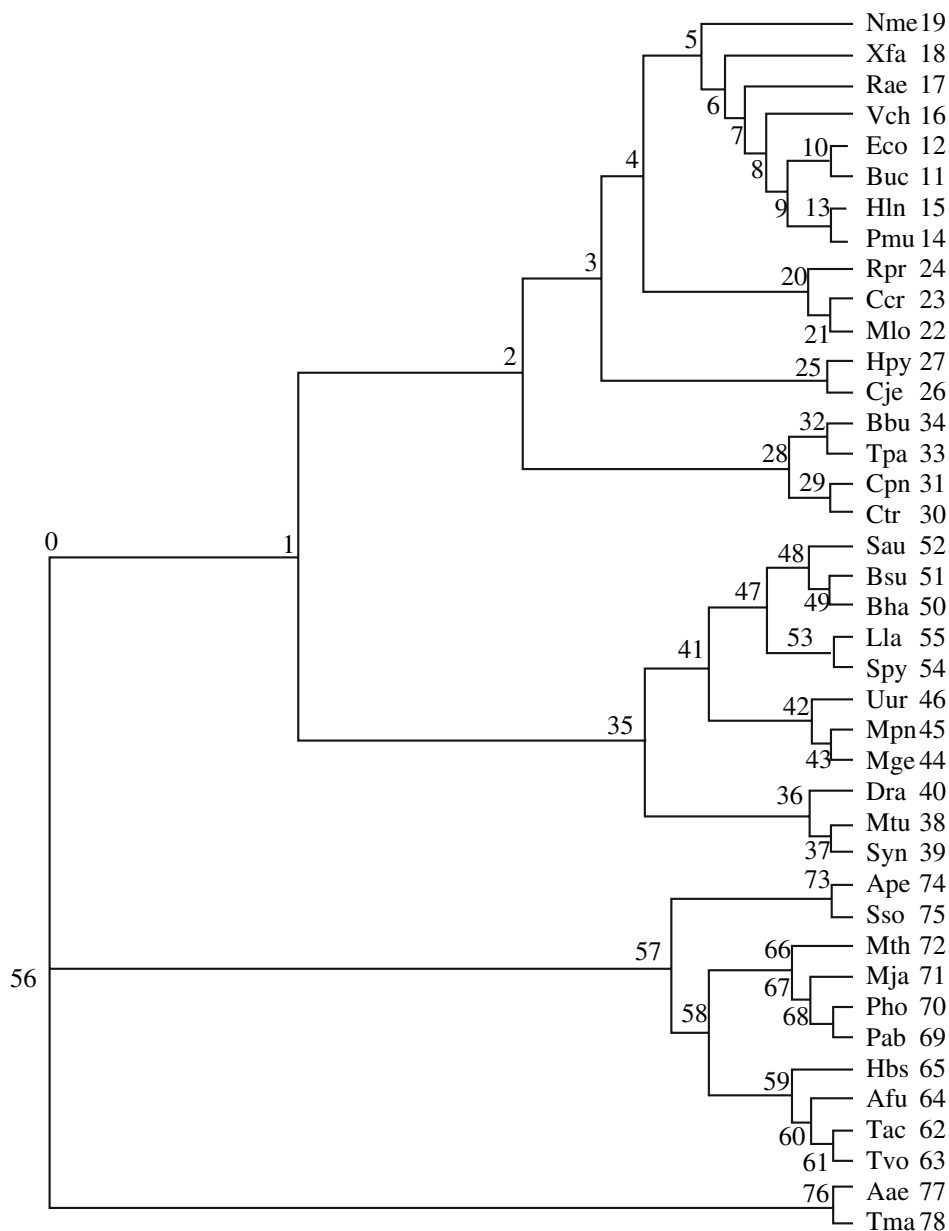


Рис. 4. Дерево видов. Названия видов и их сокращенные обозначения приведены в начале раздела “Результаты тестирования ...”. Числа нумеруют вершины дерева, корень помечен нулем.

размера, а листья внешнего дерева помечены сокращенными именами видов. В тексте при описании внутреннего дерева перед скобками указан номер трубы (который по определению совпадает с номером ее конца), а в скобках перечислены вершины внутреннего дерева, соответствующие генам, которые в этой трубе дублировались или потерялись (число штрихов указывает номер копии исходного гена, которая потерялась).

1) COG0012 (гипотетическая ГТРаза). Дерево КОГа изображено на рис. 5а, часть внутреннего дерева – на рис. 5б. Алгоритм, взятый из [30], предсказал, что хламидии являются предполагаемым ис-

точником горизонтального переноса гена *bu191* в бактерию *Buchnera aphidicola* (группа γ -протеобактерий). Там же с рядом оговорок предполагалось, что ген *slu0245* горизонтально перенесен в геном *Synechocystis* sp. из спирохет. Значимость второго предсказания характеризовалась несколькими числами, и все они были заметно меньше, чем для гена *bu191*.

Наш алгоритм предсказал, что в эволюции данного КОГа было:

9 дубликаций: 0(0,20); 1(24,25,58); 2(26); 8(42); 20(53); 58(1);

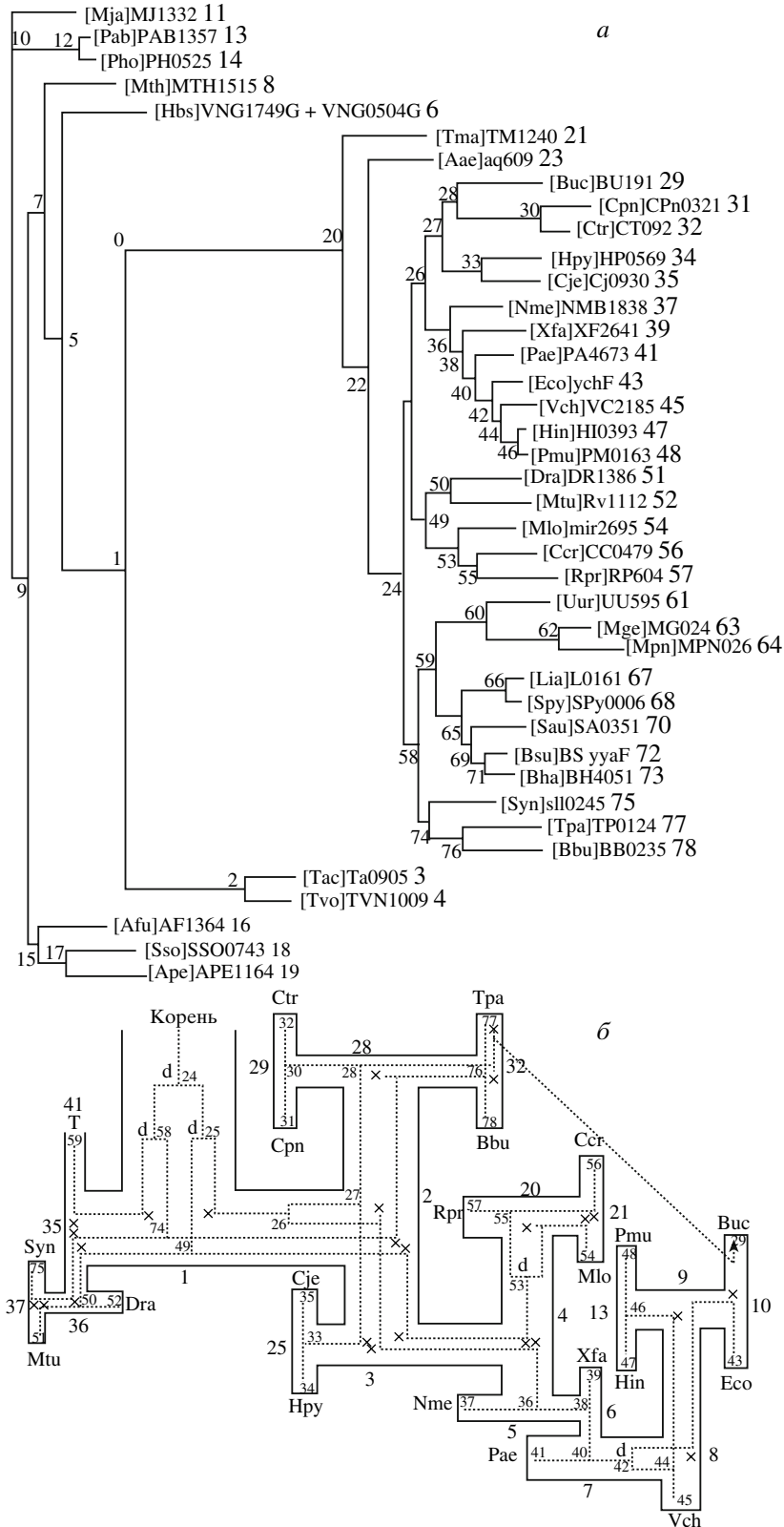


Рис. 5. Дерево КОГа COG0012 гипотетической ГТРАЗы (*a*) и фрагмент его вложения (*б*) в дерево видов. *a* – Числа нумеруют вершины дерева, корень помечен нулем; в квадратных скобках перед обозначением гена приведено название вида, которому принадлежит ген. *б* – Обозначения, как на рис. 1; вершины дерева видов помечены вне труб, внутренние вершины – числами, листья – сокращенными названиями видов; номера вершин внутреннего дерева помечены числами внутри труб, шрифтом меньшего размера; дерево, растущее из вершины, помеченной буквой Т, не показано.

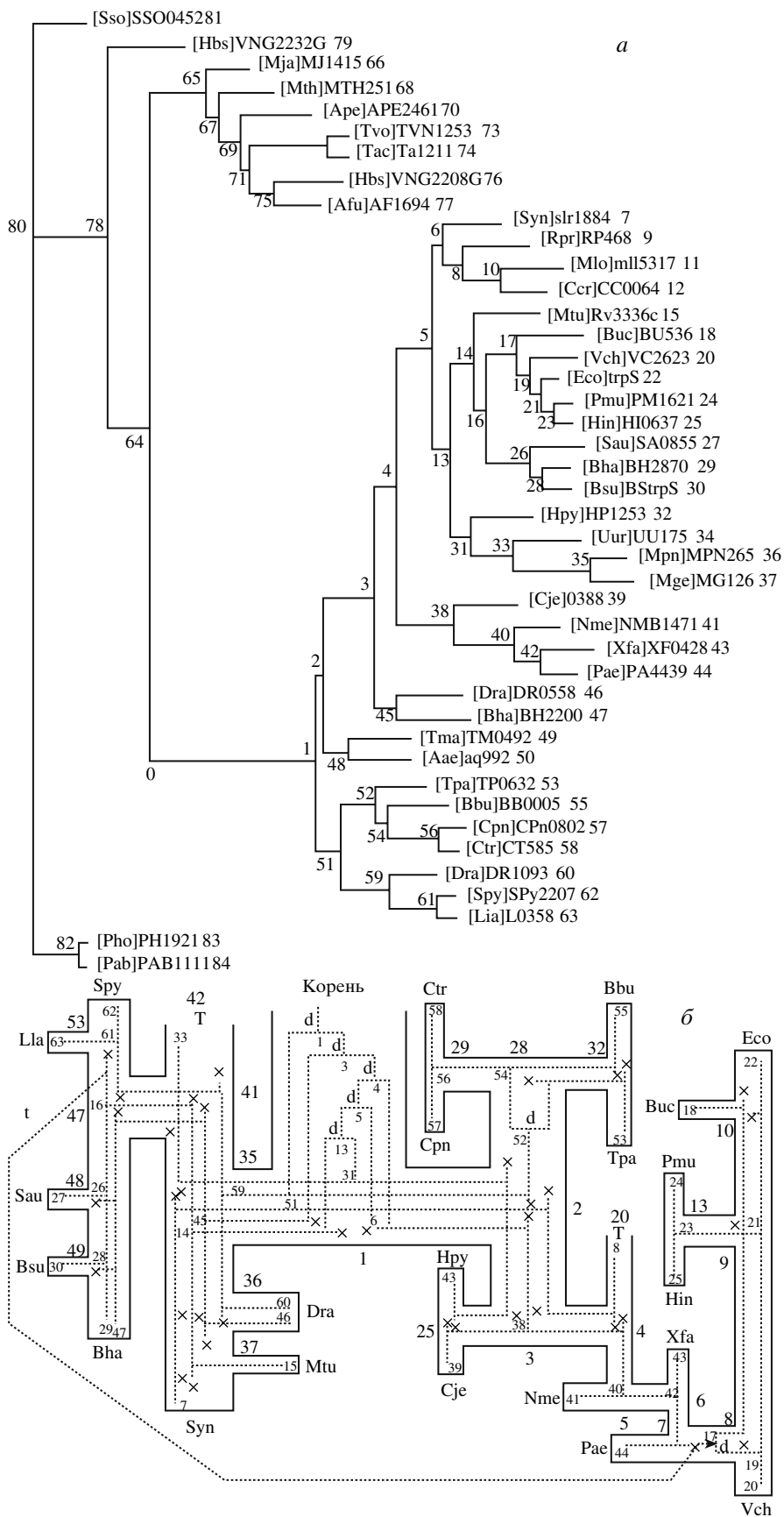


Рис. 6. Дерево КОГа COG180 триптофанил-тРНК-синтетазы (а) и фрагмент его вложения (б) в дерево видов. а – Обозначения, как на рис. 5а; б – обозначения, как на рис. 5б.

38 потеря: 1(1',21'); 2(59'); 3(76'); 4(33'); 5(53'); 10(46'); 11(43'); 13(43'"); 16(43'"); 20(36'); 22(56'); 23(54'); 24(54'"); 25(36',53'"); 28(36'","53'"); 29(76'"); 34(29'); 35(26'); 36(59'"); 38(75'); 39(52'); 40(75'"); 41(50',75'"); 57(21'',23'"); 60(6'); 61(16'); 64(2'); 65(2'"); 66(2'"); 73(1'"); 76(1'"); 77(21'"); 78(23'");

два горизонтальных переноса с сохранением гена 15 из трубы 67 в трубу 60; гена 17 из трубы 60 в трубу 73;

один горизонтальный перенос без сохранения гена 29 из трубы 33 (или в случае другого внутреннего дерева этот перенос произошел из смежной трубы 34) в трубу 11.

Мы видим, что последний перенос соответствует ранее предсказанному переносу гена *bu191*. Напротив, ранее предсказанный перенос гена *sl10245* не подтверждается: в рамках настоящей модели заключаем, что этот "перенос" является следствием трех древних дубликаций (лист Syn на рис. 5б). В [30] это предсказание имело существенно более слабые показатели значимости и в той работе не было возможности сравнить перенос со следствием дубликаций.

Предложенный здесь алгоритм также нашел два новых потенциальных переноса с сохранением, которые произошли внутри группы архей. Это перенос общего предка генов *AF1364*, *SSO0743* и *APE1164* из общего предка видов *Mja*, *Pho* и *Pab* в общий предок видов *Afu*, *Tac* и *Tvo*, за которым (после дубликации в той же трубе) последовал перенос общего предка двух последних генов в общий предок видов *Sso* и *Ape*.

На рис. 5б показан фрагмент построенного вложения, ограниченный поддеревом дерева видов с корнем в вершине 1 и, соответственно, поддеревом дерева КОГа с корнем в вершине 24 (тем самым не показаны обособленная группа архей и древние виды *Tma* и *Aae*, положение которых и в дереве видов не вполне однозначно). Вложение поддерева дерева КОГа с корнем 59 в поддерево дерева видов с корнем 41 оказалось тождественным (т.е. без нетривиальных событий) и не показано (вложение корня этого поддерева помечено буквой "Т"). Буквой "d" обозначены дубликации, единственный горизонтальный перенос указан стрелкой и буквой "t". Потери гена отмечены крестиками.

Здесь имеются два внутренних дерева *G'* с одним и тем же минимальным значением цены вложения, различающихся тем, из какой трубы, *Tra* или *Vbu*, произошел указанный перенос (на рис. 5б показан вариант с *Tra*).

2) COG0180 (триптофанил-тРНК-синтаза). Дерево КОГа изображено на рис. 6а. Алгоритм, приведенный в [3, 15], указал на возможный горизонтальный перенос между предками {*Bha*, *Bsu*, *Sau*} и {*Vch*, *Eco*, *Vuc*, *Hin*, *Rmu*}. Наш алгоритм предсказал, что в эволюции данного КОГа было:

10 дубликаций: 0(0,1); 1(3); 8(17); 28(52); 57(64,65,67,78); 59(71);

46 потеря: 1(64'); 2(45'); 3(52'); 8(44'); 11(22'); 12(18'); 13(18'"); 16(18'"); 20(40'); 26(32'); 27(39'); 28(38'); 29(53'); 33(55'); 34(53'"); 37(46',60'); 38(7'); 42(47',61'); 48(61'"); 51(47'"); 52(47'"); 53(47'"); 56(51'); 57(48'); 59(66',68',82'); 60(79'); 61(77'); 64(72'); 65(72'"); 66(71',79'"); 67(68'"); 68(66'"); 71(82'"); 72(66'","82'"); 73(66'","68'","79'"); 74(81'); 75(70'); 76(64'");

два горизонтальных переноса с сохранением гена 17 из трубы 48 в трубу 8 и гена 32 из трубы 42 в трубу 25; и

два горизонтальных переноса без сохранения гена 8 из трубы 40 в трубу 20 и гена 15 из трубы 53 в трубу 38.

Чтобы уточнить полученный сценарий и выявить наиболее надежные переносы, мы **увеличили вес переноса** на 1, т.е. вес переноса с сохранением положили равным 12, а отсюда вес переноса без сохранения – 14. И получили следующий модифицированный сценарий:

13 дубликаций: 0(0,1); 1(3,4,5,13); 8(17); 28(52); 57(64,65,67,78); 59(71);

60 потеря: 1(64'); 2(14',45'); 3(52'); 4(32'); 5(8'); 8(44'); 11(22'); 12(18'); 13(18'"); 16(18'"); 20(40'); 25(8'"); 26(32'"); 27(39'); 28(8'","32'","38'); 29(53'); 33(55'); 34(53'"); 35(38'"); 36(33'); 37(46',60'); 38(7'); 39(15'); 40(7'',15'"); 41(7'"); 42(16',47',61'); 47(33'"); 48(61'"); 51(47'"); 52(47'"); 53(47'"); 56(51'); 57(48'); 59(66',68',82'); 60(79'); 61(77'); 64(72'); 65(72'"); 66(71',79'"); 67(68'"); 68(66'"); 71(82'"); 72(66'","82'"); 73(66'","68'","79'"); 74(81'); 75(70'); 76(64'");

и **один горизонтальный перенос без сохранения** гена 17 из трубы 53 в трубу 8 (или для другого внутреннего дерева этот перенос произошел из смежной трубы 48). Этот перенос соответствует первому переносу из предыдущего сценария (трубы 48 и 53 смежные) и **сохраняется при дальнейшем увеличении** веса переноса с сохранением вплоть до значения 18 (тогда без сохранения до 20) включительно. Именно этот предковый перенос соответствует переносу между предками {*Bha*, *Bsu*, *Sau*} и {*Vch*, *Eco*, *Vuc*, *Hin*, *Rmu*}, найденному в [15].

На рис. 6б показан фрагмент второго из этих внутренних деревьев (без группы архей и видов *Tma* и *Aae*). Обозначения такие же, как и на рис. 5б.

Здесь имеются два дерева *G'* с одним и тем же минимальным значением цены вложения: в одном случае (как показано на рис. 6б) перенос без сохранения происходит из трубы 53; в другом случае в трубе 53 сразу после видообразования происходит потеря, а перенос с сохранением происходит из смежной трубы 48.

ЗАКЛЮЧЕНИЕ

До сих пор связь дерева генов *G* и дерева видов *S* осуществлялась вложением α первого из них во второе. Это вложение привязывает вершины в *G* (события с генами) к вершинам в *S* (событиям с видами). В результате события дубликации и пропуска (а кос-

венно, и потери) гена привязываются к событиям видообразования. Однако ясно, что в процессе развития видов с генами происходят и многие другие важные события. В этой работе мы предлагаем модель, которая привязывает дубликации и потери к моментам видообразования более явным образом, а главное, привязывает горизонтальные переносы к видообразованию. Такая привязка при данных G и S осуществляется внутренним деревом G' . Его роль состоит в том, чтобы определенным образом расположить G “внутри” S : представим себе ребра дерева S как полые трубы, тогда G' в прямом смысле располагает G внутри S . При этом появляются возможности детализировать события, например, различать события горизонтального переноса гена с сохранением копии в источнике и без ее сохранения.

Тестирование и обсуждение, которое приведено в разделе “Результаты тестирования ...”, показывает, что предсказания горизонтальных переносов, полученные на основе этой модели, не расходятся с предсказаниями в других работах, которые характеризуются в них как надежные. В случае нескольких менее надежных предсказаний, сделанных в других работах, наша модель заменяет горизонтальные переносы на несколько древних дубликций. Таким образом, по крайней мере, возникает возможность в рамках точно сформулированной модели “разменивать” горизонтальные переносы на определенное число дубликций; возможность говорить, что событие происходит внутри трубы или на ее конце и т.д. Мы предполагаем расширить модель, аналогичным образом включив в нее и другие события молекулярного уровня, происходящие с соответствующими нуклеотидными последовательностями.

Авторы глубоко благодарны В.В. Алешину за ценное обсуждение этой работы.

СПИСОК ЛИТЕРАТУРЫ

1. *Mathematics of Evolution and Phylogeny*. 2005. Ed. Gascuel O. Oxford US: Oxford University Press.
2. Интернет-сайт <http://evolution.genetics.washington.edu/phylip/software.serv.html>.
3. Lyubetsky V.A., Gorbunov K.Yu., Rusin L.Y., V'ugin V.V. 2006. Algorithms to reconstruct evolutionary events at molecular level and infer species phylogeny. In: *Bioinformatics of Genome Regulation and Structure II*, Springer Sci. & Business Media, Inc., 189–204.
4. Nei M., Kumar S. 2000. *Molecular Evolution and Phylogenetics*. Oxford US: Oxford University Press.
5. Gascuel O., Steel M. 2007. *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford US: Oxford Univ. Press.
6. Page R.D.M., Holmes E.C. 1998. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Scientific, Oxford.
7. Wolf Y., Rogozin I., Grishin N., Tatusov R., Koonin E. 2001. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 1–22.
8. Durand D., Haldorsson B.V., Vernot B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* **13**, 320–335.
9. Hallett M.T., Lagergren J. 2000. New algorithms for the duplication-loss model. *Proc. Fourth Annu. Internat. Conf. Comput. Mol. Biol. RECOMB 2000*, ACM, 138–146.
10. Chauve C., Doyon J.-P., El-Mabrouk N. 2007. Inferring a duplication, speciation and loss history from a gene tree (extended abstract). in: *Comparative Genomics, RECOMB 2007 International Workshop*. Eds Tesler G., Durand D. Springer, 4751 of LNCS, 45–57.
11. Willson S. 2004. Constructing rooted supertrees using distances. *Bull. Math. Biol.* **66**, 1755–1783.
12. Guigo R., Muchnik I., Smith T.F. 1996. Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.* **6**, 189–213.
13. Eulenstein O., Mirkin B., Vingron M. 1998. Duplication-based measures of difference between gene and species trees. *J. Comput. Biol.* **5**, 135–148.
14. Novozhilov A.S., Karev G.P., Koonin E.V. 2005. Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.* **22**, 1721–1732.
15. Горбунов К.Ю., Любецкий В.А. 2005. Поиск предковых генов, нарушающих согласованность деревьев белков и видов. *Молекуляр. биология*. **39**, 847–858.
16. Горбунов К.Ю., Любецкий В.А. 2007. Эволюция предковых регуляторных сигналов вдоль дерева эволюции фактора транскрипции. *Молекуляр. биология*. **41**, 918–925.
17. Smith M.W., Feng D.F., Doolittle R.F. 1992. Evolution by acquisition: The case for horizontal gene transfers. *Trends Biochem. Sci.* **17**, 489–493.
18. Page R.D.M., Charleston M.A. 1997. Reconciled trees and incongruent gene and species trees. In *Mathematical hierarchies in biology*. Eds Mirkin B., McMorris F.R., Roberts F.S., Rzhetsky A. *Amer. Math. Soc.* **37**, 1–14.
19. Любецкий В.А., Жижина Е.А., Рубанов Л.И. 2008. Гиббсовский подход в задаче эволюции регуляторного сигнала экспрессии гена. *Проблемы передачи информации*, **44**, 52–71.
20. Mirkin B.G., Fenner T.I., Galperin M.Y., Koonin E.V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 1–34.
21. Johnston A.W., Todd J.D., Curson A.R., Lei S., Nikolaidou-Katsaridou N., Gelfand M.S., Rodionov D.A. 2007. Living without Fur: the subtlety and complexity of iron-responsive gene regulation in the symbiotic bacterium *Rhizobium* and other alpha-proteobacteria. *Biometals*. **20**, 501–511.
22. Gerasimova A.V., Gelfand M.S. 2005. Evolution of the NadR regulon in Enterobacteriaceae. *J. Bioinform. Comput. Biol.* **3**, 1007–1019.
23. Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. 2005. Comparative analysis of RNA regulatory elements

- of amino acid metabolism genes in Actinobacteria. *BMC Microbiol.* **5**, 1–14.
24. Vitreschak A.G., Mironov A.A., Lyubetsky V.A., Gelfand M.S. 2008. Functional and evolutionary analysis of the T-box regulon in bacteria. *RNA*. **14**, 717–735.
25. Горбунов К.Ю., Любецкая Е.В., Асарин Е.А., Любецкий В.А. 2009. Реконструкция эволюции бактериальных регуляторных сигналов, основанных на вторичной структуре. *Молекуляр. биология*, **43**, 527–541.
26. Zghidi W., Merendino L., Cottet A., Mache R., Lerbs-Mache S. 2007. Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the psbN gene in plastids. *Nucleic Acids Res.* **35**, 455–464.
27. Favory J.-J., Kobayshi M., Tanaka K., Peltier G., Kreis M., Valay J.-G., Lerbs-Mache S. 2005. Specific function of a plastid sigma factor for *ndhF* gene transcription. *Nucleic Acids Res.* **33**, 5991–5999.
28. Seliverstov A.V., Lyubetsky V.A. 2006. Translation regulation of intron containing genes in chloroplasts. *J. Bioinform. Comput. Biol.* **4**, 783–793.
29. Lyubetsky V.A., V'yugin V.V. 2003. Methods of horizontal gene transfer determination using phylogenetic data. *In silico Biol.* **3**, 17–31.
30. Вьюгин В.В., Гельфанд М.С., Любецкий В.А. 2003. Идентификация горизонтально перенесенных генов на основе филогенетических данных. *Молекуляр. биология*. **37**, 673–687.
31. Вьюгин В.В., Гельфанд М.С., Любецкий В.А. 2002. Согласование деревьев: реконструкция эволюции видов по филогенетическим деревьям генов. *Молекуляр. биология*. **36**, 807–816.