МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2003, том 37, № 4, с. 674–687

БИОИНФОРМАТИКА

УДК 577.2.08:681.31:575.852

ИДЕНТИФИКАЦИЯ ГОРИЗОНТАЛЬНО ПЕРЕНЕСЕННЫХ ГЕНОВ НА ОСНОВЕ ФИЛОГЕНЕТИЧЕСКИХ ДАННЫХ

© 2003 г. В. В. Вьюгин*, М. С. Гельфанд[1], В. А. Любецкий

Институт проблем передачи информации Российской академии наук, Москва, 101447

[1] Федеральное государственное унитарное предприятие "ГосНИИгенетика", Москва, 113545

Поступила в редакцию 29.10.2002 г.

Предлагается способ для поиска генов, в истории которых имел место горизонтальный перенос. Такой поиск основан на учете различия в топологиях между деревьями эволюции групп генов (белков) и соответствующих им видов. Эта рассогласованность измеряется с помощью введенных в работе численных характеристик. Предлагаемая методика применялась к генам из 40 геномов прокариот, объединенным в 132 кластера ортологов. В результате выделен список генов, относительно которых гипотеза о событиях горизонтального переноса в ходе их эволюционной истории представляется правдоподобной.

Ключевые слова: горизонтальный перенос генов, эволюционное событие, статистика поиска эволюционных событий, филогенетическое дерево видов, филогенетическое дерево белков, согласование деревьев, квадратичная сложность.

Известно, что филогенетические деревья различных семейств белков из одних и тех же организмов часто не совпадают друг с другом, а также с известными из эволюционной биологии деревьями эволюции видов (организмов), содержащих эти белки. Причиной этого могут быть как неточности в построении деревьев эволюции белков (генов), вызванные, например, различиями в скорости эволюции одного гена в разных филетических линиях, так и тот принципиальный факт, что деревья генов могут отличаться от дерева видов из-за событий, происходивших на молекулярном уровне в истории геномов и не связанных с дивергенцией видов. К таким событиям относятся, в частности, дупликации и потери генов, а также горизонтальный перенос генов. Подробнее деревья эволюции белков, генов, видов и события дупликации и потери генов обсуждаются, например, в наших работах [1, 2].

Нашей целью является разработка методов получения информации о таких эволюционных событиях на основе филогенетических данных. В этой работе мы предлагаем новый подход для отбора генов, подозреваемых участников горизонтального переноса, в ходе их эволюционной истории, на основе вызываемой ими рассогласованности между деревьями групп генов и видов.

Горизонтальный перенос между геномами бактерий происходит систематически. Бактерия может получить ген в свою хромосому непосредственно из окружающей среды, в результате фаговой инфекции, от другой бактерии посредством плазмид [3–5]. Для медицины важное значение горизонтального переноса определяется тем, что многие плазмиды переносят гены устойчивости к антибиотикам или содержат островки вирулентности, включающие гены токсинов, белков инвазии и другие факторы патогенности. Некоторые авторы считают, что горизонтальный перенос генов является одним из основных факторов эволюции микроорганизмов [6–9]. Геном *Escherichia coli* содержит до 18% горизонтально перенесенных генов [10]. В геноме *Thermotoga maritima* 25% генов более родственны генам архебактерий, чем бактерий, и предполагают, что они попали в этот геном в результате горизонтального переноса [3, 11]. Еще раньше аналогичные результаты были получены при исследовании генома другой бактерии, *Aquifex aeolicus* [12].

Постановка задачи компьютерного поиска горизонтально перенесенных генов на основе филогенетических данных рассматривалась ранее [13, 14]. Однако в этих работах не были предложены какие-либо методы ее решения. В нашей работе предлагается метод отбора генов – кандидатов на горизонтальный перенос. Соответствующая компьютерная программа составляет списки генов, послуживших причиной значительной рассогласованности деревьев эволюции групп соответствующих генов и видов. Конечно, дальнейший отбор должен производиться экспертным путем на основе анализа функций отобранных генов, а также их сходства с другими генами организма, в котором они сейчас находятся, и организ-

* Эл. почта: vyugin@iitp.ru

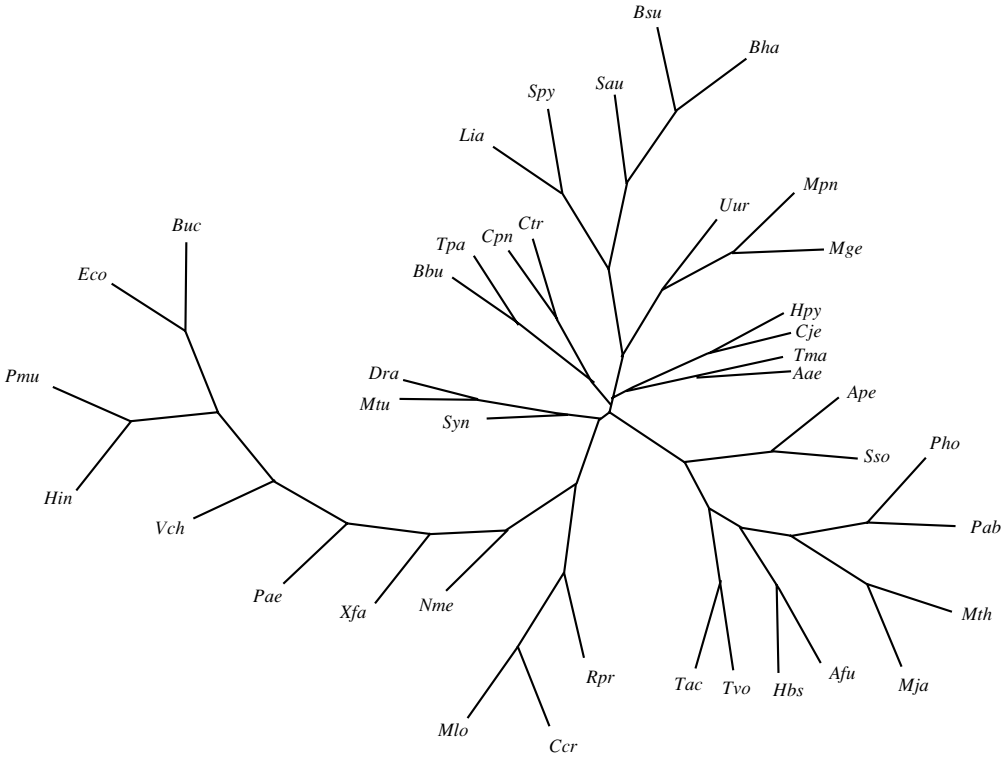


Рис. 1. Дерево *S** эволюции 40 микроорганизмов из выборки Ф.

ма, откуда в свое время, как можно предположить, осуществился горизонтальный перенос.

Когда эта статья уже прошла окончательное рецензирование, появилась публикация [19], к которой мы даем следующий краткий комментарий. В работе [19] предлагается алгоритм для наиболее экономичного вложения филогенетического паттерна (множества геномов, содержащих данный ген) в дерево видов. В качестве элементарных операций в ней рассматриваются также потери гена, дупликации и горизонтальные переносы гена. Показано, что частота горизонтальных переносов в ходе бактериальной эволюции сравнима с частотой потерь генов. При этом не рассматриваются перестройки дерева генов вследствие горизонтальных переносов, так как основная задача этой работы – это восстановление набора генов последнего общего предка рассматриваемых геномов. Тем самым, подход работы [19], в определенном смысле, дополнителен к нашему.

ИСХОДНЫЕ ДАННЫЕ

Предлагаемый метод применяют к списку (который будем обозначать Ф), состоящему из 40 микроорганизмов из 13 групп организмов.

Археи: (*Afu*) *Archaeoglobus fulgidus*; (*Hbs*) *Halobacterium* sp. NRC-1; (*Mja*) *Methanococcus jannaschii*; (*Mth*) *Methanobacterium thermoautotrophicum*; (*Tac*) *Thermoplasma acidophilum*; (*Tvo*) *Thermoplasma volcanium*; (*Pho*) *Pyrococcus horikoshii*; (*Pab*) *Pyrococcus abyssi*; (*Ape*) *Aeropyrum pernix*; (*Sso*) *Sulfolobus solfataricus*.

Гамма-протеобактерии: (*Eco*) *Escherichia coli*; (*Buc*) *Buchnera* sp.; (*Pae*) *Pseudomonas aeruginosa*; (*Vch*) *Vibrio cholerae*; (*Hin*) *Haemophilus influenzae*; (*Pmu*) *Pasteurella multocida*; (*Xfa*) *Xylella fastidiosa*.

Бета-протеобактерии: (*Nme*) *Neisseria meningitidis* MC58.

Альфа-протеобактерии: (*Mlo*) *Mesorhizobium loti*; (*Ccr*) *Caulobacter crescentus*; *(Rpr*) *Rickettsia prowazekii*.

Эпсилон-протеобактерии: (*Hpy*) *Helicobacter pylori*; (*Cje*) *Campylobacter jejuni*.

Грамположительные бактерии (Firmicutes и Mollicutes): (*Spy*) *Streptococcus pyogenes*; (*Bsu*) *Bacillus subtilis*; (*Bha*) *Bacillus halodurans*; (*Lla*) *Lactococcus lactis*; (*Sau*) *Staphylococcus aureus*; (*Uur*) *Ureaplasma urealyticum*; (*Mpn*) *Mycoplasma pneumoniae*; (*Mge*) *Mycoplasma genitalium*.

Хламидии: (*Ctr*) *Chlamydia trachomatis*; (*Cpn*) *Chlamydia pneumoniae*.

Спирохеты: (*Tpa*) *Treponema pallidum*; (*Bbu*) *Borrelia burgdorferi*.

Группа DMS: (*Dra*) *Deinococcus radiodurans*; (*Mtu*) *Mycobacterium tuberculosis*; (*Syn*) *Synechocystis*.

Термотога и аквифекс: (*Aae*) *Aquifex aeolicus*; (*Tma*) *Thermotoga maritima*.

Расчеты проводили с использованием базы COG, содержащей кластеры ортологичных генов (http://www.ncbi.nlm.nih.gov / COG/). Каждый кластер (сокращенно КОГ) содержит группу генов и соответствующих им белков, имеющих общее происхождение и ответственных за единую функцию. Каждому КОГу также соответствует множественное выравнивание его белковых последовательностей, по которому разными методами можно построить филогенетическое дерево генов (белков) этого кластера. Эти выравнивания и деревья любезно предоставлены Ю. Вульфом и Е. Куниным (Национальный центр биотехнологической информации США). Деревья построены с помощью комбинации дистанционного метода и метода максимального правдоподобия, в результате чего строилось филогенетическое дерево генов, составляющих этот КОГ, таким образом, что расстояния по дереву отражали степень сходства белков [15].

Каждое из 132 полученных таким образом деревьев имеет на каждом своем ребре число, отражающее предположительное эволюционное время между событиями, приписанными концам этого ребра.

Дерево видов S строилось нами с помощью нашего алгоритма согласования деревьев генов [2] как дерево, наиболее близкое к этим 132 деревьям генов G_i (где i меняется от 1 до 132; в соответствии с описанием задачи 1 ниже). Так полученное дерево видов S^* (рис. 1) практически совпадает с деревом видов, которое строится на основе пятого метода из работы Вульфа и др. [16]. Оно также любезно передано нам авторами последней работы; именно дерево видов S^* использовалось нами для получения результатов по идентификации горизонтальных переносов, которые далее приводятся.

Графические изображения дерева видов S^*, а также филогенетических деревьев генов для тех КОГов, которые специально обсуждаются в разделе "Результаты", приведены ниже. В таблице содержатся все гены, отобранные нашим методом как подозреваемые участники горизонтального переноса.

МЕТОДЫ

Филогенетические деревья и их вложения

Рассматриваются деревья G, G_1, …, G_n, S, каждое с каким-то своим числом m концевых вершин. Деревья, обозначаемые G (с индексом или без), понимаются как филогенетические деревья белков (генов), а обозначаемые S (с индексом или без), – как филогенетические деревья видов (кластеров, организмов). Существуют разные естественные способы измерения степени (цены) отличия каких-то двух деревьев G и S; *цена отличия* двух деревьев (синоним: цена вложения G в S) обозначается $c(G, S)$. Отличие какого-то набора деревьев G_1, …, G_n от какого-то дерева S тогда естественно измеряется как

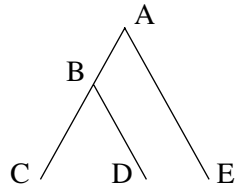
$$F(G_1, \ldots, G_n, S) = \sum_{i=1}^{n} c(G_i, S).$$

Функционал, подобный такому F, будем называть *ценой отличия* (синоним: *качество согласования*) данного набора деревьев G_1, …, G_n. Дерево S, минимизирующее этот функционал, называется результатом согласования соответствующего набора деревьев (синоним: их *консенсусом*).

Пример определения цены отличия $c(G, S)$ двух деревьев G и S таков. Определяется вложение α: $G \longrightarrow S$ как тождественное на концевых вершинах и равное

$$\alpha(x \cup y) = \alpha(x) \cup \alpha(y),$$

где знак $\cup$ обозначает супремум (наименьшую точную верхнюю грань множества $\{\alpha(x), \alpha(y)\}$, т.е. множества из двух слагаемых в правой части равенства). Нужно различать понятия "отец" и "супремум", что поясняется на следующем примере (примере 1):



Здесь A – отец B, E, но супремум A, B, C, D, E; B – отец C и D, но супремум B, C, D.

Точнее, термин "тождественное" означает здесь следующее. Если в каждом организме берется по одному гену, то α действительно тождественное отображение, т.е. отображающее каждый ген в его организм. В общем случае все гены, взятые из одного организма, отображаются в него; если из какого-то организма не были взяты гены, то ни одна из концевых вершин G не отображается в него.

Возможны и другие определения такого типа вложения, учитывающие более тонкие представления об эволюции. Заметим, что для этой тематики существенна разработка даже очень гипотетических представлений об эволюции, а отсюда – и ограничений на вид модели, положенной в основу вычислений.

Затем $c(G, S)$ определяется как сумма штрафов за каждое нарушение инъективности (склейку двух аргументов в одно значение, т.е. за каждую пару x, y, для которой $\alpha(x) = \alpha(y)$) и каждое нарушение сюръективности (значение z, в кото-

рое никто не отображается, т.е. за каждое z, для которого не выполняется $z = \alpha(x)$ ни при каком x). Иными словами, – за каждую дупликацию (двухстороннюю или только за одностороннюю) и за каждую промежуточную вершину (пропуск). Здесь может быть много вариантов с разными коэффициентами и параметрами, отражающими то или иное представление об эволюции.

Вычисление цены отличия двух филогенетических деревьев

В расчетах, проведенных для этой работы, величину $c(G, S)$ вычисляли следующим образом. Пара (g, s), где $\alpha(g) = s$, называется *односторонней дупликацией*, если выполняется одно и только одно из условий $\alpha(g) = \alpha(cg)$ или $\alpha(g) = \alpha(c'g)$, где cg – *левый сын* вершины g, а $c'g$ – *правый сын* вершины g. (Если выполнены оба этих условия, то пара (g, s) называется *двусторонней дупликацией*.) Множество всех односторонних дупликаций обозначим $O(G, S)$. Вершина s из S называется *g-промежуточной*, если она расположена строго между $\alpha(g)$ и $\alpha(pg)$, где pg – *отец* вершины g. Множество всех g-промежуточных вершин обозначим I_g, а объединение множеств I_g по всем g обозначим $M(S, G)$.

Алгоритмы построения филогенетического дерева G белков (генов) иногда позволяют приписывать длины ребрам этого дерева, эти длины будем обозначать $c(g, g')$, где g и g' – соседние вершины в дереве G. В некоторых случаях такие длины могут интерпретироваться как время, прошедшее между двумя событиями g и g' в дереве генов G, в предположении, что скорость эволюции рассматриваемых генов постоянна. Такие длины вычислялись нами либо из *показателя сходства* последовательностей, приписанных алгоритмом (построения дерева G) вершинам g и g', либо из *бутстрепа поддержки* соответствующего кластера (части дерева G под ребром gg'). Приводимые ниже расчеты использовали именно показатель сходства. Итак, в этих расчетах принималось

$$c(G, S) = \sum_{(g, \alpha(g)) \in O(G, S)} c(g, pg) + \gamma \sum_{(g, \alpha(g)) \in M(G, S)} c(g, pg)|I_g|.$$

Если какое-то слагаемое здесь не определено, то оно полагалось равным нулю.

Первый член этой суммы характеризует потери от односторонних дупликаций, а второй – от пропущенных вершин. Мультипликативный параметр γ регулирует соотношение значимости дупликаций и пропусков в ходе эволюции. Во всех приведенных здесь расчетах принимали $\gamma = 0.1$.

Нами проводились вычисления и в случае, когда к двум предыдущим суммам добавлялось еще третье слагаемое, характеризующее качество дерева генов G (например, качество КОГа, по которому строится это дерево, – мы называем его *рассеянностью* КОГа и обозначаем R_G). Сейчас приводятся результаты вычислений без этого третьего слагаемого.

При каждом вычислении функционала F и величины c (или аналогичных им) для любого дерева генов G перед вычислением F применяли *нормализацию длин* ребер дерева G с целью пропорционального сокращения слишком длинных (по сравнению со средней длиной) ребер концевых вершин. Для этого вычисляли среднюю длину $l_{ср}(G)$ всех ребер концевых вершин в дереве G. И затем длины всех ребер концевых вершин, у которых $l(g) > l_{ср}$, изменяли по формуле

$$l(g) = (l(g) - l_{ср})(1 + \mu)^{-(l(g)/l_{ср})} + l_{ср},$$

где в левой части находится новая длина $l(g)$ такого ребра, а в правой части – старая длина $l(g)$ того же ребра. Во всех приведенных здесь расчетах принималось $\mu = 0.7$. Смысл такой нормализации в том, что она, уменьшая вклад наиболее длинных ребер в величину функционала потерь F, уменьшает влияние таких нежелательных здесь причин рассогласования деревьев, как различия в скорости эволюции генов одного семейства в разных филетических линиях.

Постановка двух задач

Заметим, что, таким образом, функционал F имеет вид $F_{\gamma, \mu}$, и нами рассматривалась также задача подбора таких значений параметров в функционал $F_{\gamma, \mu}$, при которых в первой задаче (см. ниже) получается дерево видов S, наиболее близкое к биологически признанному (в данном случае) дереву видов S^* (мы пробовали вводить и другие параметры, отражающие представления об эволюции, или варьировать F в пространстве функций). Такая постановка представляет собой самостоятельную задачу, которая не будет здесь обсуждаться.

Итак, возникают две задачи.

Первая задача. Дан набор деревьев белков $G_1, \ldots, G_n$ и ищется дерево видов S, наименее отличающееся от них (т.е. такое, при котором цена отличия F на этом S принимает наименьшее возможное значение, локальное или глобальное).

Вторая задача. Даны дерево белков G и дерево видов S, цена различия $c(G, S)$ между которыми велика. Найти концевую вершину в G (т.е. белок или ген) или группу таких вершин, за счет кото-

рой (которых) возникает это различие, т.е. таких, что после удаления этой вершины (или этих вершин) цена различия $c(G', S')$ оставшихся (после удаления) деревьев G' и S' будет мала.

Таким образом, мы хотим, в частности, предсказывать гены подозреваемых участников горизонтального переноса в ходе их эволюционной истории. Существенно также уметь оценивать качество КОГов (см. [2]).

В этой публикации далее рассматривается только вторая задача.

Ген g (концевая вершина в дереве генов G), кандидат на горизонтальный перенос, определяется нами следующим образом. В дереве видов обычно фиксируется таксономия, т.е. структура групп родственных видов. Эта структура может быть одноуровневой или многоуровневой (разного типа), т.е. в свою очередь определяется некоторым графом, вершинам которого приписаны группы видов.

Окрестность радиуса r с центром в концевой вершине g по определению состоит из всех концевых вершин g' в дереве G, расстояние до которых от g меньше или равно r; *расстояние* между двумя вершинами в G определяется как число ребер в кратчайшем пути между ними в G. Окрестность радиуса r с центром в g (в дереве генов G, состоящую из его только концевых вершин) без самого центра g назовем *выколотой*. Если при отображении α образ выколотой окрестности содержится в одной или нескольких соседних группах видов, а сам центр g переходит в далекий от этих групп вид (расстояние берется по дереву S видов), то g считается *подозреваемым в участии в горизонтальном переносе* (по данному критерию). Для поиска таких генов вычисляются и сравниваются несколько величин (*статистик* рассогласования), отражающих идею этого определения.

Перечисленные выше в разделе "данные" группы организмов рассматривались нами как *одноуровневая таксономия* в S^*.

Для любой такой группы или множества групп в дереве видов (организмов) S^* имеется ровно одна вершина, поддерево (кластер) которой включает объединение этих групп, и это поддерево – наименьшее (по включению) среди всех таких поддеревьев. Это поддерево назовем *абстрактной группой* (видов, организмов) в S^*, а соответствующую ему вершину назовем *корнем* этой группы. Абстрактная группа, образованная по одной исходной группе этой таксономии, конечно, совпадает с ней. *Размер* абстрактной группы определим как наибольшее (или иногда – как статистически значимое) расстояние от ее корня до какой-то входящей в нее концевой вершины. По любой абстрактной группе и любой вершине в дереве S^* определим *расстояние* между ними как число ребер кратчайшего пути между корнем группы и этой вершиной; расстояние берется со знаком плюс, если сама вершина не входит в группу, и со знаком минус – в ином случае.

Итак, *определение* горизонтального переноса, приведенное выше, говорит, что образ выколотой окрестности гена g должен образовать абстрактную группу маленького размера, у которой расстояние до образа самого гена g большое.

Нами рассматривалась и более сложная, чем приведенная выше, многоуровневая таксономия в дереве S^*, что будет отражено в другой публикации.

Статистики для идентификации генов, подозреваемых участников горизонтального переноса

Теперь определим статистики для **второй** из перечисленных выше задач, а затем приведем (ниже в таблице) результаты их вычисления для некоторых генов из выборки Ф. Полные результаты вычислений для всех генов из этой выборки находятся по электронному адресу http://www.iitp.ru/lyubetsky. В таблице статистики расположены в соответствии со столбцами этой таблицы, поэтому определения статистик приводятся нами в том же порядке.

Самый левый столбец нумерует строки таблицы. В следующем, **первом столбце** приводятся номер КОГа, затем сокращенное название организма и имя гена (перед этим указывается номер данного гена в данном КОГе).

Итак, для данных деревьев G и S вычисляли *относительное приращение цены вложения* (сокращенно: опцв) F_g для гена g (**второй столбец**), а именно, в этом столбце приводится результат вычисления (в процентах) величины

$$F_g = ((c_g - c)/c) \times 100,$$

где c – цена вложения G в S, а c_g – цена вложения G_g в S. Здесь G_g получается из G удалением концевой вершины g. При удалении концевой вершины возникают новые ребра, длины которых определяются как суммы длин ребер на "исчезнувшем" пути. Затем в скобках указывается p-значение, вычисляемое по формуле:

$$p(g) = card(\{g' : F_{g'} \le F_g\})/m.$$

Здесь $card(X)$ обозначает число элементов в множестве X, а m – число концевых вершин (генов) в дереве G. В программе предусмотрен отбор и печать генов в G, для которых выполняется условие: значение F_g достаточно мало в смысле порога P_0 и $p(g) \le p_0$, где p_0 – любой заданный (подходящий) порог.

Результат поиска генов в организмах, представленных в выборке Ф, подозреваемых участников горизонтального переноса в ходе их эволюционного развития (подробные описания столбцов см. в тексте)

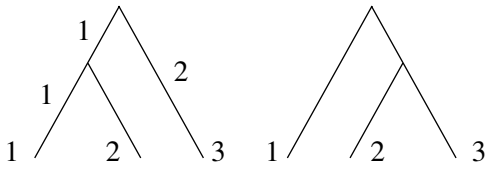
N	1	2	3	4	5	6	7
1	38. COG0012[*Syn*] *sll0245*	–3.78% (0.1)	2.67 (0.1)	–1.37 –		(3/8 3/8)	1.3
2	41. COG0012[*Buc*] *BU191*	–10.74% (0.02)	3.14 (0.03)	–3.73	хламидии	(3/12 3/12 4/10 4/10)	1.59
3	38. COG0018[*Syn*] *sll0502*	–7.07% (0.07)	2.64 (0.1)	–1.74	хламидии	(3/8 3/8 4/11 4/10)	1.07
4	40. COG0018[*Mtu*] *Rv1292*	–10.74% (0.02)	2.9 (0.05)	–2.62		(2/10 4/10 4/9)	1.38
5	35. COG0061[*Xfa*] *XF2090*	–9.29% (0.03)	1.89 (0.15)	–2.75	альфа-протеобактерии	(2/6 3/6 4/5)	0.93
6	40. COG0072[*Mtu*] *Rv1650_2*	–11.45% (0.03)	2.16 (0.03)	–3.49	альфа-протеобактерии	(3/9 4/9 4/10 4/10 4/3)	1.52
7	41. COG0080[*Bha*] *BH0409*	–6.91% (0.02)	2.17 (0.02)	–2.77	спирохеты	(4/10 4/10 4/6)	1.85
8	38. COG0085[*Aae*] *aq_1939*	–15.91% (0.03)	2.33 (0.08)	–3.77	эпсилон-протеобактерии	(3/7 3/7)	1.76
9	40. COG0102[*Dra*] *DR0174*	–12.29% (0.03)	2.8 (0.03)	–4.43	гамма-протеобактерии	(2/9 4/11 4/8)	1.73
10	37. COG0126[*Aae*] *aq_118*	–11.59% (0.03)	2.0 (0.05)	–2.87		(3/7 3/7 4/7 4/7)	1.34
11	40. COG0143[*Mtu*] *Rv1007c*	–8.95% (0.05)	3.8 (0.05)	–2.32	альфа-протеобактерии	(2/10 3/9)	1.12
12	41. COG0143[*Mlo*] *mlr5926*	–12.92% (0.02)	4.6 (0.02)	–3.45		(2/11 3/12)	1.53
13	42. COG0162[*Pae*] *PA4138*	–8.12% (0.02)	2.22 (0.05)	–3.05		(2/7 3/7 4/6)	1.47
14	30. COG0173[*Mtu*] *Rv2572c*	–17.47% (0.03)	2.37 (0.03)	–2.89	альфа-протеобактерии	(3/9 4/8 4/10 4/10 4/8)	1.5
15	33. COG0178[*Hbs*] *VNG2636G*	–8.38% (0.03)	2.6 (0.03)	–2.18	DMS	(2/8 4/9 4/9)	1.47
16	30. COG0193[*Dra*] *DR2372*	–16.18% (0.03)	2.67 (0.03)	–2.41	гамма-протеобактерии	(2/7 4/9)	1.26
17	40. COG0198[*Bbu*] *BB0489*	–15.81% (0.03)	2.63 (0.03)	–4.23	гамма-протеобактерии	(3/7 4/12 4/10 4/12 4/9)	1.75
18	38. COG0200[*Bbu*] *BB0497*	–6.64% (0.05)	2.36 (0.08)	–2.75		(3/8 4/9 4/9)	1.55
19	30. COG0203[*Mtu*] *Rv3456c*	–11.92% (0.03)	2.67 (0.03)	–2.73		(3/8 3/8)	1.53
20	38. COG0215[*Xfa*] *XF0995*	–13.86% (0.05)	1.89 (0.05)	–2.5	альфа-протеобактерии	(2/6 3/6 4/5)	1.27
21	39. COG0215[*Hbs*] *VNG1097G*	–20.64% (0.03)	2.5 (0.03)	–3.78	DMS	(2/8 4/9 4/9 4/9)	1.52
22	29. COG0221[*Mlo*] *mlr8562*	–9.12% (0.03)	3.29 (0.1)	–2.24		(3/10 4/13)	1.67

Таблица. Окончание

N	1	2	3	4	5	6	7
23	30. COG0222[*Dra*] *DR2043*	–10.31% (0.03)	3.4 (0.03)	–2.23	альфа-протеобактерии	(2/9 3/8)	1.74
24	35. COG0242[*Syn*] *slr1549*	–10.14% (0.03)	3.2 (0.03)	–2.64	спирохеты	(2/8 3/8)	1.47
25	40. COG0250[*Aae*] *aq_1931*	–9.29% (0.02)	2.2 (0.09)	–3.55		(3/7 3/7 4/8)	1.73
26	27. COG0272[*Rpr*] *RP720*	–8.8% (0.1)	1.67 (0.16)	–1.16		(3/5 3/5)	1.4
27	31. COG0272[*Eco*] *yicF*	–29.13% (0.03)	4.8 (0.03)	–4.22	спирохеты	(2/12 3/12)	1.64
28	30.COG0292[*Tma*] *TM1592*	–20.24% (0.03)	2.8 (0.03)	–3.04	DMS	(2/7 3/7)	1.76
29	35. COG0294[*Ccr*] *CC3224*	–7.61% (0.06)	3.22 (0.06)	–1.6	DMS	(3/10 3/10 3/9)	0.89
30	30. COG0335[*Dra*] *DR0755*	–6.19% (0.07)	3.6 (0.03)	–1.7 –	Chlamidiae	(2/9 3/9)	1.77
31	36. COG0343[*Afu*] *AF1485*	–8.58% (0.06)	3.33 (0.03)	–2.2	хламидии	(3/10 3/10)	1.6
32	30. COG0359[*Aae*] *aq_2042*	–13.19% (0.03)	2.8 (0.03)	–2.64	DMS	(2/7 3/7)	1.39
33	40. COG0441[*Ape*] *APE0809*	–14.09% (0.02)	3.0 (0.07)	–3.25	DMS	(2/8 3/7)	1.79
34	29. COG0452[*Bbu*] *BB0812*	–8.69% (0.03)	2.33 (0.03)	–2.34		(2/7 4/7)	1.35
35	37. COG0504[*Tpa*] *TP0305*	–21.5% (0.03)	3.2 (0.05)	–3.95	DMS	(2/8 3/8)	1.64
36	40. COG0525[*Rpr*] *RP687*	–16.16% (0.03)	3.67 (0.03)	–3.99		(3/12 3/12 3/9)	1.62
37	33. COG0547[*Cje*] *Cj0346_2*	–12.66% (0.03)	3.6 (0.09)	–2.6		(2/9 3/9)	1.28
38	39. COG0552[*Dra*] *DR2260*	–13.72% (0.03)	2.27 (0.05)	–3.14	альфа-протеобактерии	(3/8 4/9 4/9 4/8)	1.52
39	29. COG0556[*Hbs*] *VNG2390G*	–9.23% (0.03)	2.33 (0.03)	–1.93 –	DMS	(3/9 3/8 3/4)	1.53
40	28. COG0571[*Syn*] *slr0346*	–19.72% (0.07)	2.67 (0.1)	–2.43	хламидии	(3/8 3/8 3/8)	1.56
41	35. COG0587[*Bsu*] *BS_yorL*	–10.28% (0.03)	3.11 (0.03)	–2.5		(3/10 3/10 3/8)	1.6

Итак, здесь при достаточно малом значении p_0 отбирались все случаи экстремального положения гена g в дереве генов, т.е. отбирались все гены g, для которых величина F_g *экстремально мала по сравнению с большинством* других значений $F_{g'}$. Для таблицы мы, как правило, брали $P_0 = -7$ и $p_0 = 0.1$.

Пример 2А.

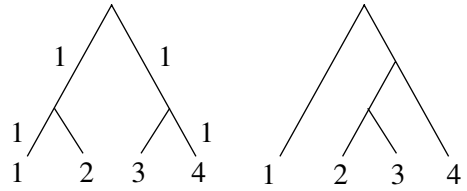


Здесь слева дерево генов G (в нем ребро к гену 2 тоже длины 1), а справа дерево видов S. Стоимость вложения G в S после нормализации G равна $c = 1.16$ (здесь везде $\gamma = 0.1$, $\mu = 0.7$).

G	1	2	3
F_g (без нормал.)	–100%	–83.3%	–83.3%
F_g (с нормал.)	–100%	–82.8%	–82.8%

Возникает упорядочение генов дерева G, при котором ген 1 наиболее подозрительный, а гены 2 и 3 одинаковы в этом отношении.

Пример 2B.



Здесь слева дерево генов G (в нем длины всех ребер равны 1) и справа дерево видов S. Стоимость вложения G в S после нормализации G равна $c = 1.3$ (везде $\gamma = 0.1$, $\mu = 0.7$).

G	1	2	3	4
F_g (без нормал.)	0%	–92.3%	7.7%	23.1%
F_g (с нормал.)	–2.85%	–92.3%	4.85%	14.5%

Возникает упорядочение генов дерева G, при котором ген 2 более подозрителен, чем ген 1 (конец примера).

В третьем столбце вычисляется обратное отношение среднего расстояния от данного g до всех его соседей g' из окрестности радиуса r на дереве генов к среднему расстоянию между $\alpha(g)$ и $\alpha(g')$ на дереве видов (в таблицах радиус окрестности r равен 4). Это отношение обозначим R_g и назовем *рассеянностью* гена g. А именно,

$$R_g = \frac{\left(\sum_{g'} \rho(\alpha(g), \alpha(g'))\right)/(m'_g - 1)}{\left(\sum_{g'} \rho(g, g')\right)/(m_g - 1)},$$

где m_g – число элементов в этой окрестности, а m'_g – число элементов в образе этой окрестности. Напомним, что расстояние ρ между g и g' вычисляется как число ребер в кратчайшем пути между g и g' в G (или между $\alpha(g)$ и $\alpha(g')$ в S).

Затем, как в столбце 2, вычисляется и указывается в скобках p-значение:

$$p(g) = card(\{g': R_{g'} \geq R_g\})/m.$$

Здесь в качестве кандидатов на горизонтальный перенос отбирались все гены g с экстремально большим значением R_g по сравнению с большинством других значений статистики R_g, для которых, кроме того, выполнялось следующее соображение.

Если филогенетически близкие соседи (в дереве генов) некоторого гена g находятся в геномах организмов (в дереве видов), далеко отстоящих от организма – носителя гена g, то мы считаем, что g мог быть привнесен в предок организма $\alpha(g)$ из предков других организмов. Если, к тому же, организмы, содержащие гены из выколотой окрестности гена g, содержатся в единой таксономической группе, то это подтверждает такую возможность. Тогда этот ген отбирается и соответствующая таксономическая группа указывается в *5-ом столбце* таблицы (как возможный источник горизонтального переноса гена g).

Итак, по этой статистике отбирались гены g, для которых R_g достаточно велико, в смысле порога P_0, и $p(g)$ достаточно мало, в смысле порога p_0 (в этих расчетах, как правило, выбиралось $P_0 = 2$ и $p_0 = 0.1$), и выполнялось условие из предыдущего абзаца. Интересно, что эмпирическое распределение статистики R_g близко к одному и тому же стандартному распределению для всех рассмотренных КОГов.

Пример 3. Для данных из примера 2B и $r = 3$.

G	1	2	3	4
R_g	2	2	1.5	1.5

Упорядочение генов: гены 1 и 2 одинаково и тот, и другой "подозрительные" и более "подозрительные", чем гены 3 и 4 (конец примера).

В **четвертом столбце** находится *отклонение* $var(g)$ (в единицах среднеквадратичного отклонения), т.е. "относительное приращение цены вложения F_g для данного g" минус "среднее ($F_{g'}^-$) значение этой величины $F_{g'}$ по всем концевым вершинам g' данного дерева генов G", деленное на "сигма". Здесь σ рассчитывается по обычной формуле, в которой m – число листьев в данном КОГе G. Итак,

$$var(g) = \frac{F_g - (F_{g'}^-)}{\sigma},$$

$$\sigma = \sqrt{(1/m - 1)\sum_g (F_g - (F_{g'}^-))^2}.$$

Оказалось, что эмпирическое распределение статистики F_g (для выборки Ф) близко к нормальному. Поэтому здесь применяется статистически обоснованная процедура: отбираются гены g, для которых уклонение $var(g)$ больше 2 по абсолютной величине. Вероятность такого события равна 0.05, поэтому соответствующие гены могут рассматриваться как экстремально расположенные.

В четвертом столбце знаком "–" отмечали нарушение этого условия $|var(g)| \geq 2$.

Статистики из 2-го и 4-го столбцов дают приблизительно одинаковые результаты отбора генов в случае нормального распределения статистики F_g. Статистика из 2-го столбца – более универсальная, так как не зависит от типа распределения F_g, а статистика из 4-го столбца имеет лучшее статистическое обоснование.

Пример 4. Для примера 2В. Здесь получается $(\bar{F}_{g'}) = -18.95$, $\sigma = 49.4$.

G	1	2	3	4
$Var(g)$	+0.33	–1.485	0.48	0.68

Упорядочение генов: только ген 2 подозрителен (конец примера).

В **пятом столбце** находится группа организмов, возможных источников горизонтального переноса гена g (здесь берется $r = 4$, можно пробовать и другие значения r, в том числе столь большие, что соответствующая окрестность охватывает множество всех концевых вершин; конечно, используется образ выколотой окрестности радиуса r с центром в точке g). В качестве такого источника, естественно, берется та таксономическая группа или абстрактная группа организмов, которая определяется образом этой выколотой окрестности. Затем указывается размер этой группы и расстояние между ею и образом g.

В **шестом столбце** находится список относительных приращений цен вложений $F_{g'}$ у соседей g' гена g, представляемых в форме: в числителе – расстояние от g до какого-то соседа g' в дереве генов G, в знаменателе – расстояние от $\alpha(g)$ до $\alpha(g')$ в дереве видов S, т.е. в форме

$$\rho(g, g')/\rho(\alpha(g), \alpha(g')).$$

В примере 5 приведено также опцв $F_{g'}$ в ситуации, когда ген g' удаляется, а ген g остается. В таблице мы не приводим эту величину $F_{g'}$, так как вся эта информация является здесь дополнительной к вычислениям в других столбцах.

Пример 5. Данные из примера 2В и $r = 3$.

G	1	2	3	4
$Var(g)$	$2/4 : F_2 = -92.3$	$2/4 : F_1 = -2.85$	$2/3 : F_4 = 14.5$	$2/3 : F_3 = 4.85$

(Конец примера.)

В **седьмом столбце** этой таблицы вычисляется отношение длины ребра концевой вершины (гена) g (после ее нормализации, см. пример 2) к средней длине ребра концевой вершины (гена) g' по всем генам g' данного дерева генов G, т.е. вычисляется $l(g)/l_{ср}(G)$. Это отношение в какой-то мере характеризует отсутствие длинных ветвей для данных дерева генов и всех его генов – его концевых вершин.

Итак, алгоритм указывает в качестве подозреваемых участников горизонтального переноса на те гены, которые отбираются по указанным выше правилам, использующим первые четыре столбца таблицы и соответствующие статистики; столбцы 5-й и 6-й содержат вспомогательную информацию, которая может быть полезна эксперту.

Конечно, отбор указанных выше порогов и, в целом, использование приведенной в таблице информации нельзя признать совершенно формальной, компьютерной процедурой. Но на данном этапе исследований этого вопроса вряд ли можно представить себе появление строго формальных процедур отбора горизонтально перенесенных генов. Скорее мы решали задачу отбора характеристик, в нашем случае статистик, которые могли бы участвовать в будущей более формальной процедуре.

Наконец, в **таблице** КОГи располагаются в порядке возрастания номера КОГа, а при одинаковом номере КОГа отобранные из него гены выписываются в том порядке, как они идут в соответствующем дереве генов G слева направо (конец описания столбцов таблицы).

Заметим, что для биологического анализа бывает полезно переупорядочить таблицу по геномам (т.е. по организмам), в которые произошел предполагаемый перенос, а иногда – по таксономическим группам организмов, из которых произошел горизонтальный перенос гена. Аналогичная более полная таблица уже для всех генов из выборки Ф приведена на сайте http://www.iitp.ru/lyubetsky.

РЕЗУЛЬТАТЫ

Теперь рассмотрим подробнее несколько отдельных поучительных случаев (рис. 2).

Два гена-кандидата на горизонтальный перенос уверенно выделяются нашим методом в COG0012 (гипотетические GTPазы) (рис. 2*а*). *Buchnera aphidicola*, эндосимбионт тли, является гамма-протеобактерией, ближайшим родственником кишечной палочки, однако ее ген *BU191* кластеризуется на дереве с генами хламидий – они и являются предполагаемым источником горизонтального переноса. Аналогично, ген *sll0245*, по-видимому, был перенесен в геном цианобактерии *Synechocystis* sp. из спирохет.

В COG0215 (цистеинил-тРНК–синтетазы) ген *VNG1095G* из галофильной архебактерии *Halobacterium* sp. (рис. 2*б*) – эубактериального происхождения. Источником, возможно, послужил геном, родственный *Deinococcus radiodurans*. В этом же кластере не исключен горизонтальный перенос гена *XF0995* из альфа-протеобактерии, родственной *Caulobacter crescentus*, в геном гамма-протеобактерии *Xylella fastidiosa*. Перенос в противо-

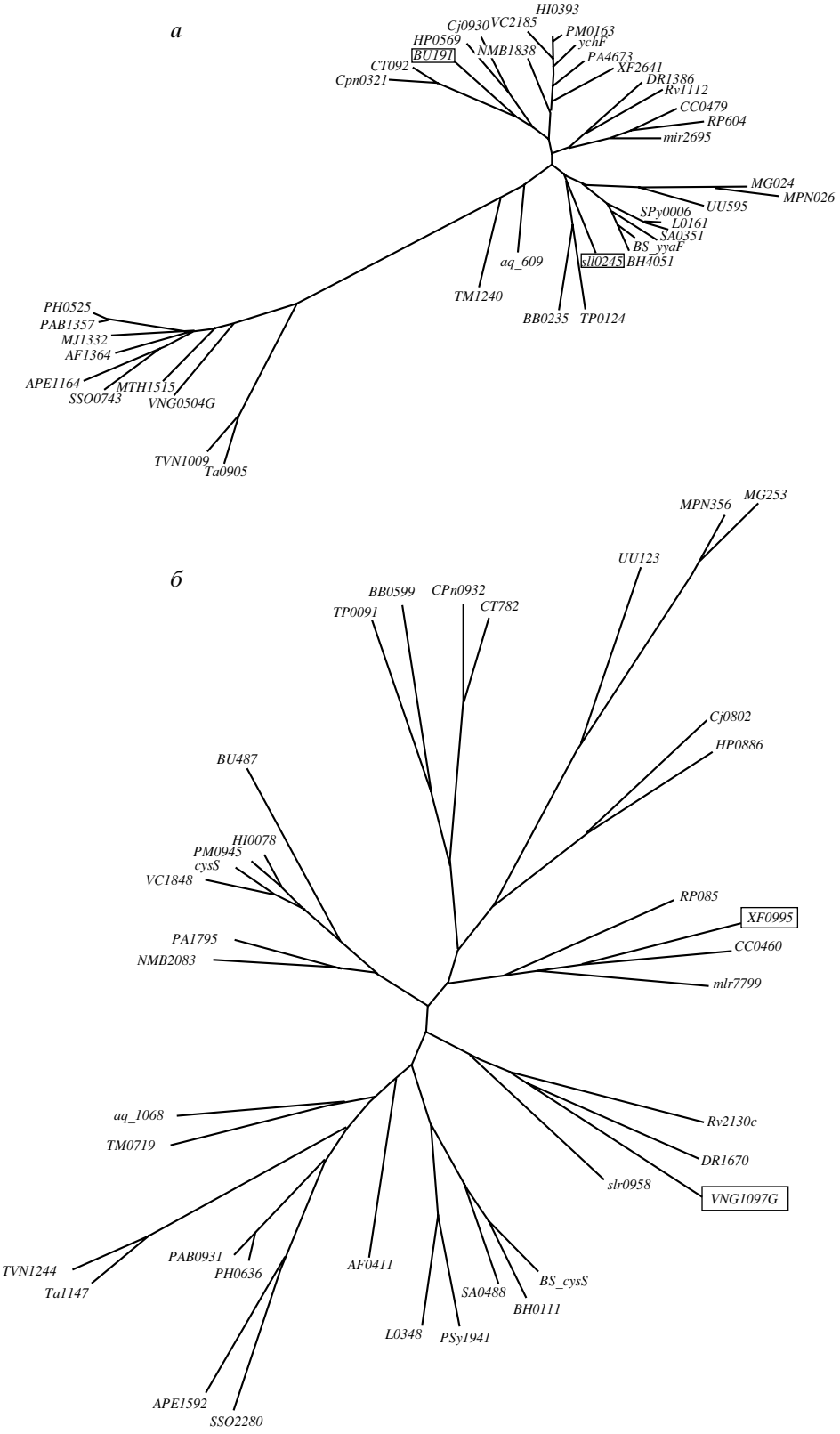


Рис. 2. Примеры горизонтальных переносов. *а* – COG0012 (гипотетические GTPазы); *б* – COG0215 (цистеинил-тРНК–синтетазы), ген *VNG1095G* из галофильной архебактерии *Halobacterium* sp.; *в* – COG0143 (метионил-тРНК-синтетазы), ген *mlr5926*.; *г* – COG0102, случай горизонтального переноса генов рибосомных белков, ген *DR0174* из генома *Deinococcus radiodurans*. *д* – COG0198, ген *BB0489* из спирохеты *Borrelia burgdorferi*; *е* – COG0272, ген кишечной палочки *yicF*, кодирующий NAD-зависимую ДНК-лигазу; *ж* – COG0343, ген кьюоин/археозин-тРНК–рибозилтрансферазы *AF1485* из генома *Archaeoglobus fulgidus*.

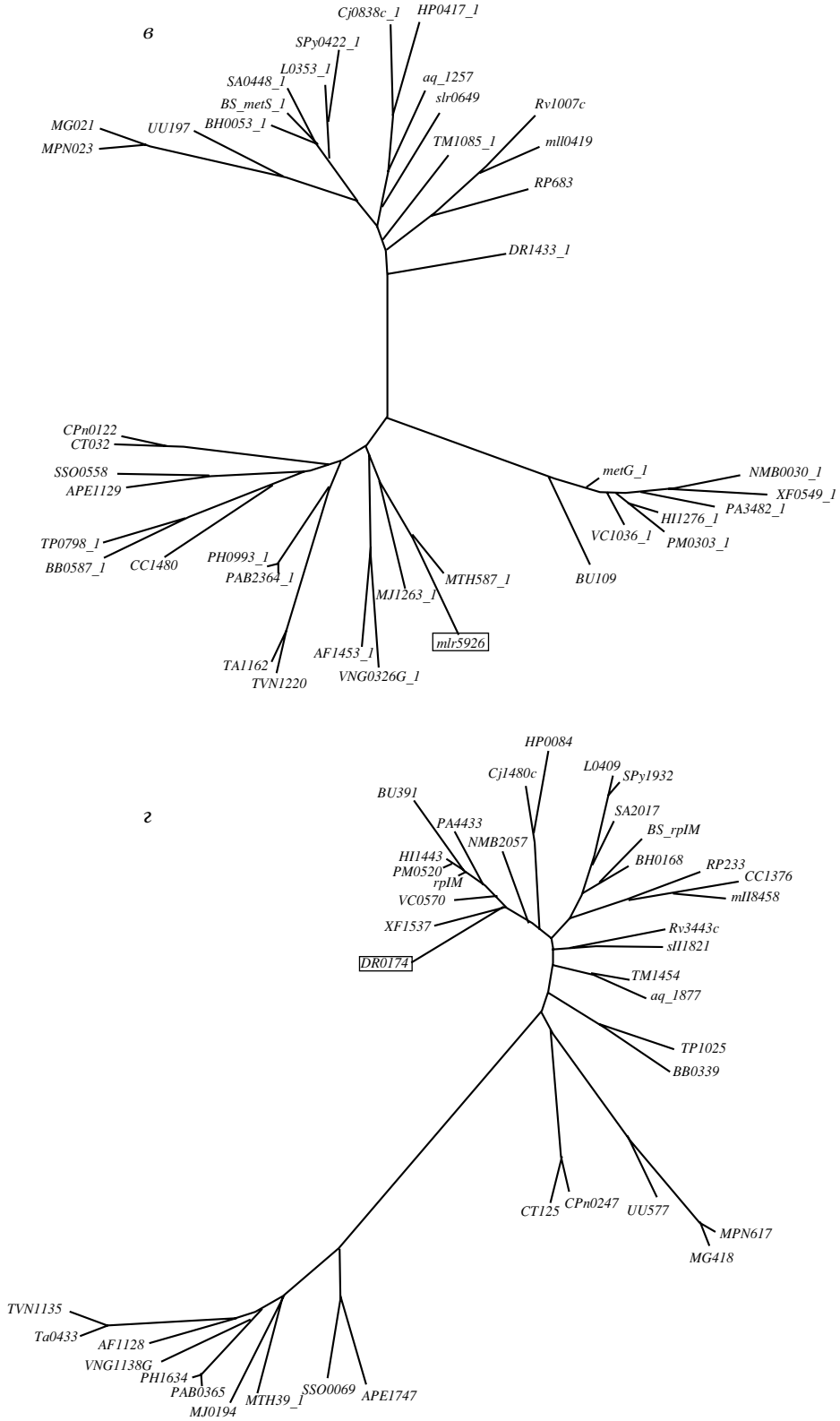


Рис. 2. Продолжение.

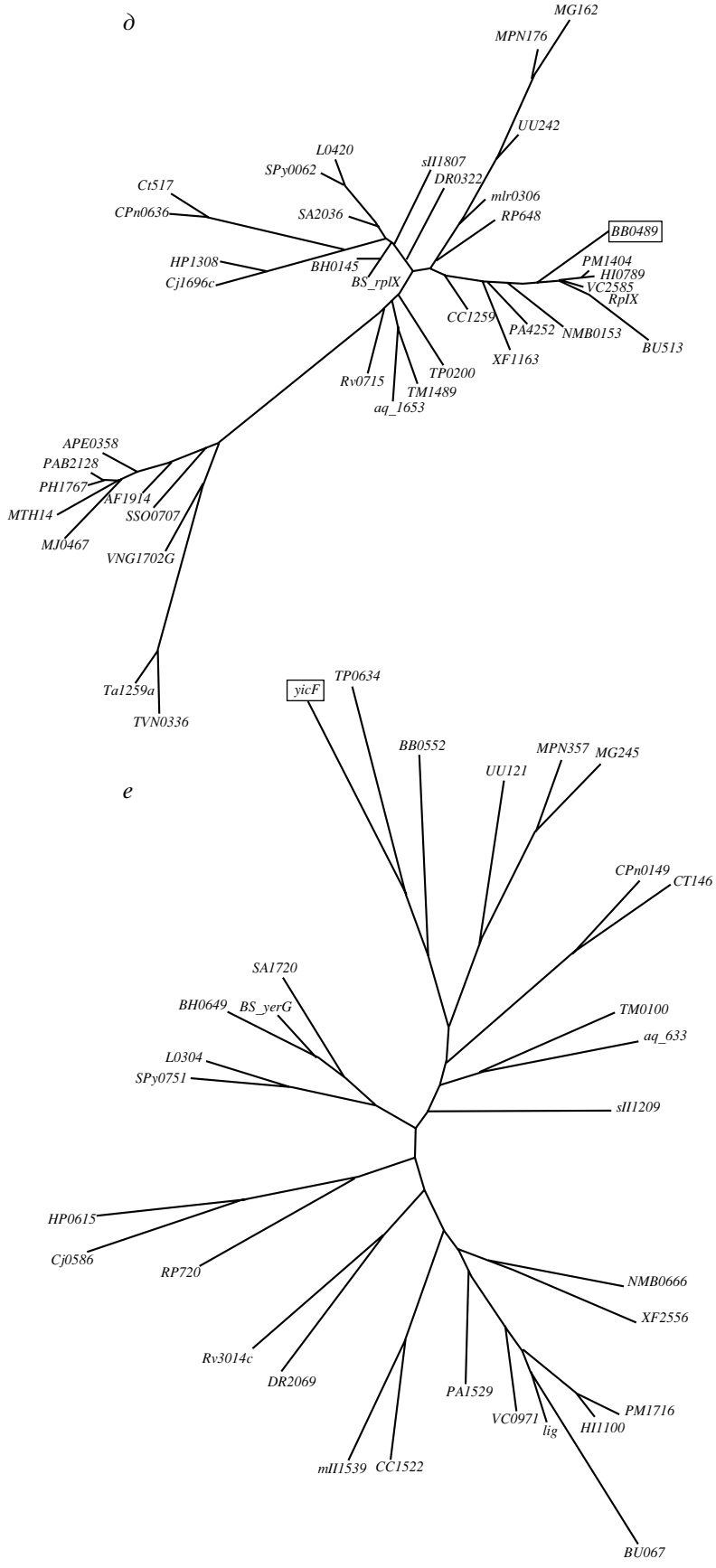


Рис. 2. Продолжение.

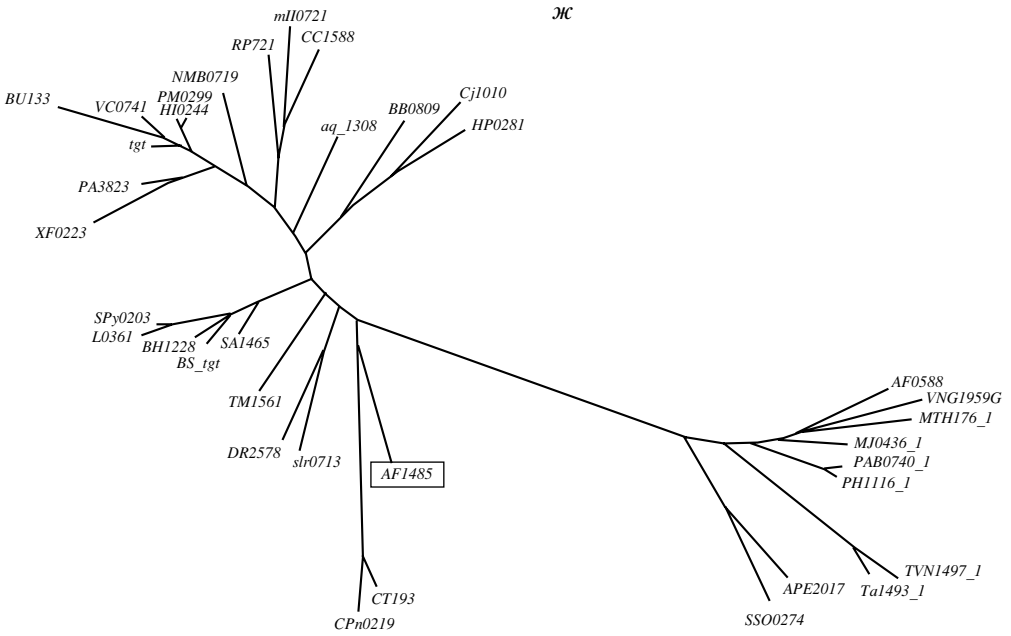


Рис. 2. Окончание.

положном направлении – от архебактерий к эубактериям – наблюдается в COG0143 (метионил-тРНК–синтетазы) (рис. 2*в*), где ген *mlr5926* из альфа-протеобактерии *Mesorhizobium loti*, происходящий из генома метаногенной архебактерии, возможно, был перенесен от архебактерий к эубактериям. Следует отметить, что это привело к появлению в геноме *M. loti* паралогичных генов, поскольку в этом геноме сохранился и первоначальный ген *mll0419*. Вообще, вопреки первоначальным представлениям, гены аминоацил-тРНК–синтетаз часто имеют сложную эволюционную историю, включающую дупликации и горизонтальные переносы [12].

В двух случаях наблюдается горизонтальный перенос генов рибосомных белков (рис. 2*г*). Ген *DR0174* из генома *Deinococcus radiodurans*, кодирующий белок L13, происходит из гамма-протеобактерий, а ген BB0489 (рис. 2*д*), кодирующий рибосомный белок L24, горизонтально перенесен из бета- или гамма-протеобактерий. Существование дупликаций и горизонтальных переносов в истории рибосомных белков показано также ранее [17].

И наконец, ген кишечной палочки *yicF*, кодирующий NAD-зависимую ДНК-лигазу (COG0271), перенесен из генома какой-то спирохеты; при этом в геноме *E. coli* содержится и "родной" ген *lig* (рис. 2*е*). Ген кьюоин/археозин-тРНК–рибозилтрансферазы *AF1485* из генома *Archaeoglobus fulgidus* (COG0343), по-видимому, имеет эубактериальное происхождение (рис. 2*ж*).

Экспертные оценки часто совпадают с результатами, которые получаются при нашем компьютерном анализе. Однако экспертный просмотр многих сотен эволюционных деревьев невозможен, поэтому одно из основных применений предлагаемого нами метода, по-видимому, может состоять в массовом анализе белковых семейств и выделении случаев, где имеются основания подозревать наличие горизонтального переноса, для дальнейшего более подробного анализа.

Хотя результаты приведенных здесь расчетов носят в известной мере предварительный характер, уже они позволяют сделать нетривиальные выводы. Во-первых, среди кластеров ортологов, в которых наблюдаются горизонтальные переносы, большинство таких, белки которых участвуют в основных информационных процессах; в основном, – в трансляции. По-видимому, это объясняется не тем, что такие гены специфически подвержены горизонтальному переносу, а тем, что соответствующие филогенетические деревья имеют ясную и хорошо разрешаемую структуру, что позволяет уверенно идентифицировать горизонтальные переносы. Выделение горизонтальных переносов в семействах транспортеров или регуляторов транскрипции требует более кропотливого анализа.

Во-вторых, часто наблюдается горизонтальный перенос без вытеснения первоначального гена. Это можно пытаться объяснить двумя способами. Либо мы наблюдаем некоторое промежуточное состояние, когда еще не зафиксировался выбор между двумя генами, выполняющими одну функцию, либо функции двух белков несколько различны. Это поддается экспериментальной проверке – как путем последовательной инакти-

вации двух вариантов, так и путем имитации горизонтальных переносов – внесения (перенесенного) варианта в геном, с или без инактивации оригинального гена.

Настоящая работа основана на предварительных публикациях [1, 2, 18].

Авторы благодарят Ю. Вульфа, предоставившего исходную выборку для счета, и Е. Кунина за полезное обсуждение, помощь и сотрудничество, а также рецензента за интересные и полезные для нас замечания.

Работа получила финансовую поддержку грантов Медицинского Института Говарда Хьюза (HHMI, 55000309) и Института исследования рака Людвига (LICR, CRDF RBO-1268).

СПИСОК ЛИТЕРАТУРЫ

1. Вьюгин В.В., Гельфанд М.С., Любецкий В.А. 2002. Согласование деревьев: реконструкция эволюции видов по филогенетическим деревьям генов. *Молекуляр. биология*. **36**, 650–658.
2. Вьюгин В.В., Любецкий В.А. 2001. Об одном алгоритме поиска горизонтального переноса генов на основе филогенетических деревьев белков. *Информационные процессы*. **1**, 167–177.
3. Logsdon J.M., Faguy D.M. 1999. Thermotoga heats up lateral gene transfer. *Curr. Biol*. **9**, 747–751.
4. Bacterial Conjugation. 1993. Ed. Clewell D.B., New York: Plenum Press.
5. Bergh O., Borsheim K.Y., Bratbak G., Heldal M. 1989. High abundance of viruses found in aquatic environments. *Nature*. **340**, 467–468.
6. Boucher Y., Doolitle W.F. 2000. The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol. Microbiol*. **37**, 703–716.
7. Lawrence J.G. 1997. Selfish operons and speciation by gene transfer. *Trends Microb*. **5**, 355–359.
8. Lawrence J.G. 1999. Gene transfer, speciation, and the evolution of bacterial genomes. *Curr. Opin. Microbiol*. **2**, 519–523.
9. Doolitle W.F. 1999. Lateral Genomics. *Trends Cell. Biol*. **9**, 5–8.
10. Lawrence J.G., Ochman H. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA*. **95**, 9413–9417.
11. Nelson K.E., Clayton R.A., Gill S.R. et al. 1999. Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thetmotoga maritima*. *Nature*. **399**, 323–329.
12. Yanai I., Wolf Y., Koonin E. 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*. **3**, Research 0024.
13. Koonin E.V., Makarova K.S., Aravind L. 2001. Horizontal gene transfer in prokaryotes: quntification and classification. *Annu. Review Microbiol*. **55**, 709–742.
14. Page R.D.M., Charlstone M.A. 1998. From gene to organismal phylogeny: reconciled trees and gene tree / species tree problem. *Mol. Phyl. Evol*. **7**, 231–240.
15. Page R.D.M. 1998. Genetree: comparing gene and species phylogenies using reconciled trees. *Bioinform. Appl. Notes*. **14**, 819–820.
16. 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**, 8.
17. Makarova K.S., Ponomarev V.A., Koonin E.V. 2001. Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol*. **2**(9), Research 0033.
18. Lyubetsky V.A., V'yugin V.V. 2002. Method of horizontal gene transfer determination using phylogenetic data. *Proc. Third Internat. Conf. Bioinform. Genome Regulat. Struct*. **2**, IC&G. Novosibirsk, 60–62.
19. 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 eukaryotes. *BMC Evol. Biol*. **3**, 2.

Identification of Horizontal Gene Transfer from Phylogenetic Gene Trees

V. V. V'yugin[1], M. S. Gelfand[2], and V. A. Lyubetsky[1]

[1] Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, 101447 Russia; E-mail: vyugin@iitp.ru

[2] State Research Center GosNIIGenetika, Moscow, 113545 Russia

Received October 29, 2002

We suggest a new procedure to search for the genes with horizontal transfer events in their evolutionary history. The search is based on analysis of topology difference between the phylogenetic trees of gene (protein) groups and the corresponding phylogenetic species trees. Numeric values are introduced to measure the discrepancy between the trees. This approach was applied to analyze 40 prokaryotic genomes classified into 132 classes of orthologs. This resulted in a list of the candidate genes for which the hypothesis of horizontal transfer in evolution looks true.