Информационные процессы, Том 10, № 2, 2010, стр. 140–144.

Об одном алгоритме согласования деревьев генов и видов с учетом дупликаций, потерь и горизонтальных переносов генов

К.Ю.Горбунов, В.А.Любецкий

Институт проблем передачи информации им. А.А. Харкевича, Российская академия наук, Москва, Россия

Поступила в редколлегию 28.04.2010

Аннотация—Для специалистов по информатике без биологических подробностей излагается и обсуждается алгоритм из работы авторов [Реконструкция эволюции генов вдоль дерева видов. Молекулярная биология, 2009, том 43, №5, с. 946-958]. Алгоритм согласует эволюционные события в данных деревьях генов и видов с учетом дупликаций, потерь и горизонтальных переносов генов.

1. ВВЕДЕНИЕ

Авторы в январе (в Москве) и в июне (в Монпелье, Франция) 2009 года читали курсы, в которых подробно излагали результаты работы [1], полученной журналом 09 Декабря 2008г. Слушатели — специалисты по computer science хотели иметь отдельное изложение алгоритма, оценку его сложности и доказательство корректности, которые содержатся в [1], но там соединяются с биологическими подробностями и тестированием алгоритма (эта работа публиковалась в биологическом журнале и была написана для биологов). Ниже содержится такое изложение. Чтобы сделать его менее зависимым от [1], напомним для специалиста по computer science на формальном уровне термины, биологическое содержание которых поясняется во многих работах по молекулярной эволюции и, в частности, тех, которые цитировались в [1]. Ген — это последовательность в четырехбуквенном алфавите, фиксированная в каждый момент времени. Дерево генов — дерево, вершинам которого приписаны гены, имеющие определенную степень родства. Вид (геном) — это множество генов, фиксированное в каждый момент времени. Дерево видов — дерево, вершинам которого приписаны виды, обычно имеющие определенную степень родства. В дереве генов и соответственно в дереве видов листьям соответствуют современные гены/виды (относящиеся к нашему времени), а предковым вершинам соответствуют предковые гены/виды (относящиеся к прошлым моментам времени). Ребра дерева видов S можно подразбить так, что образуются последовательные слои из новых и старых ребер, от корня (исходного состояния) к листьям (современному состоянию); в один слой будут входить ребра, “жившие в один период времени”, и, тем самым, между ними возможны одновременные события. В результате относительно длинные ребра в S будут разбиты на подребра; старые и новые ребра будем по прежнему называть ребрами в S. Алгоритм подразбиения ребер дерева видов на временные слои изложен в [1], он обладает естественным свойством: любое ребро располагается ровно в одном слое и любые два ребра, одно из которых является потомком другого, попадают в разные слои. В результате временной слой в дереве видов состоит из несравнимых ребер в S, находящихся на, так сказать, одинаковом расстоянии от корня. Сами деревья (без всяких приписываний) играют роль дискретного времени. Различия между деревом генов и деревом видов соответствует тому, что на генах и видах “время сказывается по разному”. С геном, входящим в данный вид, могут произойти события трех типов: он может скопироваться и остаться в том же виде (“дупликация”), он может

быть исключен из состава вида (“потеря”), он может скопироваться и одна копия перенестись в какой-то другой вид, а другая сохраниться в этом виде (“перенос с сохранением копии”) или потеряться (“перенос без сохранения”). С видом может произойти событие одного типа: он как множество может скопироваться и затем два полученных множества будут эволюционировать во времени независимо друг от друга (“видообразование”). Эволюция вида — это совместная эволюция всех составляющих его генов; в далеких видах эволюция идет независимо друг от друга в большей степени. В общем виде задача состоит в том, чтобы соотнести во времени события с генами с событиями с видами. Иными словами, соотнести время, в котором эволюционируют гены, и время, в котором эволюционируют виды. Такое соотнесение выполняется в форме вложения β дерева генов G в дерево видов S. Вложение β, которое вычисляет наш алгоритм, находится как минимум функционала H — взвешенной суммы трех величин: числа всех дупликаций, числа всех потерь и числа всех переносов (при данном вложении β из G в S); таким образом, H является функцией от β. Иными словами, эволюция описывается как развитие предкового гена в предковом виде, при котором происходит минимальное число эволюционных событий (“принцип парсимонии: parsimony”). Здесь предковые ген/вид определяются как те, которые приписаны соответственно корням деревьев G и S. Множество всех событий: дупликаций, потерь и переносов — вместе с моментами времени, когда они произошли, называется сценарием. Минимизация H означает поиск оптимального по H сценария. Значение H на некотором сценарии называется его ценой; соответственно задается цена каждого эволюционного события.

Опишем технические детали указанной выше картины. В исходных данных каждому листу дерева G приписан ген и каждому листу дерева S (за одним исключением, см. ниже) — вид. Для каждого гена в листьях указано, какому виду в листьях он принадлежит, одному виду может принадлежать несколько генов; это отношение между листьями двух деревьев называется ген-вид. Алгоритм строит оптимальное в смысле соответствующего сценария вложение одного бинарного дерева генов G в другое бинарное дерево видов S. Это вложение удобно представлять себе с помощью вспомогательного дерева G', которое имеет все вершины из G и кроме них имеет дополнительные вершины, у которых один из сыновей обязательно является листом и помечен крестиком; ребро от такой вершины к ее сыну назовем отростком; крестик означает потерю гена, приписанного этому сыну. Каждая развилка в S сопровождается развилкой в G' (видообразование сопровождается копированием всех генов). Вершины и ребра дерева G и дерева G', не помеченные крестиками, не различаются. Основное свойство G': без отростков G' изоморфно исходному дереву G. Дерево S может иметь вершины лишь с одним сыном (их появление связано с введением временных слоев). Ребра деревьев G, G' и S представляются идущими сверху вниз, т.е. корень находится сверху; ко всем деревьям добавляется по корневому ребру: это — ребро, идущее вверх от корня. Ребро дерева S называется трубой. Дерево G' располагается внутри труб дерева S; корневое ребро дерева G' (которое символизирует общего предка всех генов в G') располагается в корневой трубе дерева S (которая символизирует общего предка всех видов в S), а вершина дерева G' располагается в трубе или в развилке трубы. Труба не включает своих концов; термин “развилка” относится к концу трубы или ребра. В одной трубе может находиться несколько ребер дерева G', событие переноса изображается ребром дерева G', которое начинается в одной трубе и заканчивается в другой трубе, лежащей в том же временном слое. Перенос бывает двух типов: с потерей сына, оставшегося в исходной трубе в ней же, или без его потери в этой трубе. В первом случае перенос называется переносом без сохранения (копии в источнике), во втором — с сохранением. При вложении β каждому листу в G соответствует лист в S — тот вид, из которого взят данный ген; т.е. на листьях β совпадает с отношением ген-вид. Иными словами, каждый лист в G', не помеченный крестиком, располагается в соответствующем листе дерева S. В дереве S имеется так называемая аутгруппа, т.е. труба, идущая от корня сразу в отдельный лист, ко-

торому не приписан никакой вид; она также разбита временными слоями на меньшие трубы. Эта труба символизирует внешние по отношению к остальной части S виды, или возникновения/потери гена за счет мутаций. События, суммарную цену которых мы минимизируем, следующие: дупликация (т.е. развилка в дереве G' внутри трубы с сохранением обеих копий в этой трубе); потеря (т.е. отросток, помеченный крестиком); два типа переноса, описанные ранее; возникновение гена (т.е. перенос из аутгруппы, различия между переносом с сохранением и без сохранения здесь не делается) и перенос гена в аутгруппу без сохранения. Можно выделить еще три типа событий, цену которых мы считаем равной нулю, хотя это несущественно для нашего алгоритма: развилка в S, которая не сопровождается потерей ни одного из сыновей соответствующей развилки в G', дупликация в аутгруппе и дупликация в корневой трубе, все потомки хотя бы одной из копий которой перешли в аутгруппу уже в корне.

2. ФОРМУЛИРОВКА АЛГОРИТМА ИЗ [1]

Итак, даны деревья G и S. Напомним, что деревья G и G' (без отростков) изоморфны, поэтому говоря о ребре e в G, можно одновременно говорить и о ребре e в G'. Алгоритм перебирает пары $\langle$ребро e в дереве G, труба d в дереве $S\rangle$ и строит вспомогательную функцию $f(e,d)$, значение которой указывает, какое событие произошло на ребре e из G' в трубе d, в предположении, что e находится в d и вложение в дерево S поддерева дерева G с ребром e в качестве корневого оптимально. Когда $f(e,d)$ будет определено для всех пар $\langle e,d\rangle$, оно очевидным образом определит искомое оптимальное вложение β из G в S и одновременно определит дерево G'; для этого $f(e,d)$ используется как система ссылок, начиная с $f(e_0,d_0)$, где e_0 — корневая труба в G и d_0 — корневая труба в S.

Перебор происходит от листьев к корню, для каждого ребра e перебираются все трубы d, начиная от более поздних временных слоев, а внутри одного слоя первой берется труба из аутгруппы (т.е. одна из труб, возникших после подразбиения аутгруппы). Индукция начинается со случая, когда e и d ведут в листья, тогда $f(e,d) =$ "e принадлежит d", если e и d соответствуют друг другу как ген-вид; иначе $f(e,d') =$ "e перенесено без сохранения в ту d, которая соответствует e в отношении ген-вид". В индуктивном шаге для каждой пары $\langle e,d\rangle$ рассматривается несколько случаев, соответствующих тому, что может произойти с ребром e в трубе d и из них выбирается оптимальный, т.е. лучший в смысле функционала H:

1. ребро e разветвляется в трубе d на ребра e_1 и e_2 без потерь на них в этой трубе;
2. ребро e доходит до развилки трубы d и на его сыновьях e_1 и e_2, расположенных соответственно в сыновьях d_1 и d_2 трубы d, не произошли потери;
3. ребро e доходит до следующей трубы d_1 — единственного сына трубы d и переходит в него;
4. ребро e доходит до развилки трубы d с потерей ровно одного сына e_1 или e_2 соответственно в d_1 или d_2;
5. ребро e разветвляется в d на e_1 и e_2 с переносом одного сына e_1 или e_2 в другую трубу временного слоя, в котором находится d, и не лежащую в аутгруппе, а другой сын не теряется в d;
6. ребро e разветвляется в d на e_1 и e_2 и один из этих сыновей переносится в другую трубу временного слоя, в котором находится d и не лежащую в аутгруппе, а другой сын теряется в d; если перенос без сохранения считается самостоятельным событием, то в цену сценария такая потеря не включается, а если он рассматривается как композиция переноса с сохранением и потери, то его цена равна сумме цен упомянутых событий;
7. ребро e разветвляется в d на e_1 и e_2 и один из этих сыновей переносится в другую трубу временного слоя, в котором находится d и лежащую в аутгруппе, а другой сын теряется в d.

Первые пять случаев и случай 7 не нарушают индуктивное построение и легко позволяют определить $f(e, d)$ и использовать индуктивные ссылки; например, в случае 1 полагем $f(e, d) =$ "дупликация, $\langle e_1, d\rangle$, $\langle e_2, d\rangle$". Чуть сложнее шестой случай. Чтобы применить индукцию, нужно заглянуть на следующий шаг. На нем возможны опять те же семь случаев. Но два переноса без сохранения (случаи 6 и 6, идущие друг за другом) заведомо не оптимальны, так как могут быть заменены одним переносом. А оставшиеся случаи не нарушают индуктивное построение. Заметим: если труба d не лежит в аутгруппе, то можно не рассматривать и случай "6 и затем 5", так как он имеет такое же значение H, как если бы сначала из трубы d произошел перенос с сохранением, а затем из той же трубы перенос оставшейся в ней копии без сохранения. Если труба d лежит в аутгруппе, но цена переноса с сохранением не меньше цены возникновения, то по тем же соображениям случай "6 и затем 5" можно не рассматривать: сценарий с двумя возникновениями имеет не большее значение H. Далее случай "6 и затем 5" рассматривается только при d, лежащей в аутгруппе, и цене переноса с сохранением меньшей цены возникновения.

Обозначим $m.n$ последовательные случаи сначала m, затем n. Для сокращения времени работы алгоритма использовалось очевидное замечание: в случае 5 и в случаях 6.1, 6.2, 6.3, 6.4, 6.5 перебор труб, куда возможен перенос, устраняется стандартным приемом (широко известном в computer science): в процессе перебора пар $\langle e, d\rangle$ для каждого ребра e и каждого временного слоя нужно подсчитывать и хранить информацию о том, какая труба d' в этом слое является наилучшей для переноса ребра e в следующих смыслах (здесь d' не лежит в аутгруппе):

- для случая 5: в смысле минимальной цены сценария, начинающегося с пары $\langle e, d'\rangle$.
- для случая 6.1: в смысле минимальной суммы цен сценариев, начинающихся с пар $\langle e_1, d'\rangle$ и $\langle e_2, d'\rangle$, где e_1 и e_2 — сыновья ребра e;
- для случая 6.2: в смысле минимальной суммы цен сценариев, начинающихся с пар $\langle e_1, d'_1\rangle$ и $\langle e_2, d'_2\rangle$, где e_1 и e_2 — сыновья ребра e, а d'_1 и d'_2 — сыновья трубы d';
- для случая 6.3: в смысле минимальной цены сценария, начинающегося с пары $\langle e, d'_1\rangle$, где d'_1 — единственный сын трубы d';
- для случая 6.4: в смысле минимальной цены сценария, начинающегося с пары $\langle e, d'_1\rangle$, где d'_1 — какой либо из двух сыновей трубы d';
- для случая 6.5: в том же смысле, что и в случае 5.

Таким образом, на шаге алгоритма, соответствующем паре $\langle e, d\rangle$, лучшие для переноса трубы во всех случаях уже известны, так как они были вычислены при рассмотрении пар типа $\langle e_1, *\rangle$ и $\langle e_2, *\rangle$, где e_1 и e_2 — сыновья ребра e (случаи 5, 6.1, 6.2, 6.5) или при рассмотрении пар типа $\langle e, d'\rangle$, где d' находится в более позднем временном слое, чем d (случаи 6.3, 6.4). Если в каком-либо из случаев 5, 6.1, 6.2, 6.3, 6.4 лучшая труба для переноса — та самая d, в которой находится e, то перенос делать невыгодно и этот случай отпадает (в случае 5 перенос заменяется дупликацией; используется естественное предположение о том, что цена дупликации меньше цены переноса). В случае 6.5, если лучшие трубы d' и d'' для соответственно сыновей e_1 и e_2 ребра e совпадают, то этот случай отпадает, так как более выгодным становится сценарий с возникновением ребра e в трубе d' и последующей его дупликацией (снова используется упомянутое предположение). Если же d' не совпадает с d'', то случай 6.5 соответствует сценарию, в котором сначала происходит возникновение ребра e, скажем, в d', а затем разветвление ребра e и перенос ребра e_2 из d' в d'', т.е. перенос с сохранением. Отметим, что предположение о соотношении цен дупликации и переноса легко устранить, если всюду хранить не только лучшую трубу, но вторую по качеству.

3. ОЦЕНКА СЛОЖНОСТИ АЛГОРИТМА

Время работы алгоритма пропорционально произведению числа ребер дерева генов на число труб в дереве видов, уже разбитом на временные слои. При разбиении труб исходного дерева видов на временные слои алгоритмом из [1] число новых вершин не более, чем квадрат числа вершин n исходного дерева видов. Действительно, алгоритм построения временных слоев перебирает вершины от листьев к корню и каждой очередной вершине присваивает число — величину времени, на которое эта вершина отстоит от времени, соответствующем всем листьям. Все так присвоенные числа упорядочиваются по возрастанию, и в каждой трубе возникает столько новых вершин, сколько чисел попадает в интервал времени между числами, присвоенными концам этой трубы. Таким образом, каждая труба разбивается на не более, чем n труб, из чего видна искомая квадратичная оценка. Следовательно, время алгоритма имеет кубический порядок от величины: максимум из числа листьев в исходном дереве генов G и числа листьев в исходном дереве видов S.

4. КОРРЕКТНОСТЬ АЛГОРИТМА

Она следует из двух также очевидных соображений. Первое — список из семи возможных случаев, приведенный выше, является исчерпывающим, т.е. в предположении, что в каждый момент времени происходит лишь одно событие, ничего другого с ребром e в трубе d произойти не может. Второе — для пар $\langle e_1, d_1\rangle$ и $\langle e_2, d_2\rangle$, где e_1 и e_2 — сыновья ребра e, а d_1 и d_2 — возможно совпадающие трубы, оптимальные вложения независимы друг от друга и поэтому могут быть объединены в одно вложение для исходной пары $\langle e, d\rangle$. Важность этого условия легко понять, если попытаться распространить алгоритм с деревьев на филогенетические сети (т.е. ациклические ориентированные графы, у которых каждая вершина имеет не только двоих сыновей, но и двух родителей). Для такой сети наш алгоритм не работает, так как две подсети с корневыми ребрами e_1 и e_2 могут иметь общие вершины, что накладывает условие на согласованность пар соответствующих вложений: эти вершины должны попадать в одно и то же место. В то же время, если филогенетической сетью заменяется лишь дерево S, алгоритм работает и остается корректным.

СПИСОК ЛИТЕРАТУРЫ

1. Горбунов К.Ю., Любецкий В.А. Реконструкция эволюции генов вдоль дерева видов. *Молекулярная биология*, 2009, том 43, № 5, стр. 946–958.