БИОХИМИЯ, 2022, том 87, вып. 12, с. 2078 – 2088

УДК 577.29

НАСКОЛЬКО ЧАСТО ФИЛЬТРАЦИЯ КОЛОНОК ВЫРАВНИВАНИЯ УЛУЧШАЕТ РЕКОНСТРУКЦИЮ ФИЛОГЕНИИ ДВУХДОМЕННЫХ БЕЛКОВ?

© 2022 А.И. Сигорских[1], Д.Д. Латорцева[1], А.С. Карягина[2,3,4], С.А. Спирин[3,5]*

[1] *Московский государственный университет имени М.В. Ломоносова, факультет биоинженерии и биоинформатики, 119992 Москва, Россия*

[2] *Национальный исследовательский центр эпидемиологии и микробиологии им. Н.Ф. Гамалеи Минздрава России, 123098 Москва, Россия*

[3] *НИИ физико-химической биологии им. А.Н. Белозерского, Московский государственный университет имени М.В. Ломоносова, 119992 Москва, Россия; электронная почта: sas@belozersky.msu.ru*

[4] *Всероссийский научно-исследовательский институт сельскохозяйственной биотехнологии, 127550 Москва, Россия*

[5] *НИУ «Высшая школа экономики», 109028 Москва, Россия*

Поступила в редакцию 23.09.2022
После доработки 01.11.2022
Принята к публикации 01.11.2022

Реконструкция филогении белков обычно проводится по множественному выравниванию их аминокислотных последовательностей. Одной из проблем является наличие в таких выравниваниях участков различной консервативности, в том числе таких, где качество выравнивания сомнительно. Для решения этой проблемы часто применяется фильтрация колонок выравнивания, для чего разработано специальное программное обеспечение. В данной работе исследованы различные подходы к реконструкции филогении на примере белков с двумя эволюционными доменами. Последовательности таких белков заведомо неоднородны по консервативности благодаря наличию как эволюционных доменов, так и линкеров между доменами, а также *N*- и *C*-концевых частей. Показано, что фильтрация колонок выравнивания в среднем улучшает качество реконструкции только при использовании полноразмерных последовательностей и только при работе с эукариотическими белками. Показано также, что ограничение выравнивания на эволюционные домены с отбрасыванием менее консервативных линкеров и концевых последовательностей в среднем ухудшает качество филогенетической реконструкции.

КЛЮЧЕВЫЕ СЛОВА: филогенетическая реконструкция, эволюционные домены, фильтрация множественного выравнивания.

DOI: 10.31857/S0320972522120223, **EDN:** NIYEQG

ВВЕДЕНИЕ

Практически любое современное исследование в области молекулярной биологии и генетики включает в себя молекулярно-филогенетический анализ, основанный на множественном выравнивании последовательностей нуклеиновых кислот или белков. Распространённым методом подготовки множественного выравнивания белков к филогенетическому анализу является фильтрация (или «тримминг») колонок выравнивания. Это делается с целью удалить колонки выравнивания, в которых «шум» превалирует над «филогенетическим сигналом». Шум может возникнуть, например, в очень быстро эволюционирующих регионах, поскольку в них велика вероятность возвратных мутаций и, как следствие, гомоплазии. Кроме того, качество выравнивания (то есть соответствие распределения остатков по колонкам реальной эволюционной истории) на таких участках заметно ниже.

Для фильтрации выравниваний разработано многочисленное программное обеспечение, отбирающее «информативные» колонки

* Адресат для корреспонденции.

по таким признакам, как принадлежность к блоку без участков идущих подряд неконсервативных позиций (программа Gblocks [1]), процент гэпов ниже определённого порога (trimAl [2]) или предсказанный на основе математической модели уровень гомоплазии (NOISY [3]). Большинство опубликованных протоколов реконструкции филогении белков включают фильтрацию в качестве необходимого этапа, см., например, статью Jermiin et al. [4]. В частности, фильтрация входит по умолчанию в популярные программные конвейеры phylogeny.fr [5] и NGPhylogeny.fr [6].

В статье Tan et al. [7] было проведено тестирование ряда популярных программ фильтрации и показано, что они в среднем не улучшают качество реконструкции, а большинство их (например, довольно популярная программа Gblocks [1]), наоборот, заметно его ухудшают. По результатам этого исследования наибольшие шансы улучшить результат имеет фильтрация программой NOISY; это единственная из семи протестированных программ, которая при параметрах по умолчанию не приводит к достоверному ухудшению качества реконструкции. В связи с этим возникает вопрос о возможных условиях, при которых фильтрация выравниваний программой NOISY имеет смысл.

В качестве материала для тестирования эффективности фильтрации представляется разумным выбрать некоторый класс белков, чьи последовательности были бы неоднородны по своей эволюционной консервативности. Таким классом могут быть белки, в которых банк Pfam [8] диагностирует два эволюционных домена. Междоменный линкер, а также *N*- и *C*-концевые последовательности, как представляется, должны быть подвержены заметно более высокой частоте замен аминокислот по сравнению с доменами. Таким образом, последовательности вне доменов могут содержать больше гомоплазий, а их выравнивание может быть менее достоверным. Поэтому, как представляется, их фильтрация должна быть в среднем более целесообразна, чем фильтрация последовательностей доменов. Не исключено, что плохие результаты фильтрации в тестах из работы Tan et al. [7] вызваны превалированием в тестовых данных однодоменных белков.

В данной работе мы тестируем программу NOISY на наборе семейств белков, в которых банк Pfam диагностирует два эволюционных домена. Для этого мы создали шесть новых тестовых наборов, состоящих из ортологических рядов двухдоменных белков, при этом три набора содержат прокариотические и три – эукариотические белки. На том же материале, помимо тестирования результатов фильтрации, нами был проверен тезис о повышенной скорости накопления аминокислотных замен в междоменном линкере и концевых последовательностях по сравнению с последовательностями доменов.

МАТЕРИАЛЫ И МЕТОДЫ

База последовательностей двухдоменных белков. Из шести таксонов: царства многоклеточных животных (Metazoa), царства зелёных растений (Viridiplantae), царства грибов (Fungi), надцарства архей (Archaea), отдела актинобактерий (Actinobacteria) и отдела протеобактерий (Proteobacteria) было отобрано по 80 видов так, чтобы количество семейств эволюционных доменов, согласно версии 35 банка Pfam [8], представленных в белках всех 80 организмов, являлось максимально возможным. Для всех наборов, кроме полученного из архей, накладывались ограничения на близость видов, а именно: для Metazoa все виды были взяты из разных отрядов, для Fungi – из разных семейств, для Viridiplantae, Actinobacteria и Proteobacteria – из разных родов. Отбор видов проводился написанной нами программой, доступной по адресу https://github.com/belozersky321/PhyloBench/blob/main/selectmnems.py.

Для каждого набора из 80 видов были отобраны наборы ортологичных белков по следующим критериям:

- каждый белок включает ровно два эволюционных домена согласно Pfam;
- белки, включающие ровно два домена из этих же семейств в том же порядке, представлены не менее чем в 15 из 80 видов того же набора;
- для каждого из 80 видов имеется не более одного белка, включающего два домена из тех же семейств в том же порядке.

Тем самым отобранные белки образуют ортологические ряды объёмом от 15 до 80 белков. Списки белков, составляющих эти ортологические ряды, с указанием границ доменов доступны по адресу: https://github.com/belozersky321/2-Domain_proteins_data. Количество полученных ортологических рядов двухдоменных белков см. в таблице 1. Аминокислотные последовательности белков брались из файла pfamseq банка Pfam (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam35.0/pfamseq.gz).

Таблица 1. Ортологические ряды двухдоменных белков

Таксон	Число ортологических рядов	Среднее число белков ортологического ряда
Archaea	332	34,3
Actinobacteria	529	35,5
Proteobacteria	683	35,0
Fungi	1045	40,0
Metazoa	1742	29,6
Viridiplantae	1152	28,0

Использованное программное обеспечение. Для множественного выравнивания последовательностей использовалась программа MUSCLE [9]. Для фильтрации выравниваний использовалась программа NOISY [3]. Для филогенетической реконструкции использовалась программа FastME [10]. Все перечисленные программы использовались с параметрами по умолчанию. Бутстреп-поддержка ветвей деревьев вычислялась программой FastME с опцией -b.

Подходы к филогенетической реконструкции. Сравнивались следующие подходы. Во-первых, реконструкция делалась по: (1) выравниванию полноразмерных последовательностей белков каждого ортологического ряда; (2) выравниваниям первого или (3) второго эволюционных доменов; (4) конкатенированному выравниванию обоих доменов; (5) выравниванию обоих доменов вместе с линкером между ними, но без концевых последовательностей (отбрасывались участки до первого и после второго доменов). Во-вторых, в каждом из перечисленных вариантов реконструкция делалась по нефильтрованному и фильтрованному выравниваниям. Другой обработки выравниваний не проводилось, в частности, не удалялись никакие последовательности, даже очень близкие к другим последовательностям того же выравнивания.

Сравнение филогенетических деревьев и статистическая обработка результатов. Для каждого реконструированного филогенетического дерева было вычислено расстояние Робинсона–Фоулдса [11] до дерева соответствующих видов, полученного из таксономии NCBI [12] (деревья видов в формате Newick для всех шести сформированных нами наборов доступны по адресу https://github.com/belozersky321/2-Domain_proteins_data). Расстояние считалось написанной нами программой rf_dist_n, код которой доступен по адресу https://github.com/belozersky321/TreeDist. Считалось, что из двух реконструкций лучшей является та, расстояние от которой до дерева видов меньше. Для двух наборов реконструкций (например, полученных из фильтрованных и нефильтрованных выравниваний) набор разностей расстояний сравнивался с 0 посредством Z-критерия, то есть средняя разность делилась на стандартную ошибку средней разности, и полученное Z-значение рассматривалось как распределённое при нулевой гипотезе (эквивалентности методов реконструкций) по стандартному нормальному закону.

Для оценки статистической значимости зависимости эффекта фильтрации от различных характеристик выравнивания (числа последовательностей, длины) и связи эффекта фильтрации с бутстреп-поддержкой ветвей реконструированных деревьев использовался критерий хи-квадрат для таблиц сопряжённости.

Подбор оптимального порога по толщине выравнивания. Для подбора оптимального порога толщины выравнивания использовался принцип подбора наиболее достоверного порога, подобный описанному в работе Kalinina et al. [13]: для всех порогов была составлена таблица сопряжённости 3×2, по строкам – количества выравниваний, у которых фильтрация привела к ухудшению качества филогенетической реконструкции, не изменила качества или улучшила качество; по столбцам – толщина выше порога или толщина не выше порога. Для всех таблиц были посчитаны *p*-значения по критерию хи-квадрат. За оптимальный порог было принято значение толщины, при котором полученное *p*-значение было минимальным.

Визуализация результатов. Для построения визуализаций использовались библиотеки языка Python, а именно matplotlib версии 3.5.2 и seaborn версии 0.11.2 (функции kdeplot для плотностей распределений и barplot для столбчатых диаграмм).

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Характеристики материала исследования. Из шести таксонов было отобрано 5483 ортологических ряда двухдоменных белков, в каждом было от 15 до 80 белков, см. таблицу 1. Приведём некоторые характеристики последовательностей этих белков. Для наглядности на иллюстрациях мы объединили вместе три прокариотических набора и отдельно – три эукариотических. Для всех характеристик различия между различными прокариотическими

и различными эукариотическими наборами выравниваний оказались минимальны.

Доля последовательностей белков, занятой доменами. На рис. 1 приведены плотности распределения долей последовательностей, покрытых двумя эволюционными доменами, для отобранных нами белков. Для прокариотических белков эта доля в большинстве случаев велика, эукариотические белки более разнообразны по этой характеристике.

Скорость эволюции различается в эволюционных доменах и вне их. Для проверки предположения о более высокой скорости аминокислотных замен вне доменов мы сравнили расстояния между парами полноразмерных последовательностей, тех же последовательностей без *N*-концов (до первого домена) и *C*-концов (после второго домена), а также слитых последовательностей двух доменов без линкера между ними. Результаты в виде плотностей распределений доли расстояний, которые оказались больше в одном варианте по сравнению с другим, приведены на рис. 2. Видно, что во всех случаях чем большие части последовательностей вне доменов используются, тем чаще расстояния оказываются больше, чем при использовании только последовательностей доменов. Следовательно, исходные предположения, повлиявшие на выбор материала исследования, верны. Между различными таксономическими группами внутри прокариот и эукариот существенных различий не наблюдается, см. плотности для всех групп отдельно: https://github.com/belozersky321/2-Domain_proteins_data

Результат фильтрации. На рис. 3 приведены плотности распределения долей колонок выравниваний полноразмерных белков,

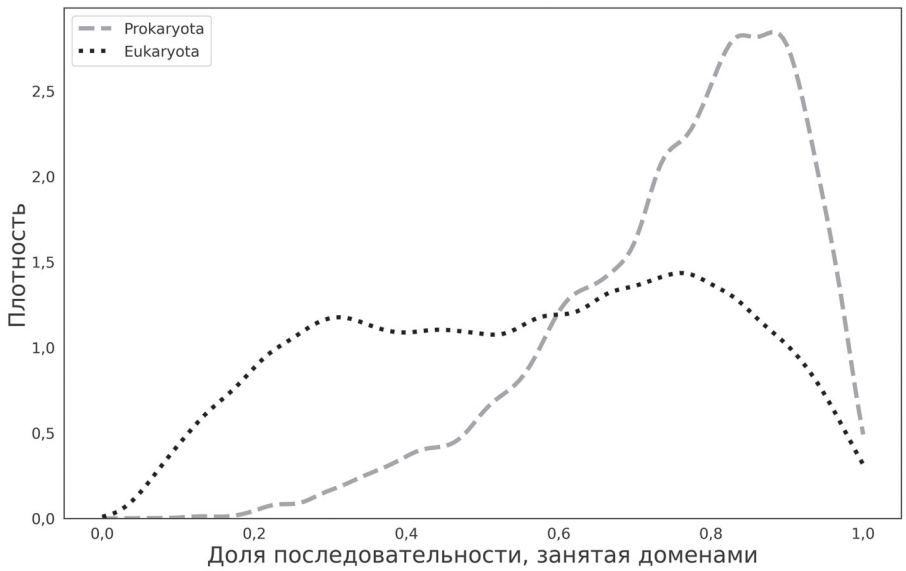

Рис. 1. Плотность распределения доли последовательностей, занятых двумя эволюционными доменами

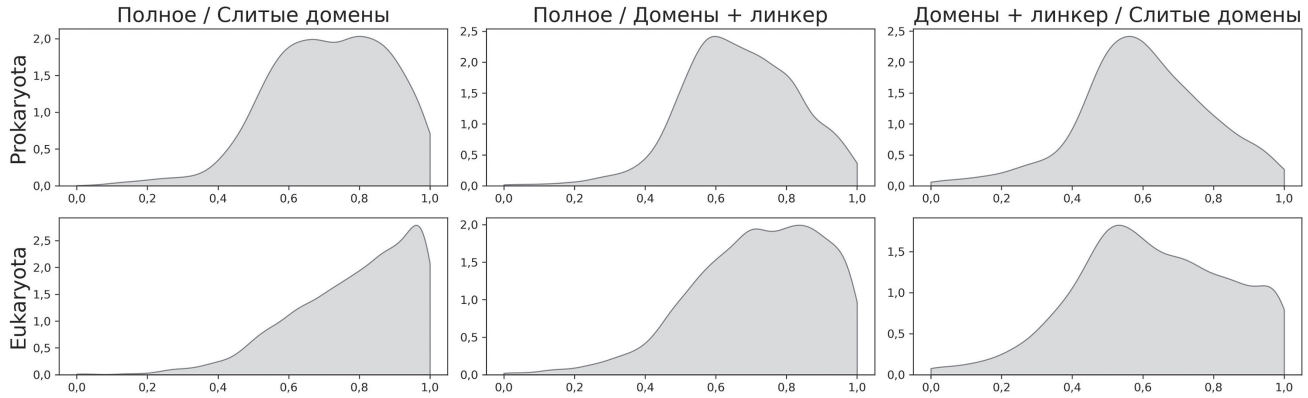

Рис. 2. Плотность распределения доли расстояний между последовательностями, которые оказываются больше в случае использования большей части последовательностей белков

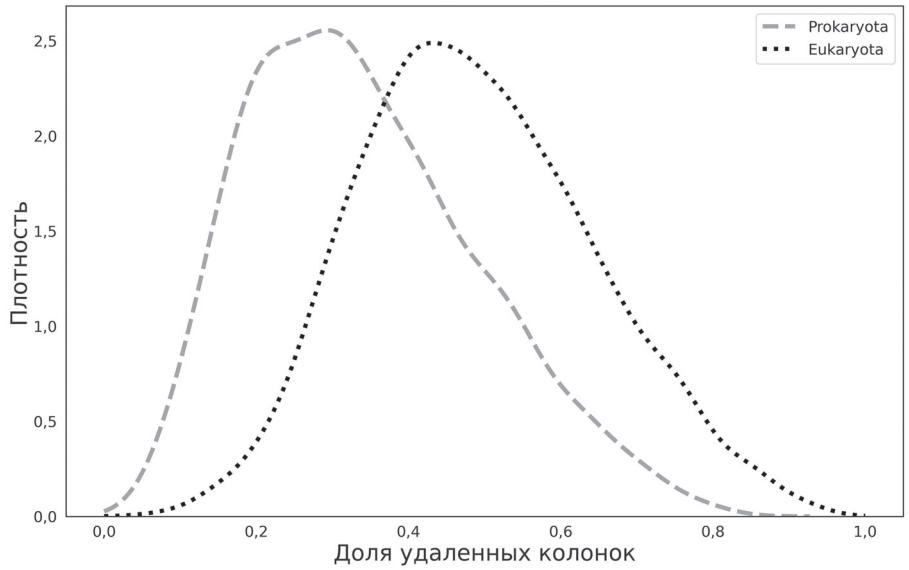

Рис. 3. Плотность распределения доли удалённых при фильтрации колонок выравниваний

которые удаляются программой фильтрации. Видно, что фильтрация затрагивает различные выравнивания в очень разной степени, причём на эукариотические влияет в среднем существенно сильнее.

Сравнение результатов филогенетической реконструкции. Результаты сравнения вариантов филогенетической реконструкции, а именно: по полноразмерным белкам, отдельным доменам, слитым доменам и доменам с линкером, с фильтрацией и без, представлены в табл. 2 и 3, а также на рис. 2. Таблица 2 содержит средние расстояния до референсного дерева для всех вариантов. Чтобы продемонстрировать статистическую значимость различий, в таблице 3 мы приводим Z-значения для сравнения результатов по разным выравниваниям с вариантом «по умолчанию», то есть с реконструкцией по полному выравниванию без фильтрации. Эффект от фильтрации для разных вариантов выравниваний проиллюстрирован на рис. 4. Показано, что фильтрация даёт в среднем преимущество только для эукариотических наборов (белков грибов, животных и растений) и только на полноразмерных белках, в остальных случаях средний результат реконструкции ухудшается от применения фильтрации. При этом достоверное улучшение ($p < 0{,}001$, Z-критерий) наблюдается только для полноразмерных белков Metazoa. Кроме того, из таблицы 2

Таблица 2. Среднее расстояние от реконструированного дерева до референсного дерева видов

Фильтрация	Выравнивание	Archaea	Actinobacteria	Proteobacteria	Fungi	Metazoa	Viridiplantae
Нет	полное	**0,450**	**0,538**	**0,602**	0,453	0,523	0,552
	домен 1	0,523	0,6231	0,657	0,605	0,644	0,664
	домен 2	0,538	0,618	0,658	0,615	0,633	0,654
	слитые домены	0,470	0,549	0,613	0,497	0,552	0,589
	домены + линкер	0,460	0,547	0,606	0,470	0,525	0,562
Да	полное	0,458	0,538	0,604	**0,447**	**0,516**	**0,550**
	домен 1	0,530	0,630	0,661	0,606	0,650	0,668
	домен 2	0,542	0,625	0,659	0,622	0,637	0,660
	слитые домены	0,479	0,557	0,616	0,501	0,556	0,591
	домены + линкер	0,469	0,548	0,608	0,471	0,527	0,562

Примечание. Жирным выделено наименьшее значение в каждой колонке.

Таблица 3. Z-Значения для сравнения разных вариантов реконструкции с реконструкцией по полному выравниванию без фильтрации

Фильтрация	Выравнивание	Archaea	Actinobacteria	Proteobacteria	Fungi	Metazoa	Viridiplantae
Нет	домен 1	12,0	13,4	13,6	29,5	31,5	25,7
	домен 2	13,3	12,4	13,5	30,4	28,4	23,7
	слитые домены	5,41	2,66	4,08	11,9	9,71	11,7
	домены + линкер	3,02	2,44	1,94	5,36	0,558	4,08
Да	полное	3,18	0,176	0,903	−2,41	−4,50	−0,852
	домен 1	12,8	14,7	13,9	29,3	33,4	26,1
	домен 2	14,0	13,6	13,4	31,7	28,8	25,1
	слитые домены	6,39	4,86	4,42	12,8	10,9	12,3
	домены + линкер	4,79	2,70	2,56	5,38	1,44	3,74

Примечание. Отрицательное значение означает преимущество варианта, указанного в строке, положительное – преимущество реконструкции по полному выравниванию без фильтрации.

видно, что полноразмерные белки во всех случаях дают лучшие результаты по сравнению с фрагментами.

Мы видим, что для неполноразмерных выравниваний эффект фильтрации во всех случаях отрицательный. Что касается полноразмерных белков, то у всех прокариот эффект фильтрации отрицательный, а у всех эукариот – положительный. В связи с этим далее мы ограничились полноразмерными выравниваниями и объединили все группы прокариот (т.е. археи, актинобактерии и протеобактерии) вместе и все группы эукариот (т.е. животных, растения и грибы) также вместе.

Связь между относительным качеством реконструкции и размером выравнивания. В следующей серии компьютерных экспериментов исследовали зависимость между влиянием фильтрации на качество филогенетической реконструкции и характеристиками входного множественного выравнивания: длиной (числом колонок) и толщиной (числом последовательностей). Медианная длина выравнивания прокариотических белков на нашем материале оказалась равной 512 аминокислотным остаткам (а.о.), медианная длина выравнивания эукариотических белков – 911 а.о. В табл. 4 приведено количество выравниваний при их разделении по длине (меньше или равной медианной либо большей медианной) и по эффекту фильтрации (в результате фильтрации расстояние от эталона уменьшилось, не изменилось либо увеличилось), отдельно для прокариотических и эукариотических выравниваний. Видно, что различия между строками незначительны для прокариот (p-значение по критерию хи-квадрат 0,62). При этом для эукариот $p = 1{,}9 \cdot 10^{-5}$. Таким образом, можно утверждать, что длина выравнивания связана с эффектом от фильтрации только для выравниваний эукариотических белков. Однако стоит отметить, что для эукариот длина выравнива-

Таблица 4. Количество выравниваний, для которых фильтрация программой NOISY привела к различным эффектам, в зависимости от длины входного выравнивания

Организмы	Эффект	Длина ⩽ медианы	Длина > медианы
Прокариоты	NOISY улучшает	86	75
	NOISY не даёт эффекта	591	589
	NOISY ухудшает	98	105
Эукариоты	NOISY улучшает	553	500
	NOISY не даёт эффекта	967	1106
	NOISY ухудшает	451	362

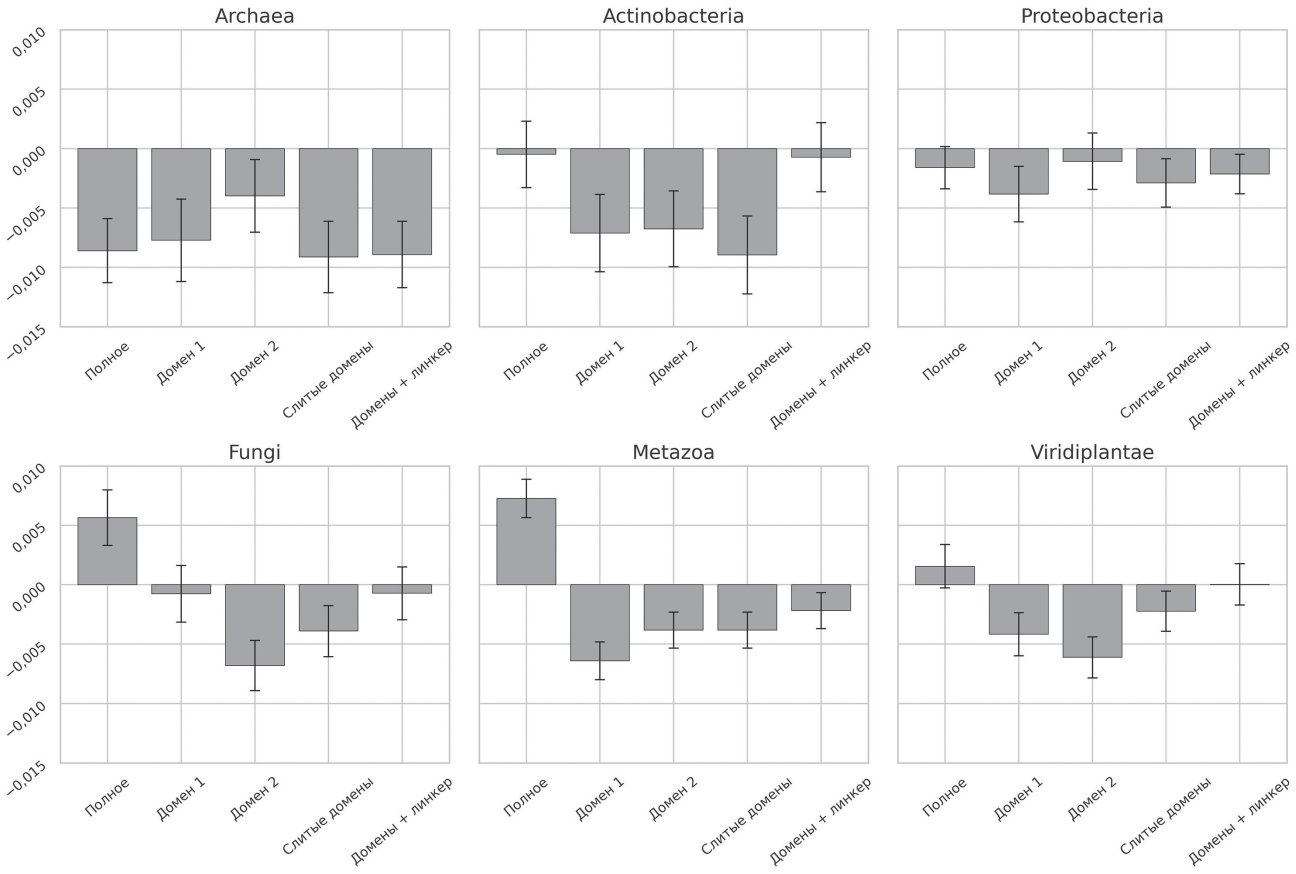

Рис. 4. Столбчатые диаграммы средних разностей между расстояниями от дерева видов до двух реконструкций: по фильтрованному и нефильтрованному выравниваниям. Из расстояния до реконструкции по нефильтрованному выравниванию вычиталось расстояние до реконструкции по фильтрованному, тем самым отрицательные значения означают преимущество нефильтрованного выравнивания. «Усы» обозначают стандартную ошибку средней разности

ния влияет прежде всего на долю случаев, когда фильтрация не имела никакого (ни положительного, ни отрицательного) эффекта.

Медианная толщина выравнивания оказалась равной 27 как для прокариот, так и для эукариот. Таблица 5, устроенная аналогично таблице 4, показывает, что толщина имеет значение: фильтрация даёт положительный эффект на выравниваниях из большого числа последовательностей существенно чаще, чем на выравниваниях из малого числа последовательностей. *P*-Значения по критерию хи-квадрат составляют 0,013 для прокариот и $6{,}6 \cdot 10^{-38}$ для эукариот.

Видно, что для эукариотических выравниваний преимущество фильтрации существенно для «толстых» выравниваний и несущественно для «тонких». Что касается прокариот, то даже

Таблица 5. Количество выравниваний, для которых фильтрация программой NOISY привела к различным эффектам, в зависимости от толщины входного выравнивания, т.е. от числа последовательностей в нём

Организмы	Эффект	Толщина ⩽ медианы	Толщина > медианы
Прокариоты	NOISY улучшает	69	92
	NOISY не даёт эффекта	616	564
	NOISY ухудшает	89	114
Эукариоты	NOISY улучшает	393	660
	NOISY не даёт эффекта	1265	808
	NOISY ухудшает	371	442

Таблица 6. Количество выравниваний, для которых фильтрация программой NOISY привела к различным эффектам, в зависимости от изменения средней бутстреп-поддержки ветвей реконструированных деревьев

Организмы	Эффект	Поддержка возросла	Поддержка не возросла
Прокариоты	NOISY улучшает	105	56
	NOISY не даёт эффекта	694	486
	NOISY ухудшает	104	99
Эукариоты	NOISY улучшает	670	383
	NOISY не даёт эффекта	1166	907
	NOISY ухудшает	468	345

Примечание. Столбец «поддержка возросла» содержит количество выравниваний, для которых средняя бутстреп-поддержка ветвей дерева, построенного по фильтрованному выравниванию, выше, чем средняя поддержка ветвей дерева, построенного по исходному выравниванию; столбец «поддержка не возросла» – количество остальных выравниваний.

после подбора порога на число последовательностей по минимуму *p*-значений, который даёт оптимальный порог 35, мы получили для выравниваний толще полученного порога, что фильтрация улучшает реконструкцию в 81 случаях и ухудшает в 85 случаях. Следовательно, для прокариотических выравниваний фильтрацию программой NOISY нельзя рекомендовать независимо от числа последовательностей в них. При этом для эукариот подбор порога дал значение 27, то есть равное медиане, поэтому для них в таблице 5 можно видеть результат разделения по оптимальному порогу.

Связь между относительным качеством реконструкции и бутстреп-поддержкой ветвей деревьев. Для оценки качества филогенетической реконструкции часто применяется бутстреп [14], позволяющий присвоить каждой ветви реконструированного дерева число («поддержку») от 0 до 100, косвенно характеризующее достоверность реконструкции данной ветви. Естественно было предположить, что если фильтрация выравнивания улучшает реконструкцию дерева, то и средняя поддержка ветвей дерева, построенного по фильтрованному выравниванию, окажется выше, чем для дерева, построенного по исходному выравниванию, и *vice versa*. В таблице 6 приведены количества выравниваний, для которых наблюдаются различные эффекты фильтрации с точки зрения расстояния до эталона и средней бутстреп-поддержки. Видно, что имеется несильная, но достоверная связь между изменением средней поддержки, с одной стороны, и изменением качества реконструкции, с другой стороны (*p*-значения по критерию хи-квадрат – 0,02 для прокариот и 0,0005 – для эукариот). При этом из данных следует, что для прокариот даже увеличение поддержки не может служить весомым аргументом в пользу применения фильтрации.

При сравнении табл. 5 и 6 возникает естественный вопрос о сочетании двух признаков выравнивания: числа последовательностей и увеличения поддержки ветвей после реконструкции по исходному и фильтрованному выравниваниям. Как видно из таблиц, этот вопрос актуален только для эукариотических выравниваний. В таблице 7 приведены количества случаев улучшения и ухудшения качества реконструкции для всех четырёх сочетаний этих признаков.

Из таблицы 7 следует, что существенные шансы улучшить реконструкцию в результате фильтрации есть только для эукариотических выравниваний, для которых одновременно выполняются оба условия: число последовательностей больше 27, и средняя поддержка ветвей дерева после фильтрации увеличивается. Впрочем, даже при выполнении только одного из этих условий количество случаев, когда фильтрация улучшает реконструкцию,

Таблица 7. Количество эукариотических выравниваний, в которых качество реконструкции изменилось после фильтрации, в зависимости от числа последовательностей и изменения бутстреп-поддержки

Толщина	Поддержка не возросла	Поддержка возросла
Толщина $\leqslant 27$	142/126 ≈ 1,13	229/267 ≈ 0,86
Толщина > 27	203/257 ≈ 0,79	239/403 ≈ 0,59

Примечание. В каждой ячейке в числителе – количество выравниваний, для которых фильтрация ухудшила результат, в знаменателе – количество выравниваний, для которых фильтрация улучшила результат. После каждого отношения приведено его приближённое значение в виде десятичной дроби.

достоверно ($p < 0,05$, критерий знаков) превышает количество случаев с противоположным эффектом.

Мы также изучали связь относительного качества реконструкции (то есть разницы между расстояниями от эталона до деревьев, построенных по исходному и отфильтрованному выравниванию) со следующими характеристиками: доля отфильтрованных позиций, доля длины, занимаемой эволюционными доменами, доля пар последовательностей, между которыми после фильтрации уменьшаются расстояния, среднее расстояние между последовательностями. Ни для одной из перечисленных характеристик значимых зависимостей обнаружено не было (данные не приведены).

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Известно, что фильтрация входного выравнивания, то есть удаление из него предположительно неинформативных, «зашумляющих» колонок, часто применяется при филогенетической реконструкции. В популярных сервисах phylogeny.fr [5] и NGphylogeny.fr [6] фильтрация включена в вариант «One click», то есть по умолчанию. В наборе программ для построения филогенетических конвейеров ETE tools [15] фильтрация выравнивания является одним из рекомендованных шагов. Проведённое несколько лет назад [7] тестирование нескольких программ фильтрации множественных выравниваний показало отсутствие положительного эффекта фильтрации. Однако не исключено, что такой результат был следствием особенностей тестовой выборки. В частности, в работе Tan et al. [7] использовалась выборка выравниваний по 30 последовательностей в каждом, что, вполне возможно, мало для эффективного выявления неинформативных колонок. Кроме того, большинство белков той выборки были небольшой длины (менее 300 а.о.) и, вероятно, включали только один эволюционный домен.

В этой связи представлялось важным исследовать эффект от фильтрации на выравниваниях, отличающихся заведомо неравномерной по колонкам структурой. Для таких выравниваний были выбраны белки с двумя эволюционными доменами. Предположительно, эволюция последовательностей в пределах доменов идёт несколько иначе, чем вне их. Мы косвенно подтвердили это сравнением расстояний между последовательностями, измеренными по полным белкам и по слитым доменам (рис. 2). В качестве программы фильтрации была выбрана NOISY [3], поскольку только для неё в работе Tan et al. [7] не было показано достоверно отрицательного эффекта её применения, в отличие от шести других программ фильтрации.

Прежде всего мы выяснили, что выравнивания полноразмерных белков всегда дают лучшие реконструкции, чем выравнивания с удалёнными частями. Тем самым распространённая рекомендация ограничиваться при реконструкции последовательностями Pfam-доменов (см., например, протокол Song et al. [16]) входит в противоречие с нашими результатами.

На нашей выборке двухдоменных белков показано, что фильтрацию программой NOISY имеет смысл применять только к выравниваниям эукариотических белков, и при этом не во всех случаях. Так, NOISY не даёт в среднем положительного эффекта, если число последовательностей в выравнивании мало (27 или меньше). Вероятно, это связано с особенностями алгоритма, основанного на оценке числа гомоплазий в каждой колонке выравнивания, что можно сделать с достаточной надёжностью лишь при достаточном числе последовательностей. Также положительный эффект фильтрации отсутствует, если из выравнивания исключены *N*- и *C*-концы белков (до начала первого домена и после второго, см. рис. 4). Ещё одним признаком целесообразности фильтрации может служить увеличение средней бутстреп-поддержки ветвей дерева (см. табл. 6). Однако этот признак оказался надёжным только в случае выполнения остальных условий, то есть в случае выравнивания достаточно большого числа эукариотических белков (см. табл. 7).

Для эукариотических белков выявилась небольшая, но достоверная ($p = 1,9 \cdot 10^{-5}$) связь эффекта фильтрации с длиной выравнивания. Однако внимательное рассмотрение таблицы 4 показывает, что длина влияет главным образом на процент случаев, когда фильтрация не оказывает никакого, ни положительного, ни отрицательного, влияния на качество реконструкции, а на соотношение между количеством случаев с положительным и отрицательным эффектом длина практически не влияет.

На основании полученных данных мы можем с некоторой осторожностью рекомендовать применение программы NOISY в случае работы с достаточно большим числом эукариотических белков, контролируя при этом изменение бутстреп-поддержки, возникающее после фильтрации. Экстраполяция данной рекомендации с двухдоменных на произвольные белки требует подтверждения. В связи с этим мы планируем проведение аналогичного исследования на эукариотических белках с

различным числом эволюционных доменов и с рассмотрением выравниваний большего числа последовательностей, чем максимально исследованное в данной работе, то есть 80. Связи эффекта фильтрации колонок со средним расстоянием между последовательностями выявлено не было, тем самым представляется маловероятным, что результаты изменятся, если убирать из выравниваний очень близкие последовательности, но мы планируем проверить и это. Возможно, стоит также протестировать программы фильтрации колонок, появившиеся после работы Tan et al. [7] 2015 года и поэтому не исследованные в ней.

ЗАКЛЮЧЕНИЕ

В настоящей работе показано, что фильтрация колонок выравниваний последовательностей белков достаточно редко приводит к улучшению качества филогенетической реконструкции, даже если рассматривать двухдоменные белки, то есть такие, где весьма вероятно очень различное качество выравнивания на разных участках. Практически можно рекомендовать фильтрацию программой NOISY только для выравниваний достаточно большого числа (более 27) полноразмерных последовательностей эукариотических белков. При выполнении указанных условий аргументом в пользу фильтрации служит увеличение бутстреп-поддержки ветвей. В остальных случаях фильтрация чаще приводит к ухудшению, нежели к улучшению качества филогенетической реконструкции.

Вклад авторов. С.А. Спирин – концепция и руководство работой; Д.Д. Латорцева, А.И. Сигорских – проведение компьютерных экспериментов; С.А. Спирин, А.И. Сигорских – обсуждение результатов исследования; А.С. Карягина – редактирование текста статьи.

Финансирование. Работа выполнена при финансовой поддержке Российского научного фонда (грант № 21-14-00135).

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Соблюдение этических норм. Настоящая статья не содержит описания каких-либо исследований с участием людей или животных в качестве объектов.

СПИСОК ЛИТЕРАТУРЫ

1. Talavera, G., and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, *System. Biol.*, **56**, 564-577, doi: 10.1080/10635150701472164.
2. Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses, *Bioinformatics*, **25**, 1972-1973, doi: 10.1093/bioinformatics/btp348.
3. Dress, A. W., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., and Stadler, P. F. (2008) Noisy: identification of problematic columns in multiple sequence alignments, *Algorithms Mol. Biol.*, **3**, 7, doi: 10.1186/1748-7188-3-7.
4. Jermiin, L. S., Catullo, R. A., and Holland, B. R. (2020) A new phylogenetic protocol: dealing with model misspecification and confirmation bias in molecular phylogenetics, *NAR Genom. Bioinform.*, **2**, lqaa041, doi: 10.1093/nargab/lqaa041.
5. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.-M., and Gascuel, O. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist, *Nucleic Acids Res.*, **36**, W465-W469, doi: 10.1093/nar/gkn180.
6. Lemoine, F., Correia, D., Lefort, V., Doppelt-Azeroual, O., Mareuil, F., Cohen-Boulakia, S., and Gascuel, O. (2019) NGPhylogeny.fr: new generation phylogenetic services for non-specialists, *Nucleic Acids Res.*, **47**, W260-W265, doi: 10.1093/nar/gkz303.
7. Tan, G., Muffato, M., Ledergerber, C., Herrero, J., Goldman, N., Gil, M., and Dessimoz, C. (2015) Current methods for automated filtering of multiple sequence alignments frequently worsen single-gene phylogenetic inference, *System. Biol.*, **64**, 778-791, doi: 10.1093/sysbio/syv033.
8. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021) Pfam: The protein families database in 2021, *Nucleic Acids Res.*, **49**, D412-D419, doi: 10.1093/nar/gkaa913.
9. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792-1797, doi: 10.1093/nar/gkh340.
10. Lefort, V., Desper, R., and Gascuel, O. (2015) FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program, *Mol. Biol. Evol.*, **32**, 2798-2800, doi: 10.1093/nar/gkh340.
11. Robinson, D. F., and Foulds, L. R. (1981) Comparison of phylogenetic trees, *Math. Biosci.*, **53,** 131-147, doi: 10.1016/0025-5564(81)90043-2.

12. Federhen, S. (2012) The NCBI taxonomy database, *Nucleic Acids Res.*, **40**, D136-D143, doi: 10.1093/nar/gkr1178.
13. Kalinina, O. V., Novichkov, P. S., Mironov, A. A., Gelfand, M. S., and Rakhmaninova, A. B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins, *Nucleic Acids Res.*, **32**, W424-W428, doi: 10.1093/nar/gkh391.
14. Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap, *Evolution*, **39**, 783-791, doi: 10.1111/j.1558-5646.1985.tb00420.x.
15. Huerta-Cepas, J., Serra, F., and Bork, P. (2016) ETE 3: reconstruction, analysis, and visualization of phylogenomic data, *Mol. Biol. Evol.*, **33**, 1635-1638, doi: 10.1093/molbev/msw046.
16. Song, L., Wu, S., and Tsang, A. (2018) Phylogenetic analysis of protein family, in *Fungal Genomics. Methods in Molecular Biology* (de Vries, R., Tsang, A., Grigoriev, I., eds) vol. 1775, Humana Press, New York, pp. 267-291, doi: 10.1007/978-1-4939-7804-5_21.

HOW OFTEN DOES FILTERING ALIGNMENT COLUMNS IMPROVE PHYLOGENETIC INFERENCE OF TWO-DOMAIN PROTEINS?

A. I. Sigorskikh[1], D. D. Latortseva[1], A. S. Karyagina[2,3,4], and S. A. Spirin[3,5]*

[1] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119992 Moscow, Russia*

[2] *Gamaleya National Research Center of Epidemiology and Microbiology, Ministry of Healthcare of the Russian Federation, 123098 Moscow, Russia*

[3] *Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, 119992 Moscow, Russia; e-mail: sas@belozersky.msu.ru*

[4] *All-Russia Research Institute of Agricultural Biotechnology, 127550 Moscow, Russia*

[5] *National Research University Higher School of Economics, 109028 Moscow, Russia*

The phylogeny of proteins is usually reconstructed based on multiple alignments of their amino acid sequences. One of the problems of such alignments is the presence of regions of different conservation, including those regions where the quality of the alignment is questionable. To solve this problem, the filtering of alignment columns is often used, with special software developed for this purpose. In this work, various approaches to phylogeny reconstruction are investigated using proteins with two evolutionary domains as examples. The sequences of such proteins are inherently heterogeneous in conservation due to the presence of both evolutionary domains and linkers between domains, as well as N- and C-termini. It is shown that filtering the alignment columns on average improves the quality of reconstruction only when full-length sequences are used and only for eukaryotic proteins. It is also shown that restriction of an alignment to evolutionary domains with rejection of less conserved linkers and end sequences worsens the quality of phylogenetic reconstruction on average.

Keywords: phylogenetic inference, evolutionary domains, filtration of multiple sequence alignment