Информационные процессы, Том 5, № 5, 2005, стр. 380–391.

Удаление шума в множественном выравнивании белковых последовательностей[1]

В.А.Любецкий, К.Ю.Горбунов, В.В.Вьюгин, Л.Ю.Русин

Институт проблем передачи информации РАН
lyubetsk@iitp.ru, rusin@iitp.ru

Поступила в редколлегию 20.10.2005

Аннотация—Одна из основных проблем при построении дерева белкового семейства состоит в получении качественных первичных молекулярных данных (в том числе, множественного выравнивания семейства) для дальнейшего проведения филогенетического анализа. Авторами предлагается процедура удаления шума из исходного множественного выравнивания (МВ) белковых последовательностей с целью повышения качества самого МВ и филогенетического дерева белкового семейства, построенного на его основе. Для этого определены энтропия и условная энтропия (филогенетическая информативность) каждой колонки МВ, и предложен алгоритм для нахождения по исходному МВ нового МВ, наиболее информативного для построения филогенетического дерева семейства. Эффективность предложенного подхода проверена на многих белковых семействах из базы данных по кластерам ортологичных групп белков (COGs, NCBI) с помощью вычисления показателя правдоподобия для всех итеративно возникающих деревьев. В заметке численные результаты тестирования приводятся для пяти КОГов.

1. ВВЕДЕНИЕ

Одна из основных проблем при построении дерева белкового семейства состоит в получении множественного выравнивания (МВ) аминокислотных или нуклеотидных последовательностей этого исходного белкового семейства; при этом важно оценивать и качество семейства. Поскольку изучение процессов эволюции геномов во многом основано на сравнительном анализе больших наборов филогенетических деревьев белковых семейств, качество такого анализа напрямую зависит от качества МВ, по которому строится отдельное дерево G белкового семейства.

Исходные белковые семейства могут содержать некорректно аннотированные последовательности и ошибки секвенирования. С другой стороны, особенности эволюции биологических макромолекул приводят к мутационному насыщению тех их участков, в которых молекулярные замещения идут с повышенной скоростью. Это приводит к сокращению числа гомологичных признаков и к потере филогенетического сигнала в высоко вариабельных областях макромолекул [1]-[3]. Поэтому выявление и исключение из исходного МВ шума различного происхождения, т.е. удаление отдельных строк и столбцов из МВ — носителей шума, повышает качество филогенетического дерева G, которое строится на основе таким образом улучшенного множественного выравнивания $MB*$, по сравнению с филогенетическим деревом G_0, построенным по исходному множественному выравниванию MB_0.

Задача настоящей работы состоит в формальном определении признаков (колонок) множественного выравнивания, которые несут шум в противоположность колонкам, несущим филогенетический сигнал, и в разработке быстрого алгоритма удаления шума и выявления филогенетического сигнала в множественном выравнивании.

[1] Работа выполнена при частичной поддержке гранта МНТЦ №2766.

Для устранения шума из исходного МВ сначала удаляются "мусорные" строки (в нашем алгоритме — строки с низким средним значением индекса CORE), а затем и столбцы с высоким значением введенного нами показателя "вариабельности" или "рассогласованности" столбца. По МВ, оставшемуся после удаления строк и столбцов, строится дерево белкового семейства. Белковое семейство обычно берется из числа кластеров ортологичных групп белков — КОГов из базы COGs [4].

Для отбора удаляемых столбцов нами определяется численная характеристика такой "вариабельности" или "рассогласованности" столбца в МВ, называемая энтропией или условной энтропией, и столбцы с высоким значением этого показателя последовательно удаляются по несколько штук ("порциями") до тех пор, пока доля q неразрешенных квартетов (см. ниже) не достигнет первого локального минимума. Численная проверка на 50 КОГах (в таблице приведены подробные результаты работы алгоритма на пяти из этих КОГов) показала, что показатель правдоподобия деревьев G_i, получаемых на этапах $i = 0, 1, 2, \ldots$ последовательного удаления порций столбцов с использованием нашего алгоритма, достигает минимума одновременно с минимумом доли неразрешенных квартетов в текущем выравнивании MB_i. При этом показатель правдоподобия всегда улучшается, и примерно в половине случаев это улучшение статистически значимо по тестам из [13]-[16].

В разделе 2 кратко перечисляются используемые нами программные средства, которые общедоступны, но их аккуратное и комплексное изложение, включая описание соответствующих алгоритмов или хотя бы соответствующей терминологии, по крайней мере, на русском языке отсутствует. Эти программные средства вместе с нашей программой реализуют предложенную процедуру: от построения множественного выравнивания ("исходное МВ") до построения дерева КОГа по новому МВ, полученному в результате "улучшения" исходного МВ.

2. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ

1) *Построение МВ.* Для выравнивания белковых последовательностей нами использовались алгоритм и программа PROBCONS [5].
2) *Выбор модели эволюции* осуществлялся на основе информационных критериев AIC1, AIC2 и BIC [6], [7], реализованных программой ModelGenerator v.06 [8].
3) *Бутстреп-анализ.* Реплики МВ генерировались в количестве 1000 штук программой seqboot из пакета PHYLIP v. 3.63 [9]. Матрицы максимально-правдоподобных расстояний вычислялись программой TreePuzzle v.5.02 [10] с участием вспомогательной программы PuzzleBoot [11]. *Бинарные деревья реплик* строились кластерным методом объединения ближайших соседей программой neighbor, и затем объединялись в *бинарное* дерево — консенсус из клад упомянутых деревьев ("компонентный консенсус") программой consense из пакета PHYLIP v. 3.63. Или объединялись в *небинарное* дерево — 50%консенсус, который уже содержит политомические вершины. Каждой вершине последнего дерева приписан процент бутстрепа, отражающий ее "надежность" (или, что тоже самое, надежность клады, соответствующей ей). В обоих случаях так полученные деревья называются *консенсусными.*
4) Вычисление доли *q неразрешенных квартетов* от числа всех квартетов ("максимально-правдоподобное картирование", [12]) и вычисление показателя *максимального правдоподобия* последовательно возникающих деревьев G_i белкового семейства выполнялись с использованием программы TreePuzzle v.5.02. Для оценки *значимости отличия* показателя максимального правдоподобия для возникающих деревьев эта программа использует статистические тесты: ожидаемые правдоподобные веса (ELW), и более консервативные — односторонний и двусторонний Кишино-Касигава, и Шимогайра-Хасигава [13]-[16].
5) *Реконструкция предковых последовательностей* в 50%консенсусном дереве белкового семейства проводилась с использованием пакета PAML v.3.14 [17].

6) Вычисление встречаемости вхождений признаков (например, аминокислот) в МВ по сравнению с их встречаемостью во всех возможных попарных выравниваниях строк МВ, т.е. *соответствие множественного и попарных выравниваний*, проводилось с помощью индекса CORE [18], который вычислялся программой T-COFFEE v. 2.11 [19].
7) Для определения *условной энтропии* и *количества информации* в колонке МВ относительно списка надежных клад использовались метод и программа, разработанные авторами, см. ниже разделы 3.2-3.3.
8) *Реконструкция филогении*, т.е. построение дерева белкового семейства, выполнялась программой MrBayes v.3.0 [20] с параметрами марковских цепей, установленными в соответствии с выбранной моделью эволюции белков (тип эмпирической матрицы аминокислотных замещений, доля инвариантных сайтов, гамма-поправка на разницу в скоростях эволюции сайтов). *Постериорные вероятности* клад вычислялись с учетом данных по 1.000.000 генераций марковских цепей с записью параметров цепи и соответствующих топологий после каждой сотни генераций.

3. УДАЛЕНИЕ ШУМА В ИСХОДНОМ МНОЖЕСТВЕННОМ ВЫРАВНИВАНИИ

3.1. Удаление низко-гомологичных последовательностей

После построения исходного МВ аминокислотных последовательностей строки с низким уровнем взаимной гомологии удаляются. Филогенетические связи таких последовательностей не могут быть достоверно установленными, и их присутствие искажает топологию дерева из-за эффекта притяжения длинных ветвей [21], [22]. Для этого в исходном МВ для каждого вхождения буквы (кроме пробела), $R_q^{S_x}$, вычисляется индекс CORE по формуле $\mathrm{CORE}(R_q^{S_x}) = \sum_{y=1,y\neq x}^{N} \mathrm{CS}(R_q^{S_x}, R_r^{S_y})/(N-1)$, где q и r — аминокислотные остатки, выровненные друг с другом в последовательностях S_x и S_y, N — число последовательностей, и CS — нормализованный показатель (*consistency score*), отражающий встречаемость этих остатков в одной колонке множественного выравнивания по отношению к их со-встречаемости в множестве возможных (локальных и глобальных) попарных выравниваний S_x и S_y. Индекс $\mathrm{CORE}(R_q^{S_x})$, усредненный по всем остаткам в последовательности, определяет индекс sCORE и служит общим показателем надежности, с которой данная последовательность может быть встроена в множественное выравнивание. Считается, что последовательности со значением sCORE меньше порога не могут дать информативное выравнивание, они исключаются из набора данных. Детальное описание показателя соответствия CS и индекса CORE приводится в [18].

3.2. Вычисление условной энтропии и количества информации в колонке МВ относительно списка надежных клад

Здесь *задача 1* состоит в ранжировании колонок МВ по степени их согласованности со списком надежных клад; затем "худшие" в этом смысле колонки удаляются.

Простейший, но иногда эффективный способ состоит в ранжировании колонок просто по убыванию величины *энтропии*

$$I(\alpha) = -\sum_i p_i \ln p_i,$$

где p_i — доля (относительно длины колонки) i-й буквы, включая знак пробела, в столбце α из МВ. Эта величина отражает просто степень вариабельности кололнки α.

Поскольку в результате бутстреп-анализа исходного МВ имеется список надежных (по некоторому порогу) клад $\beta = \beta_1, \ldots, \beta_j, \ldots, \beta_k$ с показателями бутстрепа (в долях единицы) соответственно равными B_j, то представляется правильным использовать три характеристики,

обсуждаемые ниже с номерами формул (1)-(3). Пусть $k+1$ — номер дополнения всех клад из этого списка.

Рассмотрим какой-то один столбец α из МВ. Входящие в него аминокислоты и еще знак пробела будем называть "буквами". Пусть фиксирован набор β надежных (по некоторому порогу) клад. Все буквы, встречающиеся в столбце, нумеруем числами $i = 1, \ldots, n$. Пусть K_i — число вхождений буквы i в столбец α, и K — длина этого столбца, т.е. общее число строк в исходном МВ. Суммарное число вхождений всех букв (фиксированного столбца α) как строк (белков) исходного МВ во все R реплик, возникших в результате будстреп-анализа, равно KR. Аналогично суммарное число вхождений буквы i как строк исходного МВ во все R реплик равно K_iR.

Вычисляется величина N_{ij}, равная количеству вхождений буквы с номером i в кладу с номером $j = 1, \ldots, k$ в исходном МВ. Учитывая результаты бутстреп-анализа, вводим аналогичные количества вхождений буквы i в какой-то из экземпляров (всего их B_jR) клады j с учетом R реплик бутстрепа, как $N_{ij}B_jR$. Аналогично для характеристики дополнения $N_{i,k+1}$ клад полагаем

$$N_{i,k+1} = K_iR - \sum_j N_{ij}B_jR.$$

Эмпирические частоты встречаемости буквы в кладе и отдельно в дополнении клад вычисляются как отношение числа вхождений буквы i в кладу j с учетом показателей бутстрепа, деленное на общее число вхождений букв (с учетом показателей бутстрепа):

$$\nu_{ij} = \frac{N_{ij}B_j}{K},\ i = \overline{1,n},\ j = \overline{1,k},$$

и

$$\nu_{i,k+1} = \frac{K_i - \sum\limits_j N_{ij}B_j}{K}.$$

Вычисляются частоты встречаемости всех букв P_1, $P_2, \ldots, P_n$ в столбце α (маргинальные частоты) как

$$P_i = \sum_{j=1}^{k+1} \nu_{ij},$$

и аналогичные частоты встречаемости всех клад и их дополнения Q_1, $Q_2, \ldots, Q_{k+1}$ в столбце α как

$$Q_j = \sum_{i=1}^{n} \nu_{ij}.$$

Условную энтропию столбца α относительно клады вычисляем как:

$$H(\alpha|\beta) = -\sum_{i=1}^{n}\sum_{j=1}^{2} \nu_{ij} \ln\left(\frac{\nu_{ij}}{Q_j}\right), \tag{1}$$

где вычисление проводится для каждой клады отдельно, а затем определяется

$$f(\alpha) = \min\{H(\alpha|\beta_j)|\beta_j \in \beta\}.$$

Здесь можно использовать и другие функции вместо f. Упорядочиваем столбцы по убыванию значений функции f, и отбрасываем некоторую начальную по этому упорядочению порцию столбцов.

Порция удаляемых столбцов определяется как состоящая из всех столбцов со значением условной энтропии в интервале от максимального ее значения (по всех столбцам) до 0.1 (величина "шага") от разности этого значения и минимального значения условной энтропии.

До конца этого пункта, основываясь на том же аппарате, рассмотрим *задачу 2*, не связанную прямо с задачей 1, приведенной в начале этого пункта, но относящуюся к той же теме. Пусть дана одна клада, которая считается надежно установленной, или произвольный набор β таких клад. Как найти столбец, который наиболее хорошо согласуется с этими кладами (т.е. сайт, который подтверждает эту филогению β в наибольшей степени)? Тогда буквы этого столбца, которые характерны для этих клад и почти отсутствуют в их общем дополнении, представляют собой филогенетический сигнал молекулярного уровня. Такой столбец ищется как несущий наибольшую информацию об этой филогении β. А именно, столбец находится как тот α, на котором достигают максимума функции $I(\alpha\colon \beta)$ или $\chi^2(\alpha\colon \beta)$, определенные ниже формулами (2)-(3).

Точнее, зависимость между аминокислотным составом столбца α и данным набором уже установленных надежных клад β измеряется с помощью **количества информации** $I(\alpha\colon \beta)$. В случае *попарно непересекающихся клад* β количество информации в столбце α о наборе β вычисляется как

$$I(\alpha\colon \beta) = \sum_{i=1}^{n}\sum_{j=1}^{k+1} \nu_{ij} \ln\left(\frac{\nu_{ij}}{P_i Q_j}\right), \tag{2}$$

т.е. как расстояние Кульбака–Лейблера между эмпирическими совместным распределением и произведением распределений. Здесь считается, что $0\log 0 = 0$.

Эту зависимость можно определить и с помощью аналогичной **характеристики** χ^2 (т.е. как упомянутое расстояние, теперь измеряемое с помощью хи-квадрат):

$$\chi^2(\alpha\colon \beta) = \sum_{i=1}^{n}\sum_{j=1}^{k+1} \frac{(\nu_{ij} - P_i Q_j)^2}{P_i Q_j}, \tag{3}$$

здесь считаем $\frac{0}{0} = 0$.

Формулы (2)-(3) применяются для случаев одной клады и ее дополнения (тогда $j = 1, 2$) и для набора попарно не пересекающихся клад.

Для произвольного набора β клад функции $I(\alpha\colon \beta)$ и $\chi^2(\alpha\colon \beta)$ вычисляются относительно *разбиения*, которое образуется по β следующим образом. Для вложенных друг в друга клад берем их разности, а для самых больших в этом наборе ("внешних") клад берем пересечение их дополнений. Условие применимости метода состоит в том, чтобы каждое множество в этом разбиении было не слишком маленьким, например, содержало 3 и более элементов. Число 3 было получено из аналогии со статистическим критерием [23] без строго доказательства теоремы.

3.3. Корреляционный подход к удалению столбцов из МВ

Другой подход к задаче 1 из раздела 3.2, удаления столбцов из исходного МВ, основан на использовании коэффициента корреляции между двумя любыми столбцами МВ. Для этого каждому столбцу α из МВ сопоставляется *функция* $F(\alpha)$, оценивающая близось относительно α любых двух строк a и b из МВ, т.е. двух белков из исходного белкового семейства, в смысле

близости их расположения в будущем дереве этого белкового семейства. А именно, если у столбца α в строках a и b расположены близкие аминокислоты $MB(a,\alpha)$ и $MB(b,\alpha)$, то считаем $F(\alpha,a,b) = F(\alpha,b,a)$ и равным сходству этих аминокислот по матрице аминокислотных замен (взятой, например, с сайта www.zbh.uni-hamburg.de/research/torda/sub_mat). Если в одной из этих a или b строк находится аминокислота, а в другой пробел, то это значение полагаем равным (-4). Если в обоих строках — пробелы, то это значение полагаем равным 1. Конечно, эти числа условны и могут быть заменены на другие подходящие значения параметров.

Предполагается удалять столбец, который слабо коррелирует с большинством других столбцов из MB. Для этого вводится *коэффициент корреляции* между столбцами, отражающий, насколько два произвольных столбца α и α_1 зависимы друг от друга. Естественно представить себе функцию $F(\alpha,a,b)$ столбца α (с переменными a и b) как некоторые P значений некоей случайной величины (своей для каждого столбца, пусть $P = \frac{K(K-1)}{2}$, где K — длина столбца в MB), и использовать коэффициент корреляции между такими двумя случайными величинами, соответствующими двум любым столбцам α и α_1. Такое значение P возникает как число различных номеров a и b любых неравных между собой строк в MB. Как обычно, этот коэффициент равен математическому ожиданию произведения отклонений случайных величин от их математических ожиданий, делённому на корень квадратный из произведений их дисперсий. Используя статистический аналог коэффициента корреляции, полагаем:

$$K(\alpha,\alpha_1) = \frac{E\Big([F(\alpha,a,b) - E(F(\alpha,a,b))][F(\alpha_1,a,b) - E(F(\alpha_1,a,b))]\Big)}{\sqrt{E\Big([F(\alpha,a,b) - E(F(\alpha,a,b))]^2\Big)E\Big([F(\alpha_1,a,b) - E(F(\alpha_1,a,b))]^2\Big)}},$$

где α и α_1 — два столбца из MB, а E означает среднее арифметическое по переменным a и b. Если знаменатель равен нулю, то всё выражение полагаем равным нулю. Напомним, что "среднее" означает здесь сумму по всем значениям соответствующей функции от ее аргументов a и b (при всех неравных между собой значениях аргументов a и b) деленную на число пар a и b в ее области определения, т.е. на упомянутое выше число P.

Для каждого столбца вводится *показатель его зависимости*, например, как средний коэффициент его корреляции со всеми другими столбцами. И, наконец, столбцы с малыми значениями показателя зависимости *удаляются*.

Перейдем теперь к задаче 2 с этой точки зрения. Пусть дана некоторая заведомо надёжная клада β, и мы хотим определить, насколько столбец α различает эту кладу от ее дополнения. Показателем соответствия столбца α кладе β считаем разность двух величин: первая из них — среднее значение функции $F(\alpha,a,b)$ по a и b — всем различным парам строк, взятым из клады β, а вторая — среднее значение той же функции на парах строк, ровно одна из которых принадлежит β, а вторая не принадлежит β.

В более сложных случаях, когда дано несколько заведомо надёжных клад и они могут быть вложены друг в друга, показателем соответствия столбца α набору β таких клад можно считать коэффициент корреляции между функцией $F(\alpha,a,b)$ и некоторой функцией $G(a,b)$, отражающей степень принадлежности двух любых строк a и b (двух белковых последовательностей) к одной из этих клад, а также степень надёжности клады (процент ее бутстрепа). Сначала опишем общие требования на такую функцию. Для a и b из клады функция $G(a,b)$ должна быть тем больше, чем меньше клада и больше её дополнение, а также, чем больше степень её надёжности. Это позволит сбалансировать влияние на результат размера и надёжности клад, а также отразить значимости двух случаев: первый — обе строки a и b взяты из какой-то одной исходной клады и второй — одна строка из какой-то клады, а другая вне нее (т.е. в другой кладе или вообще вне всех исходных клад). При этом обе функции F и G определены только для таких пар a и b, что обе строки принадлежат какой-то кладе, или одна из

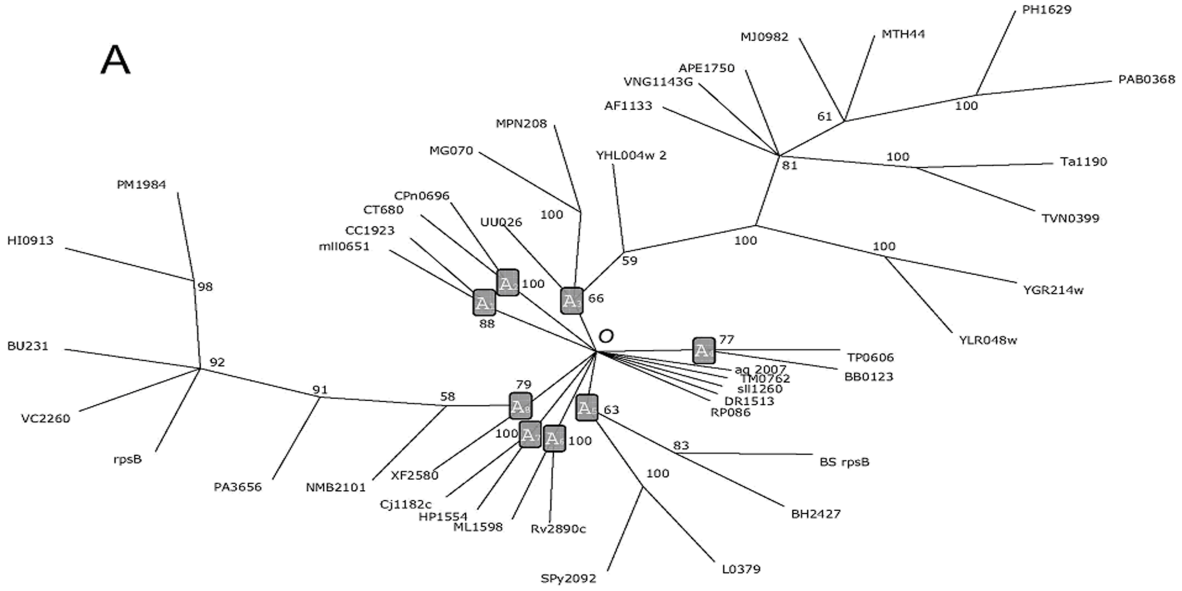
них в кладе, а вторая вне нее (эти пары a и b назовем *граничными*). Например, если исходные клады не пересекаются, то простейший вариант таков: $G(a,b) = \frac{d_j B_j}{k_j}$ для любых пар $\langle a,b\rangle$ из j-ой клады, и $G(a,b) = 0$ для граничных пар, где B_j — доля бутстрепа, k_j — мощность клады, а d_j — мощность её дополнения.

3.4. Прекращение процесса удаления порций столбцов

Для определения *итогового* шага, т.е. шага, на котором полученное текущее МВ считается окончательным, на каждом шаге работы алгоритма вычисляется доля q неразрешенных квартетов [12]. Квартет — это случайный набор из четырех последовательностей, взятых из числа анализируемых. Возможны три варианта представления квартета бескорневым бинарным деревом. Показатель максимального правдоподобия вычисляется для каждого из трех вариантов и нормализуются так, чтобы сумма значений равнялась единице. Таким образом, однозначно определяется точка в правильном трехмерном симплексе. Положение точки в одном из углов симплекса соответствует случаю, когда из трех возможных бескорневых топологий предпочтительна одна, т.е. квартет может быть однозначно разрешен. Анализ симплекса, на которых нанесены данные по большому числу случайных квартетов, информативен с точки зрения возможности вывести из данного выравнивания разрешенное дерево: чем меньше точек содержится в его центральной части (это и есть доля неразрешенных квартетов q), тем лучше разрешение дерева, которое получается на основе такого выравнивания. Детальное описание процедуры максимально-правдоподобного картирования приведено в [12]. Чтобы снизить отрицательный вклад в величину q квартетов, состоящих из видов, расположенных в хорошо разрешенных кладах 50%консенсусного дерева (т.е. направить процедуру картирования на выявление сигнала в областях именно с низким разрешением клад), такие виды исключатся из текущего МВ, что приводит к определению *редуцированного* МВ, по которому в нашем алгоритме вычисляется q. Для этого в 50%-консенсусном дереве для вершин расположенных, например, на единичном расстоянии от узла с наибольшей политомией, реконструируются предковые последовательности (с использованием выбранной модели эволюции), рис. 1А. Концевые виды такого узла (т.е. соответствующие строки текущего МВ) заменяются на одну предковую последовательность, реконструированную в этом узле. Таким образом от текущего МВ переходим к новому МВ с меньшим числом строк, которое называем редуцированным. Это МВ используется только для вычисления q, рис. 1Б.

Алгоритм заканчивает работу, когда показатель q доли неразрешенных квартетов на редуцированном выравнивании, полученном по текущему МВ, достигает первого локального минимума.

Описанный выше алгоритм процедуры улучшения исходного МВ использует критерий, основанный на вычислении доли q неразрешенных квартетов. Нами разработан и *другой критерий*, использующий статистический анализ распределения длин большого числа случайных бинарных деревьев с фиксированным набором концевых вершин, равным числу строк в матрице МВ [24]. Этот критерий основан на генерировании таких деревьев с последовательной (от меньших клад к большим) заменой в 50%консенсусном дереве, построенном по исходному МВ, небинарных поддеревьев на бинарные с последовательным уменьшением числа политомических вершин до нуля. На рис. 1А, например, для первой политомической вершины (с меткой 61) порождаются все случайные деревья с тремя концевыми вершинами, которым приписываются виды MJ0942, MTH44 и поддерево с меткой 100; затем переходят к следующей политомической вершине (с меткой 81), и т.д. Каждое так полученное случайное бинарное дерево заменяет исходное небинарное поддерево, соответствующее политомическому узлу. Для каждого так полученного случайного дерева вычисляется его длина как наименьшее число молекулярных замещений в текущем МВ, необходимое для объяснения топологии этого дерева.



Б

```
                10        20        30        40        50        60        70        80        90       100
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
aq_2007         MKELLEAGVHFGHSKSRWNPKMAPYIYGVR-H---GIHIIDLNKTIVMLEDAYHFVSDVIAQGGEILFVGTKKQ-AKDIIKEEAERCGAFYVNERWVGGL
TM0762          MKQLLEAGVHFGHRTRRWNPKMAPYIYTER-K---GIYIIDLQKTQQLLEEAYYFVREKASEGATILFVGTKKQ-AQGVIKAEAERCGAFYVNNRWLGGL
sll1260         LADLLESGVHFGHQTRRWNPRMDQYIYTAR-N---GVHIIDLVQTAQLMEDAYEYVRSSTINGKKFLFVGTKKQ-AAGIISQEAQRCGAHYVNQRWLGGM
DR1513          MKQLLEAGVHFGHETKRWNPKFKKFIFAER-N---GIFIIDLQKTLKQVDRSFDYIKDLAERGGVILFVGTKKQ-AQEIVELEARRTGMPYVTSRWLGGM
RP086           IKELLDAGVHFGHKTSRWNPKMASYIYGER-D---DVHIIDLRQSAALMSVALNAIYETVKKDGKILFVSTKIQ-ASDIIAEYAEKCGQYYVNNRWLGGM
A1              MKNLLEAGVHFGHQVRRWDPRMKRYIFAERANHNNGIHIIDLQKTLQAIKEAYDLVRRTVSDGKSVLFVGTKKQYAQEIIEQEAQRCGMFYVNNRWLGGM
A2              MKQLLEAGVHFGHQTRRWNPKMKPYIFTERANHNNGIHIIDLQKTVKLVEEAYNFVRDLAAEGGKILFVGTKKQYAQDSVKEEAERCGMYYVNHRWLGGM
A3              MKQLLDSGTHFGHQTRRWNPKMKRFIFTDRANHNNGIYIIDLQQTLTFIDKAYEFVKETVAHGGSVLFVGTKKQYAQESVAAEATRVGMPYVNQRWLGGM
A4              MKDLLECGVHFGHQTRRWNPKMKKFIFGERAKHNNGIHIIDLQKTLRYFRYTYNIVRDAAAEGKTILFVGTKKQYANETIKEYAEKCGMPYVNHRWLGGM
A5              MRQMLEAGVHFGHQTRYWNPKMKPYIFGARANHNNKIHIINLEKTLPLFNDAMNFVRRVAAKKGTILFVGTKRSYARDVIKEEAERCGMPYVNHRWLGGM
A6              MRQLLEAGVHFGHQTHRWNPKMAPYIFGARANHNNNIHIIDLSQTVPLLHQALKQVRDTAAKGGRVLFVGTKRQYASDIVADAAQRCAQYYVNHRWLGGM
A7              LKELLEAGAHFGHQTRRWNPKMKPYIFEEKANHNNGLYIIDLAKTLQQLKKALAHIRKTIGDNKTILFVGTKKQYAKQIIREAAIECGEEFVSERWLGGM
A8              IKKLLEAGVHIGMQTRRWNPKMKPYIYGERPNHNNGIHVIDLQKTLQFLEKAYNFVRDVAQNGGTILFVGTRKQYVKELIKEEAERTNAYYVTQRWLGGT

                110       120       130       140       150       160       170       180       190       200
                ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
aq_2007         LTNFQTVRKSIMKLQTLERMEEDG-IFDVLPKKEVRNLRRKMEKLRKLYNGIRDMKKLPDVIWIVDTVREKIAVSEARKLGIPIVAIADSNCDPTVIDYT
TM0762          LTNFKTIRSRIDKLIELEEMEQSG-KLEELPKKEQSRIRRVLEKLRKNLGGLKEMRRLPDIIYIVDPRKEKIAVAEANKLGIPIVAIVDTNCDPDPIDYV
sll1260         LTNWETIRKRVDRLKELEALEASG-GIDRRGKKEGSMLRRELGKLQKYLGGIKNMRKLPDVVVIVDQRREHNAIHECQKLGIPIIAMLDTNCDPDVVDVP
DR1513          LTNFKTMRTRIDRLNELDDLFESE-RINDRPKAERIQLASERERLLREVGGIRKMNRLPDAIFVVDPTKEVIAVQEANKLGIPVIALADTDSDPDVIDYI
RP086           LTNWKTIACSIEKLEKLEKTLESEEAARICYTKKEILDMSRKKDKLLLSLAGIRNLNSKPDLLVVIDTNKEHIAINEAVKLNIPVVAVVDTNSNPDNVDYP
A1              LTNFNTIRKSIQRLKKLEKMEVDGETFDHLSKKEVAQLRRERAKLEKNLGGIKDMKKLPGVIFIIDPRKEQIAIREAKKLGIPIVAVVDTNCNPDVIDYP
A2              LTNFETIQKRIKRLKELEKMEEDGETFDVLPKKEVVQLKKERERLEKFLGGIKDMKRLPDVLFVIDPRKERIAVAEAKKLGIPVVAIVDTNCDPDEIDYV
A3              LTNFSTVHKRLQRLKELEAMEQTGEGFEGRTKKEILGLTREKNKLERSLGGIRDMAKVPSAIWVVDTNKEHIAVGEAKKLGIPVIAILDTNCDPDEVDYP
A4              LTNFSTIRKSIRKLEVIEKMEESGESIDLLTKKEALMLTRKKEKLDKYLGGIRHMKKLPDMIFVIDTAKEKIAVAEAKKLRIPVVAPLDTNCDPDLVDYP
A5              LTNFKTVRQSIKRLKELETAEEDGETFDKLTKKEVLQLRRELEKLERSLGGIKDMKRLPDAIFVIDTGHEKIAIKEAKKLGIPVVAVVDTNSDPDGVDYP
A6              LTNWKTISNSIQRLRELEEMELAGEEAQGLTKKELLQLTRERERLEKLKSLGGIKDMGSIPDIMFVIDTNKEAIAILEAKKLGIPVVAIIDTNCDPDKIDYP
A7              LTNMTTIRNSIKTLNKIEKDLESAESNSGLTKKELALLAKRHQKLLRNLEGIRHMKKLPGLLIVVDPSYERIAVAEAKKLGIPVMALVDTNCDPTPIDHV
A8              LTNFKTISKSIKKLNELEKMQESEETFDKLTKKEKLLLRRETEKLEKFFGGIKNMKRLPDVLIVVDPVNDKNAVAEARKLGIPVVALCDTNSDPELIDYP

                210       220
                ....|....|....|...
aq_2007         IPGNDDAIKSIKLLTSKIADAVLEGKQR
TM0762          IPGNDDAIRSIKLITSVIADAYLEGREG
sll1260         IPANDDAIRSIKLIVGKLADAIYEGRHG
DR1513          VPGNDDAIRSIQLITHRVGDLLVEARGG
RP086           IPGNDDSIRSIRLYCSLFADAALQGLEE
A1              IPGNDDAIRSISLFTQIIADAVLEAGKE
A2              IPGNDDAIRAVKLLTSKMADAILEGRQG
A3              IPGNDDAIRSAALLTRVIASAVAEGLQA
A4              IPGNDDAIRSIQLFCKEMAEAILEGREL
A5              IPGNDDAIRAVQLYARAIADAVLEGKQG
A6              IPGNDDAARAIQLYCDLIADAVLDGLAA
A7              IPCNDDSMKSIRLIVNVIKDAVIEAKHR
A8              IPGNNDSVRSITLLLGVLADAVAEGRGG
```

Рис. 1. Редуцированное выравнивание, образованное по 50%-консенсусному дереву КОГа 52 (рибосомальный белок S2). В политомический узел O входят 5 концевых вершин (соответствующих aq2007, TM0762, sll1260, DR1513, RP086) и 8 внутренних вершин (помеченных индексами $A_1 \ldots A_8$), рис. 1А. В состав редуцированного МВ, показанного на рис. 1Б, вошли последовательности из этого КОГа, приписанные концевым вершинам, и предковые последовательности, реконструированные в вершинах $A_1 \ldots A_8$ (последние выделены рамкой). Использованы следующие сокращения: aq_2007 — *Aquifex aeolicus*, TM0762 — *Thermotoga maritime*, sll1260 — *Synechocystis sp.*, DR1513 — *Deinococcus radiodurans*, RP086 — *Rickettsia prowazekii*

Для этих чисел определяется доля встречаемости каждого из них и смещение влево кривой из этих долей. Это смещение измеряется статистикой $g_1 = \frac{\sum_{i=1}^{n}(T_i-\overline{T})^3}{ns^3}$, где n — это число всех деревьев, T_i — длина i-го дерева, $\overline{T}$ — среднее значение этих чисел и s –стандартное отклонение длин от $\overline{T}$, вычисляемое по обычной формуле $s = \frac{\sqrt{\sum_{i=1}^{n}(T_i-\overline{T})^2}}{n-1}$, [25]. Если текущее МВ достаточно информативно для установления родства составляющих его строк, то смещение g_1 кривой уменьшается, и это говорит о присутствии в выравнивании филогенетического сигнала. Действительно, в этом случае доля деревьев, для объяснения которых требуется меньшее число замещений, будет больше, чем при анализе набора случайных строк (максимально шумное выравнивание). Алгоритм заканчивает удаление порций столбцов на шаге, на котором величина g_1 достигает первого локального минимума при условии, что на всех итерациях g_1 меньше некоторого порога достоверности. Подробное изложение этого второго критерия и результаты счета с его использованием будут даны в другой публикации.

4. РЕЗУЛЬТАТЫ КОНТРОЛЬНОГО СЧЕТА

Указанная в разделах 2-3 цепочка алгоритмов тестировалась на наборе из 50-ти белковых семейств, взятых из базы данных, содержащей кластеры ортологичных групп белков (КОГи из базы COGs, NCBI). В набор были включены КОГи, содержащие примерно одинаковое число (45-50) последовательностей длиной до 500 аминокислотных остатков. Отбор КОГов был продиктован эвристическими соображениями: размер их выравниваний (длина строки и число строк) не должен быть слишком большим, так как иначе **проверка** (но не сама работа) алгоритма, состоящая в построении дерева и в вычислении показателя правдоподобия (на каждом шаге алгоритма) займет слишком много времени; выравнивание должно содержать консервативные (предположительно несущие информацию, филогенетический сигнал) позиции и вариабельные участки (предположительно несущие шум) для проверки содержательной стороны работы алгоритма. Заметим, что скорость работы нашего алгоритма полиномиальная и оценивается полином третьей степени.

В исходном МВ каждого КОГа шумные позиции удалялись за несколько последовательных шагов на основании значений функций энтропии или корреляционной функции, описанных выше. Текущее МВ на каждом шаге удаления шума с **целью контроля** работы алгоритма использовалось для построения филогенетического дерева G_i программой MrBayes v.3.0 с одинаковыми параметрами поиска. Для полученных G_i вычислялся показатель максимального правдоподобия ($-\ln L$) всякий раз *относительно исходного* МВ соответствующего КОГа. При сравнении показателей $-\ln L$ деревьев, полученных на последовательных итерациях алгоритма, для всех исследованных КОГов была выявлена следующая закономерность: величина $-\ln L$ достигает локального минимума (т.е. величина L максимального правдоподобия дерева достигает локального максимума) после удаления некоторого числа колонок, и затем начинает расти на дальнейших шагах работы алгоритма. Это наблюдение позволяет предположить, что удаленные столбцы в действительности мало информативны и, более того, препятствуют корректной реконструкции филогенетического дерева. Локальный минимум $-\ln L$ во всех исследованных случаях не приходится на дерево, построенное по исходному МВ, что свидетельствует: некоторые узлы дерева и/или длинны некоторых ветвей реконструируются с большей достоверностью после удаления шумных позиций. Для 24 из 50-ти тестированных КОГов значение локального минимума $-\ln L$ оказалось отличным от величины $-\ln L$ для дерева, построенного по исходному МВ, **статистически достоверно** по общепринятому тесту ELW. Известно, что тест ELW менее консервативен и в меньшей степени зависит от длины МВ и числа последовательностей в МВ при определении границ доверительного интервала показателя

$-\ln L$, [13]. Сравнительный анализ показывает, что эти 24 КОГа представлены наименее консервативными белковыми семействами, МВ которых содержат высоко вариабельные участки с трудно установимой позиционной гомологией аминокислотных остатков (примеры таких КОГов — это КОГи 0016, 0018 и 0020 в Таблице). Метод, использованный нами для реконструкции дерева, считается наиболее устойчивым к присутствию шума, поэтому представляется, что при использовании более чувствительных к шуму методов, например, дистанционных, применение нашего алгоритма тем более значительно повысит аккуратность филогенетического анализа. Отметим, что достоверное улучшение $-\ln L$ было зарегистрировано при условии вычисления этого показателя для деревьев относительно исходного выравнивания, в котором заведомо присутствует шум. При вычислении величины $-\ln L$ относительно выравнивания, на котором достигнут локальный минимум числа неразрешенных квартетов q, статистически достоверное улучшение показателя $-\ln L$ наблюдается в 100% случаев.

Результаты этого контрольного счета показали, что в 96% случаев (для 48-ти КОГов из 50-ти) первый локальный минимум $-\ln L$ совпадает с первым локальным минимумом величины q, вычисленной для соответствующего МВ (соответствующие шаги алгоритма выделены полужирным шрифтом в примерах из Таблицы). Таким образом, показатель q может использоваться в качестве критерия остановки алгоритма. Поскольку величина q вычисляется сравнительно быстро, время работы нашего алгоритма полиномиальное, не выше 3й степени, в то время как время минимизации показателя $-\ln L$ при построении дерева более, чем экспоненциальное.

Использование описанных функций при выявлении шумных колонок дали практически одинаковые результаты. Результаты, приведенные в Таблице для пяти КОГов получены с использованием функции условной энтропии.

Таблица. Шаги алгоритма по удалению шумных колонок в исходном множественном выравнивании для пяти КОГов.

	COG 0012			COG 0016			COG 0018			COG 0020			COG 0052		
i	n	$-\ln L$	q	n	$-\ln L$	q	n	$-\ln L$	q	n	$-\ln L$	q	n	$-\ln L$	q
1	0	19876,58*	11,13	0	19499,28	9,4	0	37059,25	3,9	0	13141,56	17,8	0	6939,79*	7,6
2	13(3,57)	19889,57	11,12	5(1,43)	19498,51	9,2	19(3,44)	37060,34	3,9	3(1,21)	13141,68	17,3	3(1,32)	6939,24*	7,4
3	**43(11,81)**	**19866,88***	**10,9**	13(4,58)	19492,43	9,2	47(8,50)	37056,30*	3,8	10(4,05)	13137,12	16,7	10(4,39)	6746,57	7,7
4	85(23,35)	19879,65*	12,8	31(8,88)	19489,18	9,0	**70(12,66)**	**37040,83***	**3,5**	13(5,26)	13120,01*	16,4	**25(10,96)**	**6734,04***	**7,0**
5	129(35,45)	19895,27	15,4	48(13,75)	19488,98	8,9	100(18,08)	37053,71*	3,7	**15(6,07)**	**13110,88***	**16,1**	39(17,11)	6739,14*	7,4
6				**72(20,63)**	**19456,13***	**8,5**	136(24,59)	37051,52*	3,8	23(9,31)	13123,03*	17,1	50(21,93)	6753,53	8,4
7				94(26,93)	19476,13*	9,0				36(14,57)	13147,59	18,9			
8				105(30,09)	19532,13	12,1									

Нулевой столбец указывает номер i шага в процедуре удаления столбцов из исходного МВ. Столбец n указывает число удаленных колонок и их относительную долю от числа всех колонок в исходном МВ в процентах; $-\ln L$ — показатель правдоподобия дерева G_i; q — доля неразрешенных квартетов в текущем МВ, определенная согласно нашей модификации метода квартетного картирования. Наименьшее значение величины q выделено полужирным шрифтом. Оно определяет номер шага, на котором полученное текущее МВ считается оптимальным. Локальный минимум показателя $-\ln L$ деревьев G_i, полученных для данного КОГа на этапах работы алгоритма, также выделен полужирным шрифтом. Из Таблицы видно, что локальные минимумы величин q и $-\ln L$ приходятся на одну итерацию. Звездочка отмечает значения показателя $-\ln L$, несущественно отличающиеся между собой согласно наименее консервативному статистическому тесту ELW.

СПИСОК ЛИТЕРАТУРЫ

1. Halanych K.M., Robinson T.J. Multiple substitutions affect the phylogenetic utility of cytochrome b and 12S rDNA data: examining a rapid radiation in leporid (Lagomorpha) evolution. *J. Mol. Evol.*, 1999, vol. 48, pp. 369–379.
2. Yang Z., Roberts D. On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, 1995, vol. 12, pp. 451–458.
3. Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. In *Molecular systematics*. 2nd edition. Eds. Hillis D.M., Moritz C., Mable B.K., Sunderland, Massachusetts: Sinauer, 1996, pp. 407–514.
4. Tatusov R.L., Koonin E.V., Lipman D.J. A genomic perspective on protein families. *Science*, 1997, vol. 278, pp. 631–637.
5. Do C.B., Brudno M., Batzoglou S. Prob Cons: probabilistic consistency-based multiple alignment of amino acid sequences. Proceedings of the Nineteenth National Conference on Artificial Intelligence, July 25-29, San Jose, California, p. 703, 2004.
6. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control.*, 1974, vol. 19, pp. 716–723.
7. Schwarz G. Estimating the dimension of a model. *Ann. Stat.*, 1978, vol. 6, pp. 461–464.
8. Keane T.M., Naughton T.J., McInerney J.O. *ModelGenerator: amino acid and nucleotide substitution model selection.* National University of Ireland, Maynooth, Ireland, 2004 (http://bioinf.nuim.ie/software/modelgenerator).
9. Felsenstein J. *PHYLIP (Phylogeny Inference Package) version 3.6.* Distributed by the author, Department of Genome Sciences, University of Washington, Seattle, 2004.
10. Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 2002, vol. 18, pp. 502–504.
11. Roger A., Holder M. *PUZZLEBOOT version 1.03.* Department of Biochemistry & Molecular Biology, Dalhousie University, Canada, 2003 (http://hades.biochem.dal.ca/Rogerlab/Software/software.html).
12. Strimmer K., von Haeseler A. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *P. Natl. Acad. Sci. Usa.*, 1997, vol. 94, pp. 6815–6819.
13. Strimmer K., Rambaut A. Inferring confidence sets of possibly misspecified gene trees. *P. Roy. Soc. Lond. B.*, 2002, vol. 269, pp. 137–142.
14. Kishino H., Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.*, 1989, vol. 29, pp. 170–179.
15. Shimodaira H., Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.*, 1999, vol. 16, pp. 1114–1116.
16. Goldman N. Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 1993, vol. 36, pp. 182–198.
17. Yang Z. PAML: a program for phylogenetic analysis by maximum likelihood version. *CABIOS*, 1997, vol. 13, pp. 555–556.
18. Notredame C., Abergel C. In *Bioinformatics and Genomes: Current Perspectives.* Ed. Andare M. Wymondham, UK: Horizon Scientific Press, 2003, pp. 30–55.
19. O'Sullivan O., Suhre K., Abergel C., Higgins D.G., Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.*, 2004, vol. 340, pp. 385–395.
20. Huelsenbeck J. P., Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 2001, vol. 17, pp. 754–755.
21. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.*, 1978, vol. 27, pp. 401–410.

22. Hendy M.D., Penny D. A framework for the quantitative study of evolutionary trees. *Syst. Zool.*, 1989, vol. 38, pp. 297–309.

23. Вьюгин В.В., Маслов В.П. Об экстремальных соотношениях для колмогоровской сложности и аддитивных функций потерь. *Проблемы передачи информации*, 2003, том. 39, № 4, стр. 71–87

24. Hillis D.M., Huelsenbeck J.P. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.*, 1992, vol. 83, pp. 189–195.

25. Sokal R.R., Rohlf F.J. *Biometry.* 2nd ed., San Francisco: W.H. Freeman, 1981.