АЛГОРИТМИЧЕСКИЙ АНАЛИЗ СТРУКТУРНОЙ ИНФОРМАЦИИ
(Вычислительные системы)
1985 год Выпуск 112

УДК 577.2

АЛГОРИТМ ПОСТРОЕНИЯ АДДИТИВНЫХ ДЕРЕВЬЕВ ПО НАБОРУ ГОМОЛОГИЧНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ. ДОСТОВЕРНОСТЬ ВОССТАНОВЛЕНИЯ ФИЛОГЕНИЙ

Л.В.Омельянчук, Н.А.Колчанов

ДНК и РНК – это линейные биополимеры, состоящие из нуклеотидов четырех типов, обозначаемых как A, U, G, C, следующих друг за другом в определенной последовательности [1]. В ходе эволюции за счет накопления нуклеотидных замен из общей предковой молекулы возникает набор сходных по последовательности (гомологичных) молекул. Этот процесс может быть описан деревом, висячие вершины которого соответствуют современным молекулам, а внутренние – предковым молекулам. Существуют два подхода к построению филогенетических деревьев (т.е. деревьев, отражающих происхождение). Один из них основан на так называемом принципе максимальной парсимонии, согласно которому процесс дивергенции семейства последовательностей описывается деревом, имеющим минимально возможное число мутаций. Разработанный нами метод основан на другом подходе, использующем принцип совместимости [2], согласно которому реальное эволюционное дерево соответствует максимальному набору совместимых позиций.

В настоящей работе дано изложение этого метода с позиций теории графов. Доказательство так называемой теоремы парной совместимости проведено существенно более простым способом, чем в работе [3]. Кроме того, сформулировано новое условие, выполнение которого позволяет избежать ошибок, возникающих при восстановлении эволюционных деревьев на основе принципа совместимости.

А д д и т и в н ы е д е р е в ь я. В качестве характеристики сходства двух последовательностей i и j используют прямое расстояние между ними D_{ij}, равное числу позиций, по которым они различаются. Помимо прямого расстояния, между последовательностями

обычно используют и понятие расстояния между двумя представителями по дереву [4]. Пусть внутренним вершинам дерева приписаны последовательности той же длины, что и последовательности гомологичного семейства. Весом ребра назовем количество приписанных ребру замен, равное прямому расстоянию между последовательностями, соответствующими концевым вершинам этого ребра. Расстоянием по дереву между вершинами i и j будем называть сумму весов ребер единственной цепи, соединяющей i и j .

Будем называть дерево аддитивным, если прямое расстояние между любыми двумя вершинами равно расстоянию по дереву между ними. Вариабельностью v позиции последовательностей гомологичного семейства будем называть число, на единицу меньшее числа различных нуклеотидов, встречающихся в этой позиции у всех последовательностей данного семейства [4]. Будем рассматривать только последовательности, образованные нуклеотидами A, U, G, C. Ясно, что позиции с v = 0 являются инвариантными и не несут информации о дивергенциях рассматриваемых последовательностей. Позиции с v = 2,3 не могут входить в набор позиций, по которым может быть построено аддитивное дерево. Действительно, пусть позиция с v = 2 содержит символы A, U, G. Минимальное число замен, в результате которых могли возникнуть эти символы, равно 2. Прямые расстояния между всеми парами последовательностей, содержащими различные символы по этой позиции, равны I. Единственное дерево в этом случае содержит 3 ребра. Очевидно, что разместить две и более замен на ребрах этого дерева таким образом, чтобы между всеми парами представителей расстояние по дереву (в числе замен) было равно единице, невозможно.

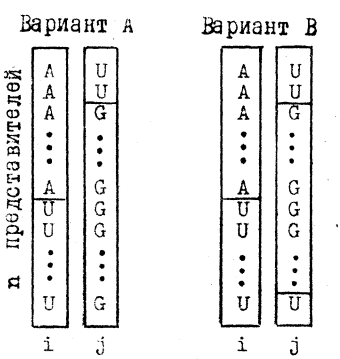


Рис.I

Ограничимся далее рассмотрением позиций с v = 1.

Рассмотрим две позиции i и j гомологичного семейства. Любому типу нуклеотида в позиции i соответствует определенный тип нуклеотида в позиции j, находящийся в той же последовательности. Будем говорить, что позиции i и j совместимы, если по этим позициям отсутствует хотя бы одна из четырех потенциально возможных комбинаций символов. Так, на рис.I вариант A соответствует паре совмести-

мых позиций, а вариант B - пара несовместимых позиций.

УТВЕРЖДЕНИЕ I. Если дерево аддитивно, то все позиции имеют v=1 и совместимы.

В силу сделанных выше замечаний рассматриваемые позиции должны иметь v =1. Предположим, что существуют две несовместимые позиции с v =1 аддитивного дерева. Это означает,что существуют четыре группы последовательностей, содержащие по рассматриваемым позициям все четыре возможные комбинации символов. Обозначим их через t_1s_1 t_1s_2 t_2s_1 t_2s_2. Матрица расстояний между этими группами последовательностей по рассматриваемым позициям имеет вид:

	t_1s_1	t_1s_2	t_2s_2	t_2s_1
t_1s_1	0	I	2	I
t_1s_2		0	I	2
t_2s_2			0	I
t_2s_1				0

Заметим, что если на каком-либо ребре дерева, имеющего четыре висячие вершины, расположены две замены, то расстояние по дереву между этой вершиной и остальными тремя больше или равно 2, что несовместимо с матрицей расстояний. Ограничиваясь случаем, когда на каждом ребре расположена одна либо ни одной замены, достаточно перебрать все возможные расположения двух,трех, четырех и пяти мутаций на деревьях,имеющих четыре висячие вершины,и убедиться, что ни одно из них не удовлетворяет рассматриваемой матрице расстояний. На рис.2 приведены все такие деревья и спектр расстояний между висячими вершинами для каждого из них. Можно видеть,что ни один из них не совпадает с набором элементов матрицы расстояний.

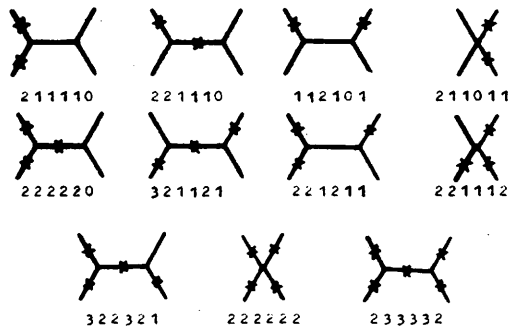


Рис.2

УТВЕРЖДЕНИЕ 2. Если набор позиций с v=1 таков, что все позиции совместимы,

то по нему можно построить аддитивное дерево.

Будем доказывать это утверждение по индукции. Рассмотрим две совместимые позиции. Поскольку позиции совместимы, то имеются следующие комбинации нуклеотидов по этим позициям: либо $s_1t_1\ s_2t_1\ s_2t_2$, либо $s_1t_1\ s_1t_2\ s_2t_2$, что одно и то же с точностью до замены s на t и t на s. Возможность построения аддитивного дерева в этом случае очевидна (рис.3,а).

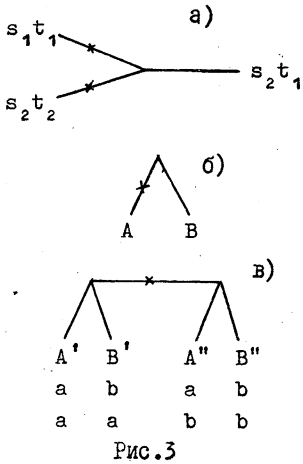


Рис.3

Пусть теперь имеется набор из n совместимых позиций, для которых существует аддитивное дерево. Добавим к этому набору новую (n+1)-ю позицию, совместимую со всеми n позициями. Покажем, что в этом случае аддитивное дерево можно построить для полного набора из (n+1)-й позиции.

Новая позиция n+1 содержит два типа нуклеотидов s и t, которые соответствуют двум непересекающимся группам последовательностей A и B. Добавление новой позиции не изменяет расстояний между последовательностями внутри групп A и B и увеличивает на единицу расстояние между последовательностями, принадлежащими различным группам. Пусть в аддитивном дереве, построенном по n позициям, есть ребро, делящее дерево на две части такие, что висячие вершины в одной части - это последовательности группы A, а другой - группы B. Тогда достаточно поместить новую замену, отличающую последовательности группы A от последовательностей группы B по (n+1)-й позиции, на этом ребре, чтобы получить аддитивное дерево, содержащее (n+1) позицию (рис.3,б).

Пусть в дереве нет ребра, которое делит последовательности указанным образом. Это означает, что существует ребро, делящее последовательности на части A'B' и A"B", где A'∪A"=A и B'∪B"=B (нет общего предка у последовательностей группы A). Если вес этого ребра равен 0, то можно заменить это дерево на другое, в котором последовательности A'∪A" и B'∪B" соединены ребром длины 0; при этом мы возвращаемся в ситуацию, описанную выше. Если длина рассматриваемого ребра не равна 0, то реализуется ситуация, изобра -

женная на рис.3,в. В этом случае любая замена, которая произошла на этом ребре (на рис.3,в эта замена обозначена "x"), в одной из n исходных позиций несовместима с (n+1)-й позицией, что противоречит условию.

Объединяя утверждения 1 и 2, получаем

УТВЕРЖДЕНИЕ 3. Для данного набора позиций аддитивное дерево существует тогда и только тогда, когда две любые позиции из этого набора совместимы и каждая позиция имеет v=1.

Это утверждение представляет собой новый вариант теоремы парной совместимости, доказанной в работе [3].

Практически находить наборы совместимых позиций можно следующим образом. Каждой позиции с вариабельностью 1 сопоставим вершину графа G . Две вершины соединим ребром, если соответствующие им позиции совместимы. Каждый максимальный набор совместимых позиций в этом случае будет соответствовать клике [5] графа G. По каждому найденному набору совместимых позиций можно построить аддитивное дерево по алгоритму, рассмотренному в данной работе при доказательстве утверждения 2, либо по алгоритму "уникальных" замен А.А.Харких [4].

Анализ аддитивных деревьев. Выше был дан способ построения всех возможных максимальных по позициям аддитивных деревьев для данного набора гомологичных макромолекул (поскольку можно найти все клики). Рассмотрим теперь вопрос о том, какое из этих деревьев соответствует процессу дивергенций от общей предковой последовательности. Введем для этого понятие аддитивной позиции, т.е. такой, в которой замена фиксировалась только один раз при дивергенции последовательностей данного семейства от общего предка.

Рассмотрим дерево, описывающее процесс дивергенций реальных последовательностей. Исключим из рассмотрения все неаддитивные позиции. Очевидно, что по оставшимся позициям реальное дерево будет аддитивным. На одном наборе аддитивных позиций можно построить только одно аддитивное дерево. (Это утверждение доказано для последовательностей, состоящих только из аддитивных позиций в [6], эквивалентное утверждение было доказано в теории графов [7].) Поэтому аддитивное дерево, построенное по аддитивным позициям, будет совпадать с реальным деревом. На рис.4 показан пример правиль-

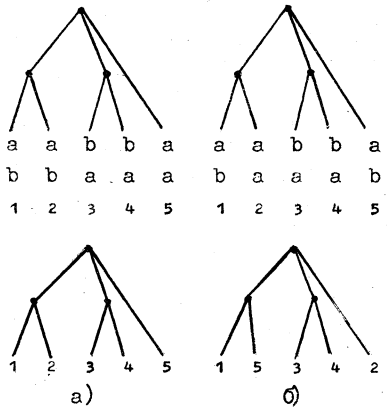


Рис.4

ного восстановления реального дерева по аддитивным позициям (а) и неправильного (б) - по набору аддитивных и неаддитивных позиций.

Таким образом, из всего множества аддитивных деревьев, которые соответствуют кликам графа G , реальному процессу будет соответствовать только то дерево, которое построено по аддитивным позициям. Поэтому вопрос о построении дерева, описывающего реальный процесс дивергенций, сводится к выбору множества аддитивных позиций.

Проблема выявления аддитивных позиций для семейств гомологичных макромолекул является сложной. Это обусловлено тем, что всякая аддитивная позиция имеет вариабельность I, но не всякая позиция с вариабельностью I является аддитивной, так как она могла промутировать два раза с возвращением в исходный нуклеотид. Поэтому, имея набор гомологичных последовательностей, можно выделить для него все позиции с $v=1$, но относительно каждой такой отдельно взятой позиции нельзя утверждать, что она является аддитивной. Однако можно сформулировать некоторый критерий, при выполнении которого максимальный набор совместимых позиций образован только аддитивными позициями.

Рассмотрим дерево, описывающее реальную дивергенцию семейства гомологичных последовательностей.

УТВЕРЖДЕНИЕ 4. Пусть каждое ребро этого дерева содержит хотя бы одну замену, которая привела к образованию аддитивной позиции. Покажем, что в этом случае все аддитивные позиции образуют клику в графе G.

ДОКАЗАТЕЛЬСТВО. Поскольку все аддитивные позиции совместимы, достаточно доказать, что для любой неаддитивной позиции найдется хотя бы одна аддитивная, такая, что эти две позиции несовместимы.

Рассмотрим произвольную неаддитивную позицию x этого семейства, которая имеет вариабельность I и в которой предковый нуклеотид

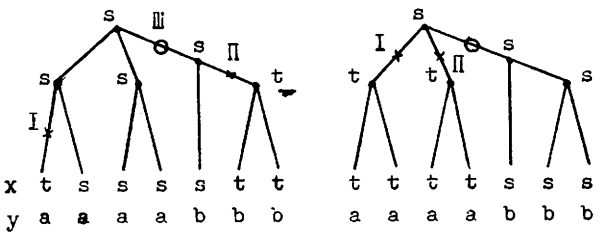


Рис.5

в два раза заменился на нуклеотид t. Пусть эта замена произошла на некоторых ребрах I и П.*) Рассмотрим аддитивную позицию y (по условию она найдется), замена в которой произошла на ребре Ш, расположенном между ребрами I и П (рис.5). По одну сторону от этого ребра все последовательности в позиции y имеют нуклеотид a , а по другую - b . Понятно, что в этом случае реализуются все комбинации типов нуклеотидов по паре позиций xy - ta sa sb tb . Поэтому рассматриваемые позиции x и y несовместимы.

Доказанное утверждение гарантирует, что аддитивные позиции соответствуют одной из клик графа G , если набор последовательностей подобран таким образом, что дерево, описывающее процесс дивергенций последовательностей, на каждом ребре имеет хотя бы одну аддитивную позицию (т.е. позицию, содержащую аддитивную замену). Однако оно не дает критерия, на основании которого можно было бы среди множества клик графа G выделить клику, образованную аддитивными позициями.

Если число замен, в результате которых образовалось гомологичное семейство, мало, то почти все промутировавшие позиции будут являться аддитивными, поскольку вероятность двукратного и более чем двукратного попадания замены в одну позицию будет мала. В этом случае клика, в которую входит наибольшее число позиций, будет соответствовать множеству аддитивных позиций.

Рассмотрим теперь отношение числа аддитивных позиций к числу позиций с вариабельностью I в зависимости от числа мутаций, фиксировавшихся в семействе макромолекул. Вероятность того, что данная позиция аддитивна (т.е. содержит одну замену), если гомологическое семейство образовалось в результате фиксации N замен (в

*) Если эти замены произошли на соседних ребрах (рис.5,б), то такая неаддитивная позиция не может исказить реальное дерево, восстанавливаемое по аддитивным позициям, так как совпадает с одной из аддитивных позиций с точностью до замены нуклеотидов.

предположении равновероятности фиксации замен по позициям гомологичного семейства), равна:

$$P_1 = C_N^1 \frac{1}{L}\left(1 - \frac{1}{L}\right)^{N-1} .$$

Пусть вероятность замены любого нуклеотида на любой другой равна I/3, тогда вероятность того, что в данной позиции с вариабельностью I произошла фиксация одной или более замен, равна:

$$\Phi = \sum_{k=1}^{N} C_N^k \left(\frac{1}{L}\right)^k \left(1 - \frac{1}{L}\right)^{N-k} \cdot \frac{1}{3^{k-1}} .$$

К-й член суммы равен вероятности иметь ровно k замен по данной позиции при условии, что после каждой замены позиция имеет вариабельность I, т.е. при последовательных заменах происходит чередование двух символов.

Выражение для Φ может быть преобразовано к виду:

$$\Phi = 3\left\{\left(1 - \frac{2}{3L}\right)^N - \left(1 - \frac{1}{L}\right)^N\right\} .$$

Отношение среднего числа аддитивных позиций к среднему числу позиций с вариабельностью I равно:

$$\frac{P_1}{\Phi} \cong \frac{N}{3L\left(\left(1 + \frac{1}{3L}\right)^N - 1\right)} .$$

Функция P_1/Φ убывает при увеличении N. Существует, однако, большой интервал значений N ($N < L$), при которых аддитивных позиций среди позиций с вариабельностью I больше, чем остальных замен. Приведенная оценка указывает границы, в которых клика графа G, максимальная по количеству вершин, соответствует дереву дивергенций реальных последовательностей.

Возможна также чисто качественная оценка достоверности построенного дерева. Каждое дерево будем характеризовать числом совместимых позиций. Распределение деревьев по числу совместимых позиций будем называть "спектром". Если "спектр" имеет вид, как на рис.6,а, т.е. существует дерево, явно выделяющееся по числу позиций, то следует ожидать, что это дерево хорошо соответствует реальному. Если же "спектр" таков, как на рис.6,б, то решение о выборе дерева, соответствующего реальным дивергенциям, не производится.

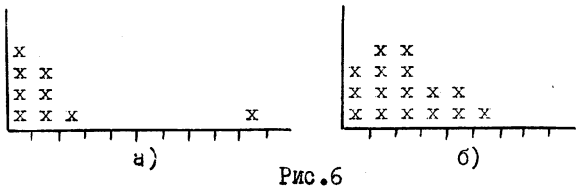
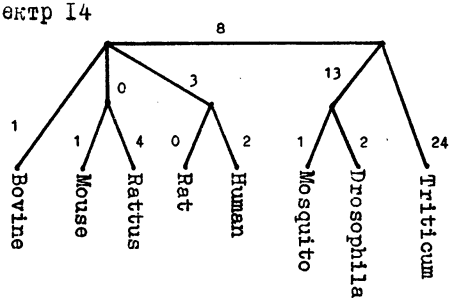
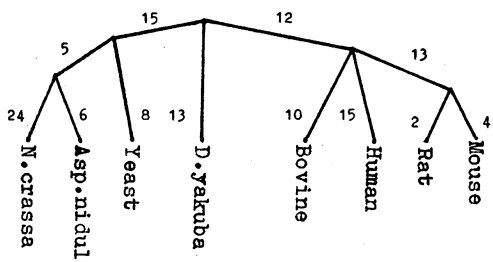


Рис.6

Спектр I4



Спектр 2448



Спектр 66778

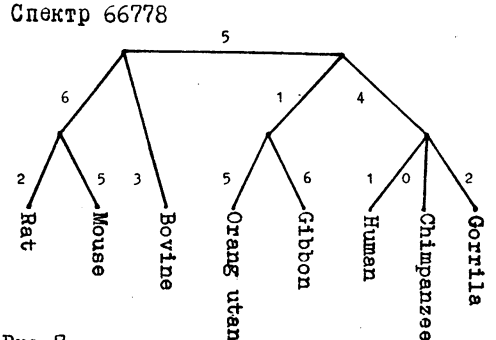


Рис.7

Алгоритм построения филогенетического дерева состоит в следующем. Осуществляется поиск позиций с вариабельностью I. Для этих позиций указанным способом строится граф совместимых позиций G . С помощью алгоритма поиска клик графа [8] находятся все клики графа G .По кликам восстанавливаются деревья приведенным выше способом.Длины ребер дерева восстанавливаются по алгоритму из работы [9]. Алгоритм реализован на языке Фортран IУ для ЭВМ СМ-4.

Филогения митохондриальных тРНК.Для иллюстрации работы алгоритма были восстановлены филогенетические деревья ряда семейств митохондриальных тРНК. На рис.7 приведены филогенетические деревья семейств Met, Val, His и их "спектры" аддитивных деревьев (соответственно спектр I4,спектр 2448,спектр 66778). "Спектр" се-

мейства Met состоит из единственного дерева, что означает высо - кую достоверность филогении. "Спектр" семейства Val состоит из четырех деревьев, и число позиций в одном из них в два раза превышает число позиций в остальных. В этом случае достоверность полученного дерева так же высока. "Спектр" для семейства His содер - жит пять деревьев, несущественно отличающихся по числу позиций. На рис.7 изображено дерево, максимальное по числу позиций. В этом случае достоверность полученной картины дивергенций сравнительно низка, хотя полученное дерево совпадает с данными классической таксономии.

Литература

1. РАТНЕР В.А., ЖАРКИХ А.А., КОЛЧАНОВ Н.А. и др. Проблемы теории молекулярной эволюции/Под ред. Р.И.Салганика.-Новосибирск: Наука,1984. - 340 с.

2. FELSENSTEIN J. Numerical methods for inferring evolutio - nary trees. - The Quarterly Review of Biology,1982, v.57, N 4,p.- 379-404.

3. ESTABROOK G.F., McMORRIS F.R. When is one estimate of evo- lutionary relationships a refinement of another? - J.Math.Biolo - gy,1980,v.10,N 4,p.367-373.

4. ЖАРКИХ А.А. Алгоритм построения филогенетических древ по аминокислотным последовательностям. - В кн.: Математические моде- ли и селекции. Новосибирск,1977,с.5-52 (ИЦиГ СО АН СССР).

5. ХАРАРИ Ф. Теория графов.- М.:Мир,1973.-284 с.

6. WATERMAN M.S., SMITH T.F.,SINGH M.,BEYER W.A. Additive evolutionary trees.- J.Theor.Biol.,1977,v.64,p.199-213.

7. СМОЛЕНСКИЙ Е.А. Об одном способе линейной записи графов.- Журнал вычислит. математики и мат.физики, 1962,т.2,№2,с.371-372.

8. БЕССОНОВ Ю.Е., СКОРОБОГАТОВ В.А. Применение относительных разбиений для поиска клик. - В кн.: Автоматизация проектирования в микроэлектронике.Теория.Методы.Алгоритмы (Вычислительные систе- мы,вып.77).Новосибирск,1978,с.24-33.

9. HARTIGAN J.A. Minimum mutation fits to a given tree. - Biometrics,1973,v.29,N 1,p.53-65.

Поступила в ред.-изд.отд.
10 июля 1985 года