

УДК 575.852'112

СРАВНЕНИЕ МЕТОДОВ РЕКОНСТРУКЦИИ ФИЛОГЕНИИ БЕЛКОВ НА МАТЕРИАЛЕ ПРИРОДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

М.С. Кривоzubов, С.А. Спирин

(факультет биоинженерии и биоинформатики и НИИ ФХБ имени А.Н. Белозерского МГУ;
e-mail: mikhailkrivozubov@yandex.ru)

На материале природных последовательностей проведено сравнение нескольких методов реконструкции филогении белков. Были протестированы программы пакета PHYLIP, а также программы FastME, PhyML и TreeTop. Вопреки исследованиям на симулированных данных наши результаты демонстрируют превосходство программ, принимающих на вход матрицу расстояний, над реализациями метода наибольшего правдоподобия.

Ключевые слова: реконструкция филогении, алгоритмы, последовательности белков.

К настоящему времени разработано множество методов реконструкции филогении белков, представляющих результаты в виде филогенетических деревьев. Все основные их типы реализованы в пакете PHYLIP [1]. Метод максимальной экономии реализован в пакете PHYLIP в программе *protpars*, метод наибольшего правдоподобия — в программах *proml* и *promlk*, метод Фитча—Марголиаша [2] и родственные методы (например, метод наименьших квадратов [3]) реализованы программами *fitch* и *kitsch*, методы объединения соседей (Neighbor-joining, [4]) и UPGMA [5] реализованы как опции программы *neighbor*. Для вычисления расстояний между выравненными белковыми последовательностями в пакете PHYLIP имеется программа *protldist*. Методы UPGMA, *kitsch* и *promlk* предполагают справедливость гипотезы молекулярных часов, остальные методы не предполагают ее.

Разные алгоритмы часто выдают деревья отличающейся топологии. Тем не менее все известные нам сравнения их качества [6, 7] проводились на искусственно генерируемых последовательностях и могут неточно характеризовать успешность их работы с настоящими белками. Мы провели сравнение методов реконструкции филогении белков путем сравнения деревьев, построенных каждым методом по ортологичным белковым семействам фиксированного набора организмов, с деревом, построенным по объединенному выравниванию всех выбранных белков.

Филогения конкретного семейства белков может отличаться от филогении организмов из-за горизонтальных переносов, потери паралогов или других причин. Однако если какая-либо ошибка отдаляет реконструкцию филогении конкретного семейства от его реальной филогении, то эта ошибка скорее отдалит реконструкцию и от филогении организмов — случайное сближение с последней маловероятно. Поэтому, по нашему мнению, филогения ор-

ганизмов, реконструированная по множеству белков, может быть использована как эталон для оценки качества реконструкций филогении отдельных белков.

Помимо программ пакета PHYLIP, были протестированы программы PhyML [8] (реализация метода наибольшего правдоподобия), FastME [9] (реализация одного из вариантов метода минимальной эволюции) и TreeTop [10, 11] (реализации топологического и кластерного методов).

Методы

Оценку эффективности методов проводили по следующей схеме:

- составляли выборку организмов, белковые домены которых использовали для построения деревьев;
- составляли выборку ортологичных рядов белковых доменов этих организмов, строили выравнивания отобранных последовательностей;
- строили эталон филогении отобранных организмов (дерево) по конкатенированному выравниванию;
- каждым из рассматриваемых методов строили дерево для каждого ряда белковых доменов;
- качество методов оценивалось путем сравнения выданных ими деревьев с эталоном.

В данной работе были использованы четыре выборки организмов. Первая (1) представлена 30 видами эубактерий, вторая (2) — 42 видами архей. Третья (3) и четвертая (4) выборки получены в результате разбиения пополам выборки 2 и включают по 21 виду. Были отобраны эволюционные домены белков, аннотированные в базе данных Pfam [12] и однозначно группирующиеся в ортологические ряды: для первой выборки организмов — 103 ортологических ряда, для второй — 88 рядов, тот же набор из 88 рядов использовали при работе с выборками 3 и 4.

Эталонное дерево для каждой выборки организмов было получено в несколько этапов по объединенным выравниваниям рядов как строгий консенсус результатов четырех программ: *fitch*, *neighbor*, *proml* и *protpars* (полное изложение процедуры см. на сайте <http://mouse.belozesky.msu.ru/~sas/phylogeny>). Строгий консенсус оказался почти полностью разрешенным: эталонное дерево выборки 1 содержит 24 ветви из 27 возможных, выборки 2 — 26 из 39, выборок 3 и 4 — соответственно 13 и 14 из 18.

Результаты и обсуждение

В таблице приведены результаты тестирования на выборке 1 следующих методов: 1) семь методов пакета PHYLIP (шесть, выполняемых программами по умолчанию, и UPGMA); методам, работающим с матрицей расстояний, подавался на вход результат программы *protdist*; 2) методы, реализованные в программах FastME, PhyML (при двух наборах параметров) и TreeTop (оба метода); 3) программа *fitch*, которой на вход давались расстояния, посчитанные по формуле Кимуры [13] и как доля разли-

чий между последовательностями. Все программы тестировали при значениях параметров, принятых по умолчанию; программа PhyML также при значениях параметров: девять категорий сайтов, автоматический подбор параметров распределения скоростей мутаций по категориям. Программа TreeTop доступна только через веб-интерфейс и была протестирована без учета времени работы.

Основные отличия результатов, полученных на выборках 2, 3 и 4, следующие.

1. Результаты UPGMA и *kitsch* на выборках 2 и, особенно, 4 (для *kitsch* также на выборке 3) оказались существенно хуже, чем на выборке 1. Программа *promlk* наоборот, показала на выборках 3 и 4 заметно лучший результат (превзошла *proml*). Причиной столь разной эффективности этих методов на разных выборках может быть разница в топологии деревьев.

2. На выборке 2 программа *protpars* оказалась худшей среди протестированных методов пакета PHYLIP (в частности, существенно уступила *proml*). Возможно, качество работы *protpars* ухудшается с ростом числа последовательностей во входном выравнивании.

Данные о работе программ пакета PHYLIP, FastME, PhyML и TreeTop с выборкой 1 (30 видов зубактерий, 103 ортологических ряда белковых доменов)

	Среднее	Медиана	> 18	< 12	≥ max - 1	< max - 2	X - fitch			X - proml			Корреляция с длиной	Время
							> 1	< -1	P	> 1	< -1	P		
fitch	17,7	18	42	1	56	21	0	0	—	66	3	$2 \cdot 10^{-16}$	0,33	3,2
fitch Kim.	16,9	17	40	4	41	36	6	18	0,02	60	7	$1 \cdot 10^{-11}$	0,15	1,93
fitch diff.	18	18	47	2	65	13	15	8	0,21	74	4	$1 \cdot 10^{-17}$	0,33	2,12
NJ	17,3	18	38	5	45	31	9	19	0,09	69	5	$2 \cdot 10^{-15}$	0,37	1,9
proml	14,7	15	18	23	16	75	3	66	$2 \cdot 10^{-16}$	0	0	—	0,42	15,7
protpars	15,1	15	13	13	6	78	2	69	$2 \cdot 10^{-18}$	32	21	0,17	0,35	1
FastME	17,4	18	36	2	47	32	16	20	0,62	67	1	$5 \cdot 10^{-19}$	0,36	1,8
PhyML D	16,1	16	21	10	20	62	6	50	$1 \cdot 10^{-9}$	42	7	$4 \cdot 10^{-7}$	0,42	6,7
PhyML M	16,1	16	22	7	18	70	5	51	$1 \cdot 10^{-10}$	50	8	$2 \cdot 10^{-8}$	0,41	61,6
TreeTop T	17,5	18	43	3	51	31	16	22	0,42	69	4	$2 \cdot 10^{-16}$	0,29	—
UPGMA	18	18	49	3	63	13	24	16	0,27	73	7	$6 \cdot 10^{-15}$	0,3	1,9
kitsch	17,7	18	43	4	55	23	17	21	0,63	68	7	$1 \cdot 10^{-13}$	0,27	2,1
promlk	13,9	14	2	19	6	92	3	87	$2 \cdot 10^{-22}$	27	41	0,11	0,31	104,2
TreeTop C	17,4	18	37	4	49	30	21	24	0,77	67	13	$6 \cdot 10^{-10}$	0,25	—

Примечание. Под жирной чертой помещены программы, предполагающие молекулярные часы. Сокращение “fitch Kim.” означает *fitch* с использованием формулы Кимуры, “fitch diff.” — с использованием доли различий, “PhyML D” — PhyML при параметрах по умолчанию, “PhyML M” — PhyML при значениях параметров, указанных в тексте, “TreeTop T” — топологический метод TreeTop, “TreeTop C” — кластерный метод TreeTop. Показателем качества результата программы на одном ортологическом ряде считается число ветвей, общих с эталонным деревом. В столбцах “Среднее” и “Медиана” приведены соответствующие характеристики показателей; в столбце “> 18” — число показателей, больших 18 (3/4 от максимально возможного), в “< 12” — меньших 12 (1/2 от максимально возможного); в столбце “≥ max - 1” — число рядов, для которых показатель отличался не более чем на 1 от лучшего по всем программам, в “< max - 2” — отличался от него более чем на 2. Три столбца “X - fitch” содержат результат сравнения данного метода и *fitch*: в столбце “> 1” указано число рядов, для которых показатели данной программы превосходили *fitch* более чем на 1, в столбце “< -1” — уступали более чем на 1, в столбце “P” — вероятность получить такое же или большее соотношение при случайном равновероятном выборе; в трех столбцах “X - proml” — соответствующие данные для *proml*; в столбце “Корреляция с длиной” — коэффициент корреляции результата с длиной выравнивания; в столбце “Время” — условное усредненное время работы (за единицу принято время работы *protpars*).

Выводы

1. Лучшим методом реконструкции филогении белков среди методов, не предполагающих молекулярные часы, является метод Фитча—Марголиаша, при этом методы объединения соседей, FastME и топологический метод программы TreeTop уступают ему мало. Метод наибольшего правдоподобия в обеих реализациях (*proml* и *PhyML*) показывает весьма низкие результаты по сравнению с перечисленными выше методами.

2. В методе наибольшего правдоподобия учет разных вероятностей мутаций для разных групп сайтов практически не улучшает результат.

3. Лучшим методом среди использующих молекулярные часы является UPGMA.

4. Программа *fitch*, которой на вход поданы расстояния, рассчитанные как доля различий между последовательностями, дает достоверно лучшие результаты, чем она же с расстояниями, рассчитанные по формуле Кимуры.

* * *

Работа проведена при поддержке Российского фонда фундаментальных исследований (грант № 90-04-92743-ННИОМ_а).

СПИСОК ЛИТЕРАТУРЫ

1. Сайт PHYLIP (URL: <http://evolution.genetics.washington.edu/phylip.html> 05.04.2010).
2. *Fitch W.M., Margoliash E.* Construction of phylogenetic trees // *Science*. 1967. Vol. 155. С. 406—416.
3. *Cavalli-Sforza L.L., Edwards A.W.F.* Phylogenetic analysis: models and estimation procedures // *Evolution*. 1967. Vol. 32. С. 550—570.
4. *Saitou N., Nei M.* The neighbor-joining method: a new methods for reconstructing phylogenetic trees // *Mol. Biol. Evol.* 1987. Vol. 4. С. 406—425.
5. *Sokal R.R., Sneath P.H.A.* Numerical Taxonomy. San Francisco: Freeman, 1963. 399 с.
6. *Hall G.B.* Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences // *Mol. Biol. Evol.* 2005. Vol. 22. С. 792—802.
7. *Kuhner K.M., Felsenstein J.* A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates // *Mol. Biol. Evol.* 1994. Vol. 11. С. 459—468.

8. *Guindon S., Gascuel O.* A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood // *Syst. Biol.* 2003. Vol. 52. С. 696—704.

9. *Desper R., Gascuel O.* Fast and accurate phylogeny reconstruction algorithms based on minimum-evolution principle // *J. of Computation Biology*. 2002. N 9. С. 687—705.

10. *Юшманов С.В., Чумаков К.М.* Алгоритмы конструирования филогенетических деревьев максимального топологического подобия // Молекулярная генетика, микробиология и вирусология. 1998. № 3. С. 9—15.

11. Веб-интерфейс программы TreeTop: (URL: http://www.genebee.msu.ru/services/phtree_reduced.html 05.04.2010).

12. *Finn R.D. et al.* The Pfam protein families database // *Nucleic Acids Research*. 2010. Database Issue 38. С. D211—222.

13. *Kimura M.* The neutral theory of molecular evolution. Cambridge: Cambridge University Press, 1983. 388 с.

Поступила в редакцию
13.04.10

COMPARISON OF PROTEIN PHYLOGENY RECONSTRUCTION METHODS ON NATURAL PROTEIN SEQUENCES

M.S. Krivozubov, S.A. Spirin

On a set of real protein sequences, comparison of several protein phylogeny reconstruction methods has been performed. The programs of PHYLIP package and programs FastME, PhyML and TreeTop have been tested. In contrast to several studies that used simulated sequences, our results demonstrate superiority of distance methods over the maximum likelihood method.

Key words: *phylogeny reconstruction, algorithms, protein sequences.*

Сведения об авторах

Кривоzubов Михаил Сергеевич — студент факультета биоинженерии и биоинформатики МГУ. Тел. (965)301-12-05; e-mail: mikhailkrivozubov@yandex.ru

Спирин Сергей Александрович — канд. физ.-мат. наук, ст. науч. сотр. НИИФХБ имени А.Н. Белозерского МГУ. Тел. (495)939-54-14; e-mail: sas@belozersky.msu.ru