Тема работы:

РАЗРАБОТКА ВЫЧИСЛИТЕЛЬНОГО КОНВЕЙЕРА ДЛЯ ПОИСКА И АНАЛИЗА ГЕНОВ СЕМЕЙСТВ МНОГОДОМЕННЫХ БЕЛКОВ В ГЕНОМАХ

Состав коллектива:

1. Бочарникова Мария Евгеньевна, ММФ НГУ (гр. 21152), ИЦиГ СО РАН, почта: m.bocharnikova@g.nsu.ru , lachynova@bionet.nsc.ru
2. Афонников Дмитрий Аркадьевич, к.б.н., в.н.с. ИЦиГ СО РАН, почта: ada@bionet.nsc.ru

Информация о гранте:

The work was funded by the Kurchatov Genome Center of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, agreement with the Ministry of Education and Science of the Russian Federation, no. 075-15-2019-1662

Научное содержание работы:

1. Постановка задачи:

Целью работы является создание автоматического вычислительного конвейера для поиска и анализа белков-ортологов в геномах групп организмов.

Задачи:

1. Анализ литературы и выбор средств для решения задачи

2. Создание вычислительного конвейера с использованием платформы Snakemake

3. Оценка точности идентификации ортологов для созданного конвейера

4. Применение конвейера для анализа белков семейства фосфолипаз А2 у плоских червей

2. Современное состояние проблемы (на момент начала работы).

Одна из важных задач для биологов при исследовании геномов – предсказание функций генов. Она решается двумя способами: с помощью поиска функциональных доменов белков и с помощью поиска генов-ортологов в геномах различных организмов. Второй метод позволяет идентифицировать функцию генов в масштабе всего генома, однако для некоторых белков, которые включают несколько функциональных доменов, этот метод дает неточную классификацию.

В настоящей работе предложено использовать комбинацию двух подходов. Это решение требует выполнения нескольких шагов обработки больших биоинформатических данных по анализу геномных последовательностей. Решение таких задач в биоинформатике основывается на использовании вычислительных конвейеров. В работе мы остановились на использовании системы управления рабочими процессами Snakemake, так как она позволяет проводить вычисления под управлением командной строки на высокопроизводительном кластере и интегрировать программные пакеты с использованием менеджера Conda.

Ключевой этап в реализуемом конвейере по анализу ортологов, отличающий его от других биоинформатических инструментов подобного типа – фильтрация полученных последовательностей по доменному составу белков. Мы предположили, что такая фильтрация позволяет получить более точные результаты, так как благодаря этому из ортогруппы исключаются последовательности с доменным составом не соответствующим функции исследуемых белков.

3. Подробное описание работы, включая используемые алгоритмы

Данный конвейер был реализован с использованием систему управления рабочими процессами Snakemake. Конвейер был реализован на суперкомпьютере на операционной системе Linux. В ходе реализации использовались – абстрактный синтаксис Snakemake, язык программирования python и его библиотеки и bash-скрипты. На рис.1 изображен граф рабочих процессов конвейера OrthoDom.

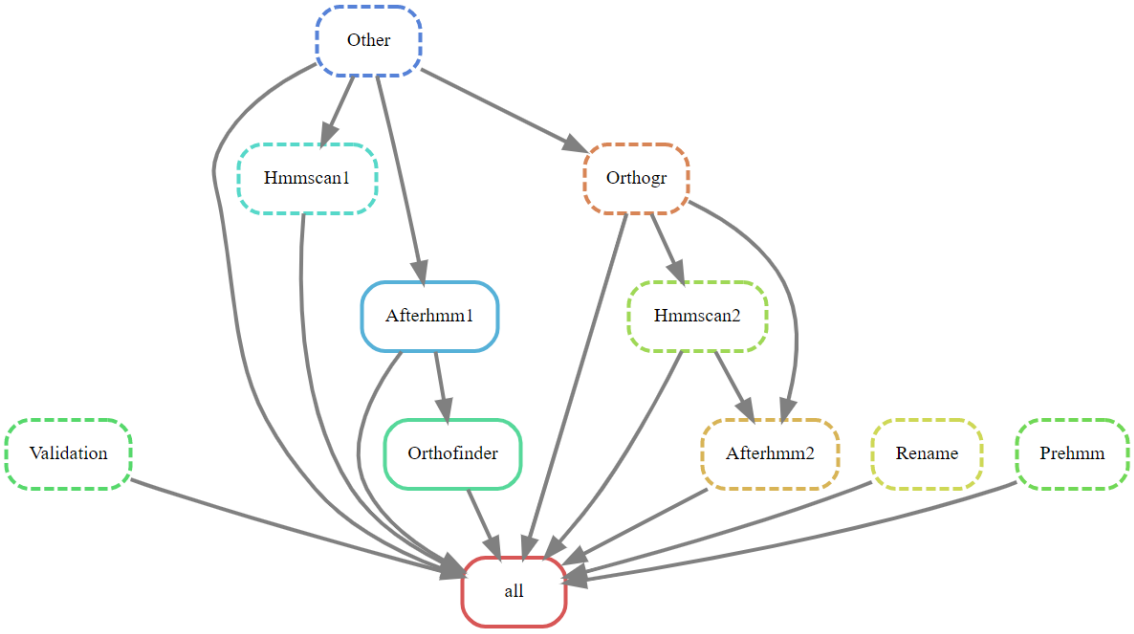

Рис.1 Зависимость рабочих процессов конвейера OrthoDom представленная в виде графа рабочих процессов

Рассмотрим подробнее структуру конвейера по этапам работы.

Этап 1 – загрузка данных:

1. Загрузка репозитория github

2. После загрузки, в корневой папке пользователя появляется папка data. В ней хранятся все входные, выходные и промежуточные данные. В папке data находятся еще четыре папки – input, output, work, scripts. Пользователю необходимо загрузить входные данные в папку input самостоятельно.

3. В папке input хранятся еще три – ref_dom, ref_fam, ref_prot. В них нужно будет загрузить следующие файлы:

- ref_prot (геномы с белковыми последовательностями, в которых будет осуществляться поиск генов). Каждая последовательность имеет формат fasta. Для того, чтобы в программе orthofinder эти геномы назывались определенным образом, название файла должно иметь следующий формат (стандартизованный для всех геномов): расширение .fa, три первых буквы берутся от названия рода, три следующих буквы - начало названия вида. При этом первые символы заглавные. Пример названия последовательностей в папке – AraTha.fa.

- ref_fam (референсные белки исследуемых семейств, которые считаются представительными). Каждая последовательность имеет формат fasta (расширение .fa). Для того, чтобы в программе orthofinder эти наборы последовательности можно было различить по организмам, названия файлов должны соответствовать таковым из папки ref_prot (см выше). Мы полагаем, что для одного организма в этом файле может быть несколько белковых последовательностей. Эти последовательности относятся к возможно нескольким ортогруппам (число ортогрупп не равно числу последовательностей). Это учитывается в названии последовательностей. Так что название ортогруппы стоит первым до знака разделителя (подчерк) - дальше - любые символы. Названия этих ортогрупп должны быть одинаковы для разных организмов референсных семейств. Но наличие ортогруппы во всех организмах не обязательно. Пример названия файла: организм Homo sapiens; название файла – HomSap.fa.

- ref_dom (профили формата hmm (расширение .hmm), из базы данных pfam, соответствующие исследуемым референсным доменам). Их может быть несколько. При этом, мы считаем корректным, что результат поиска доменов в искомой последовательности положительный, если в ней встречается хотя бы один из референсных доменов. Названия файлов соответствуют названиям доменов Pfam. Далее во всем конвейере мы будем использовать названия этих доменов из названия файлов. При этом названия доменов берутся из файлов. Пример названия файла - LCAT.hmm, где LCAT – название домена.

- Далее работает скрипт checkon.py, который проверяет – формат файлов, расширения, название файлов и соответствующее содержание, а также считает статистику (для матриц – кол-во доменов и их длины, для последовательностей – кол-во последовательностей в каждом файле и название файлов).

- При возникновении ошибок – пользователь вручную исправляет файлы. Если валидация прошла успешно - конвейер автоматически продолжает работу.

Этап 2 - Проверка наличия доменов в аннотированных последовательностях:

1. При помощи программы hmmscan, производится поиск референсных белков по загруженным на прошлом этапе матрицам в референсных последовательностях. Выводится статистика (по умолчанию программы hmm). Программа hmmscan базируется на алгоритме скрытых марковских цепей.

2. Из файлов экстрагируются последовательности с длинной меньше параметра A, e-value меньше параметра B и последовательности без референсных доменов (они хранятся в папке (works->exclude1). Параметры A и B пользователь прописывает самостоятельно в Snakemake rule HMM.

Этап 3 - Поиск ортогрупп в протеомах, содержащих референсные домены:

1. Запускается программа orthofinder. Которая кластерезует ортологичные гены. Алгоритм данной программы базируется на глобальном выравнивании последовательностей. Результаты хранятся в отдельной папке.

2. Далее экстрагируются ортогруппы содержащие референсные белки.

Этап 4 - Проверка наличия доменов в ортогруппах:

1. Аналогично этапу 2 - происходит поиск референсных доменов в последовательностях ортогрупп. Из файлов экстрагируются последовательности с длинной меньше параметра A, e-value меньше параметра B и последовательности без референсных доменов (они хранятся в папке (works->exclude2). Параметры A и B пользователь прописывает сам в Snakemake rule HMM.

Этап 5 - Анализ данных:

1. Выводится статистика в виде таблицы1: строки – названия организмов, столбцы – ортогруппы, ячейки – белки

2. Выводится статистика в виде таблицы2: строки – названия организмов, столбцы – ортогруппы, ячейки – кол-во найденных белков

3. Создаются текстовые файлы для каждой ортогруппы, в которой хранятся названия входящих в нее последовательностей

4. Полученные результаты.

Для того, чтобы оценить точность работы конвейера, мы обратились к базе данных OrthoDB, которая представляет из себя каталог групп ортологичных генов, кодирующих белки, у позвоночных, членистоногих, грибов, растений и бактерий. Ортология относится к последнему общему предку рассматриваемого вида, и, таким образом, OrthoDB явно выделяет ортологов в каждом основном направлении филогенеза вида.

Мы сравнивали получившиеся ортогруппы, после реализованного нами конвейера OrthoDom (предсказанние) со списками белков из ортогрупп БД OrthoDB (истинная классификация) для одних и тех же организмов (см. рисунок 2). Также сравнивали состав ортогрупп, полученный после выполнения Orthofinder без использования доменной фильтрации последовательностей.

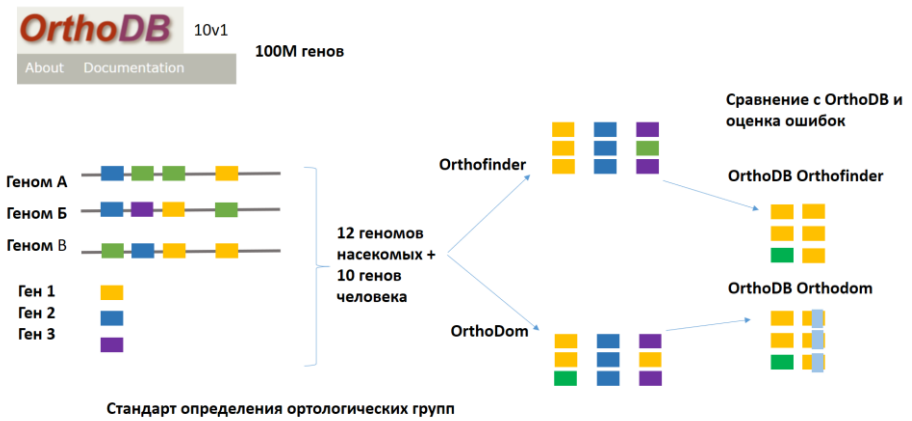

Рис.2 Схема сравнения ортогрупп, получившихся в результате работы конвейера OrthoDom и программы Orthofinder с ортогруппами из БД OrthoDB.

TP — истинно-положительное решение;
FP — ложно-положительное решение;
FN — ложно-отрицательное решение.

$$SP = \frac{TP}{TP + FP} * 100; \text{если } FP = 0, \text{то } SP = 100$$

$$SN = \frac{TP}{TP + FN} * 100; \text{если } FN = 0, \text{то } SN = 100$$

$$AC = \frac{SN+SP}{2}; \text{если } FN = FP = 0, \text{то } AC = 100$$

$$F1 = 2 * \frac{SN*SP}{SP+SN}; \text{если } FN = FP = 0, \text{то } AC = 100$$

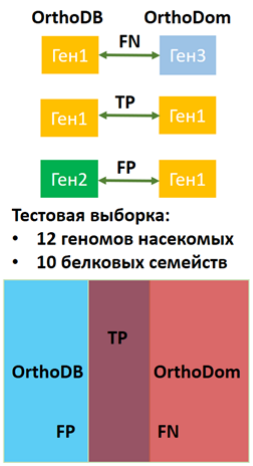

Рис.3 Формулы для расчета метрик SP, SN, AC, F1, где TP – последовательность есть и OrthoDB и в OrthoDom, FP – последовательность есть в OrthoDom, но нет в OrthoDB, FN – последовательности нет в OrthoDom, но есть OrthoDB.

5. Иллюстрации, визуализация результатов.

Таб 1. AC и F1 после запуска конвейера OrthoDom и программы Orthofinder.

Название белка	AC (Orthodom) %	AC (Orthofinder) %	F1 (Orthodom) %	F1 (Orthofinder) %
Granzyme B	21,73	0,00	2,62	0,00
PPIase A	83,01	80,46	82,05	78,10
Tektin-2	65,31	62,18	46,87	46,15
TNNI1	100,00	96,43	100,00	96,30
Cytochrome c	100,00	97,83	100,00	97,78
Low density lipoprotein receptor	96,88	89,47	96,77	88,24
Palmitoyltransferase	72,92	67,85	62,86	63,15
Transthyretin	100,00	96,42	100,00	96,29
SETX	67,83	66,86	59,46	61,54
ATPase Ca++ transporting cardiac muscle	98,00	67,86	97,96	63,16
Средние значения	81	73	75	69
		Δ 8		Δ 6

В результате, разница между ортогруппами составила 8 по accuracy и 6 по F1. То есть качество после запуска конвейера OrthoDom действительно улучшилось.

Конвейер был протестирован на 12 геномах насекомых и 12 белковых последовательностях. В результате, как описано выше, разница метрик составила 3-4 процента. Также, конвейер был запущен с 3, 6, 9, 12 белками и 3, 6, 9, 12 геномами. При варьировании белков время работы почти постоянно, а при увеличении кол-ва геномов – время работы конвейера возрастает линейно.

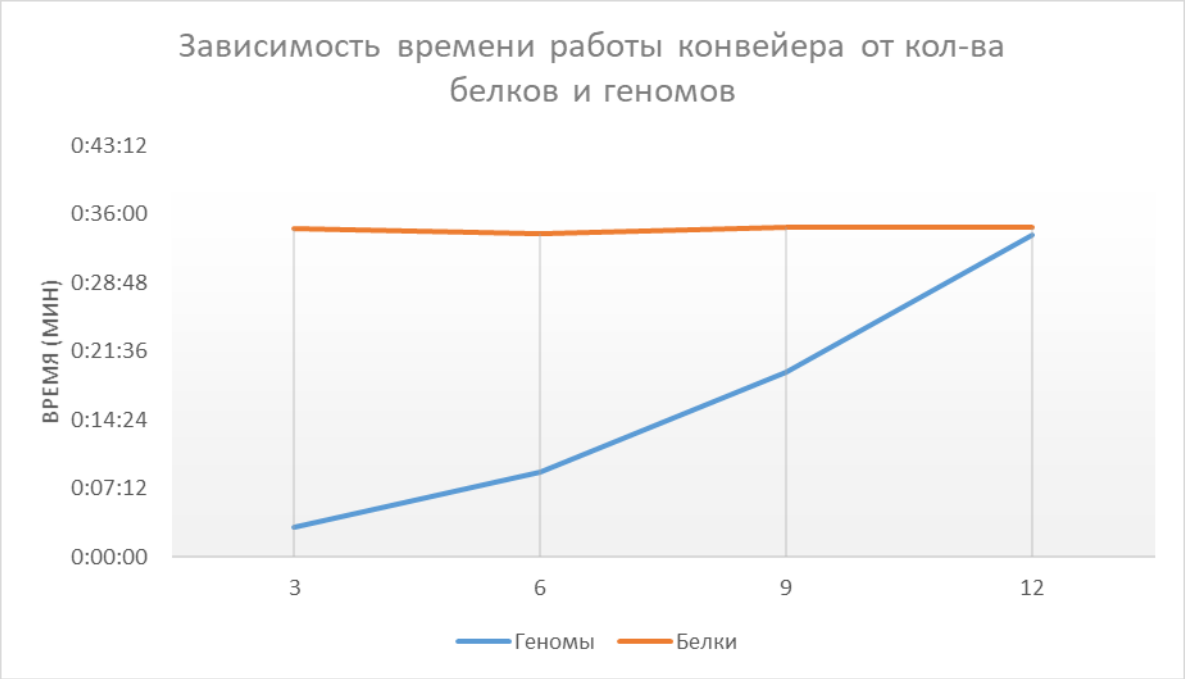

Рис 4. Зависимость времени работы конвейера OrthoDom от количества белков и геномов.

Эффект от использования кластера в достижении целей работы.

Все этапы программы, были поочередно реализованы и протестированы на кластеры. Эффект положительный.

Перечень публикаций, содержащих результаты работы (если есть). Если имеется, указать импакт-фактор журнала (Thomson Reuters, РИНЦ,…).

Турнаев И. И., Бочарникова М. Е., Афонников Д. А. Фосфолипазы А2 человека: функциональный и эволюционный анализ //Вавиловский журнал генетики и селекции. – 2023. – Т. 26. – №. 8. – С. 787-797.

Опционально: ваши впечатления от работы вычислительной системы и деятельности ИВЦ НГУ, а также предложения по их совершенствованию.

Хотелось бы иметь подключение к сети Интернет

АННОТАЦИЯ:

Геномы живых организмов содержат десятки тысяч генов, кодирующих белки. Каждый белок экспрессируется в результате транскрипции и трансляции и принимает уникальную трехмерную структуру, которая, как правило, состоит из нескольких функциональных доменов. Каждый домен отвечает за специфические взаимодействия с лигандами и обуславливает функцию белка.

Функции белка определяются комбинацией доменов, один из которых, как правило, является основным (ферментативным).

Биологов интересует распознавание функций белков, кодируемых в геномах. Поскольку количество генов велико, то решение такой задачи связано с необходимостью вычислительной обработки больших геномных данных.

Главной целью работы является создание автоматического вычислительного конвейера для поиска и анализа белков-ортологов в геномах групп организмов.

Данный конвейер может использоваться молекулярными биологами и биоинформатиками для более точного анализа многодоменных белков в геномах немодельных организмов.