Домашнее задание №0

Иванов Вячеслав, группа 699

28 октября 2018 г.

Оглавление

1 Проект по информатике:

1.1 Предисловие

Репозиторий на GitHub: https://github.com/ivanov-v-v/eQTL_analysis. Проект находится в активной разработке, подача текста на публикацию запланирована на зиму-весну 2019 года. Большая часть кода написана на Python (в формате Jufpyter Notebook), а некоторые скрипты — на R[1]. Поскольку это мой первый проект такого масштаба, и опыта промышленной разработки у меня не было, его структура не соответствует стандартам reproducible research, но у меня есть детальный план рефакторинга, который будет выполнен в осеннем семестре этого года.

На втором курсе, в рамках математического практикума ФИВТ, я, под руководством Юрия Львовича Притыкина, начал писать научную работу по системной биологии. Это междисциплинарная наука, цель которой — максимально точно смоделировать взаимосвязи элементов биологических систем, таких как: метаболические пути, клеточные структуры, клетки, ткани и т.д. Её расцвет пришёлся на последние десятилетия и связан с темпами развития экспериментальной биологии.

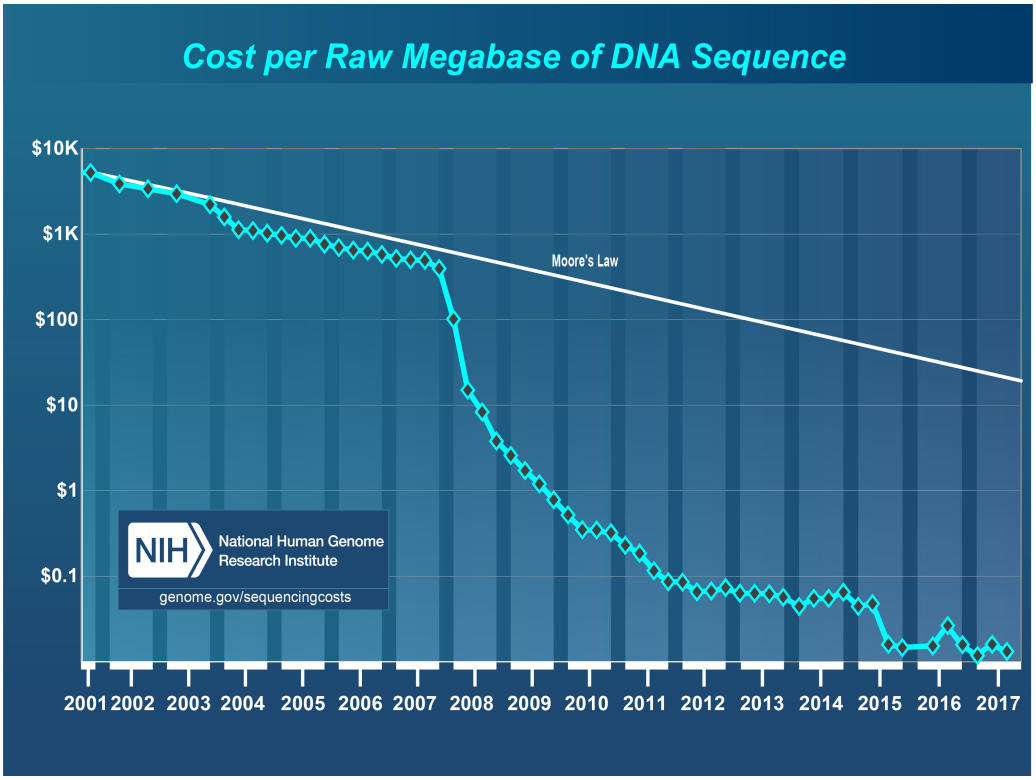

Рис. 1: с 2003 года стоимость секвенирования снижается даже быстрее, чем по закону Мура

Как и вся современная биология, это вычислительная наука, основанная на обработке огромных массивов данных. Считается нормой работать с геномом, протеомом[2] и интерактомом[3] целиком. Что мы и делаем, в частности. За подробностями можно обратиться к странице проекта.

[1] я очень не люблю, но вынужден использовать R, т.к. некоторые биостатистические пакеты не портированы под Python и не имеют аналогов

[2] протеом — совокупность белков, синтезируемых данной клеткой

[3] интерактом — сеть молекулярных взаимодействий в клетке. В контексте нашего исследования — сеть белок-белковых взаимодействий

1.2 Зачем изучать QTLs?

Мы исследуем взаимосвязь между eQTLs[4] и pQTLs — участками генома (т.н. маркерами), управляющими экспрессией[5] мРНК и белков соответственно. QTLs активно изучают, т.к. это святой Грааль биологии — точная, количественная взаимосвязь между генотипом и фенотипом. В том числе и с патологическим фенотипом: к примеру, недавняя статье в "Nucleic Acids Research"посвящена поиску eQTLs в раковых клетках, и многие из них определяют прогноз.

1.3 Зачем сравнивать eQTLs и pQTLs?

Несмотря на очевидную связь между транскрипцией[6] и трансляцией[7], почти для всех генов множества eQTLs и pQTLs практически не пересекаются. Кажется, что взаимосвязи и вовсе нет: с точки зрения регуляторных механизмов, эти процессы будто происходят сами по себе. В научном сообществе ведутся оживлённые дебаты по этому поводу, поскольку решение этой проблемы позволило бы значительно повысить эффективность персонализированной медицины. Если мы не знаем, к каким изменениям в протекании основных клеточных процессов приведёт вмешательство в геном, мы не имеем морального права использовать имеющиеся технологии, хотя именно на них основываются перспективные методы лечения онкологических заболеваний, такие как иммунотерапия.

1.4 В чём новизна нашего исследования?

Мы решили посмотреть на проблему с высоты птичьего полёта. Если какие-то гены вместе входят в функциональный модуль[8], то наборы маркеров, управляющие их экспрессией, должны быть схожи. По крайней мере, так подсказывает здравый смысл: если однотипные элементы системы ведут себя согласованно, то управляющие ими механизмы должны быть устроены похожим образом. Такой подход встречается в литературе, но у нас совершенно иное разрешение и необычные методы. Более того, наши данные по белковой экспрессии богаче тех, что имеются в общем доступе: экспрессию белков долго, дорого и сложно измерять, в отличие от экспрессии мРНК.

1.5 Инженерная сторона вопроса: какими инструментами мы пользуемся?

1. **Системы контроля версий:**
   Репозиторий проекта хранится на *GitHub* и регулярно обновляется там.

2. **Среда разработки:**
   Работа над проектом в интерактивном режиме ведётся в *Jupyter Notebook*. Работа над скриптами, структурой проекта, деплойментом и системой контроля версий на языке Python ведётся в среде *Pycharm* от компании *JetBrains*, а на языке R — в *RStudio*.

3. **Обработка и хранение данных:**
   Несмотря на то, что природа наших наблюдений не должна зависеть от конкретного организма, на данный момент мы анализируем данные, полученные для *Saccharomyces Cerevisiae* — обыкновенных дрожжей. Это заодно позволяет обойти сложности, возникающие при работе с многоклеточными организмами (например, QTLs становятся тканеспецифичны). Тем не менее, мы обязательно проверим нашу гипотезу ещё как минимум на крысах и людях.

[4] QTL — quantitative trait loci, локус количественных признаков

[5] экспрессия — процесс, в ходе которого информация, закодированная в ДНК, используется дли синтеза биологических молекул (в нашем случае — матричной РНК и белков)

[6] транскрипция — процесс синтеза мРНК на основе ДНК

[7] трасняция — процесс синтеза белка на основе мРНК

[8] функциональный модуль — множество генов, согласованно участвующих в осуществлении некоторого клеточный процесс (в нашем случае — на уровне белков)

Чрезвычайная популярность дрожжей в качестве модельных организмов в генетике привела к беспрецедентному разрешению всех типов данных, которые мы используем: экспрессии мРНК и белков, геномных аннотаций[9], графов белок-белковых взаимодействий.
Такой подход позволил нам избежать шума в данных: нам не пришлось производить их предварительную обработку, всё уже сделано до нас. Большая часть данных хранится в формате .csv. Обращение к ним происходит посредством библиотеки *pandas*.
Для обеспечения data-persistency и компактного хранения результатов мы используем *pickle*.

4. **Работа с графами:**
Интерактом дрожжей представляет собой большой псевдограф (тысячи вершин и сотни тысяч рёбер). Его размеры и необходимость его частой рандомизации для проверки достоверности результатов ставят высокие требования к производительности используемых библиотек. В данный момент мы используем *igraph*, а точнее — обёртку над ней для языка Python. Тем не менее, на Python она портирована отвратительно (мне приходилось править в ней баги вручную), в связи с чем мы планируем перейти на *graph-tool*.

5. **Работа с данными, оптимизация и статистика:**
Мы используем стандартный стек — *numpy*+*scipy*+*sklearn* с редкими вкраплениями биоинформатических библиотек на R. Для тестирования гипотез мы используем *scipy.stats.mannwhitneyu*, для FDR-коррекции — пакет *qvalue*, для построения линейных моделей — *sklearn.linear_model*, класс *LinearRegression* (аналог lm в R).

6. **Работа с кластером МФТИ:**
Объём и природа наших вычислительных задач, подразумевающая эффективную параллелизацию, естественным образом привела нас на кластер МФТИ. Папки проекта в автоматическом режиме синхронизируются с копией на кластере с помощью *rsync*. За взаимодействие с кластером отвечает slurm. Типичная рабочая сессия выглядит так: подключиться к кластеру по ssh-тоннелю, запросить вычислительную единицу с 32-64 ядрами, запустить на ней Jupyter Notebook в интерактивном режиме и подключиться к нему с локальной машины. Распараллелирование реализуется с помощью библиотеки *joblib*.

1.6 Чего мы достигли на данный момент?

Промежуточные результаты показывают, что для функциональных модулей действительно наблюдается гораздо более существенное сходство между множествами eQTLs и pQTLs. Причём, как на простейших тестах, так и на довольно специфичных, выявляющих неочевидные взаимосвязи между маркерами: например, мы обнаружили, что существенную часть pQTLs можно предсказать на основе eQTLs, что интересно само по себе.
В данный момент мы занимаемся внедрением идей и данных, опубликованных в недавней статье из журнала *eLife* — "Genetics of trans-regulatory variation in gene expression". Внушает оптимизм, что их данные, покрывающие практически весь геном дрожжей, дают более сильный результат, чем те, которые у нас имелись раньше. Тем не менее, статья содержит гораздо более современный и в некотором смысле даже спорный подход к поиску QTLs в геноме, в связи с чем окончательные выводы пока делать рано.

2 Какой предмет и преподаватель больше всего понравился за 2 курса и почему?

Не могу сказать, что какой-то из предметов вызвал у меня восторг. Всё же, моя осознанная любовь это scientific computing — математика на компьютерах. Похожие предметы появились только на третьем курсе, а по отдельности эти компоненты вызывают у меня гораздо меньше

[9] аннотация гена — совокупность стандартизированных терминов, описывающих его функцию в организме

энтузиазма. Хотелось бы отметить, пожалуй, только курс *дискретного анализа*, читаемый А.М. Райгородским. В нём собраны удивительно красивые теоремы дискретной математики, настоящие жемчужины этой науки. К ним тяжело остаться равнодушным.
Преподаватели — дело другое. Они — настоящее богатство Физтеха, его основа. В моей памяти точно останутся двое: Вадим Витальевич Редкозубов и Аркадий Борисович Скопенков. Вадим Витальевич все два года вёл у нас математический анализ. Человеку, знакомому с реалиями Физтеха, многое скажет тот факт, что его лекции стабильно собирали полный зал. Неподдельная интеллигентность Редкозубова, его любовь к уместной шутке, разбавляющей строгость и полноту, с которой он подавал материал — всё это вызывало уважение и любовь у студентов. Я дважды сдал ему экзамен на 10, а в четвёртом семестре посещал его семинары. Я очень ценю то, как он показывал нам глубину анализа, учил нестандартному подходу, искренне хотел, чтобы мы понимали, о чём говорим. В его лекциях временами встречался продвинутый материал, выходивший за рамки физтеховской программы. Мы с ним неоднократно обсуждали возможные усовершенствования курса матанализа на ФИВТе, он был открыт к новым предложениям и искренне хотел сделать особенный курс, в полной мере задействующий ту математическую базу, которую даёт наш факультет.
Акрадий Борисович Скопенков был нашим семинаристом по дискретному анализу. Как матгруппники, мы также прослушали семестровый курс общей топологии в его исполнении. Признаться, Аркадий Борисович — самый сложный преподаватель в моей жизни, и это мнение разделяют мои одногруппники. Его можно считать эталоном математической строгости: малейшая неточность доказательства не проходила мимо его ушей. Чего стоили только "идеальные письменные решения"домашних задач, приличный балл за которые совмещался с практической невозможностью сдать их с первого раза. Метка '(1)', поставленная красной ручкой и указывающая на "бессмысленную фразу после которой, ясное дело, работа далее не проверялась, прочно вошла в ФИВТовский фольклор. Скопенков перебарщивал с формализмом, с ним было тяжело, но строгость мышления, к которой меня приучили его занятия, останется со мной навсегда. Понимание опасности неточно сформулированных мыслей, пропущенных "по очевидности"логических шагов и неверно данных определений здорово помогает мне как в программировании, так и в жизни. После Скопенкова, после сотен олимпиадных задач человек не остаётся прежним. И, на мой взгляд, он всё же меняется к лучшему...

3 Академия? Индустрия? Клиника?

Мой вектор развития допускает все три варианта, но выбрать между ними я пока что не в состоянии. Одно я знаю точно: я пойду туда, где моя работа будет приносить наибольшую пользу людям, и где я смогу заниматься наукой, ведь я не могу себе представить, как я мог бы заниматься чем-нибудь ещё.
Если говорить предметно, то у каждого варианта есть свои плюсы и минусы:

- В академии больше свободы в исследованиях. При удачном стечении обстоятельств и должной квалификации можно позволить себе роскошь заниматься тем, что субъективно кажется наиболее важным. Тем не менее, в академии легко потерять связь с реальностью, а для меня это именно то мерило, которое определяет ценность моей работы. Если исследование принесёт пользу науке и технике через пару столетий, я не готов столько ждать. К тому же, меня не привлекает преподавание и я хотел бы всеми способами такой необходимости избежать.

- У индустрии есть ресурсы, которые позволяют им двигаться в ногу со временем. Есть деньги для найма лучших людей, есть чёткое понимание цели и выбора средств для её достижения. Кроме того, мне нравится соревновательный дух и желание всё время быть на горизонте событий. Тем не менее, raison d'être существования любой частной компании — извлечение прибыли. К сожалению, многие из них этим и ограничиваются. Я не готов работать только

ради денег, мне нужно знать, что я живу эту жизнь не зря. Возможно, мне бы понравилось в государственных исследовательских корпорациях в духе NASA, где работа построена вокруг цели, а не на выгоде от её достижения. Деньги — рычаг, поворотом которого можно изменить мир, но жизнь не стоит выстраивать вокруг них.

- Клиника — специфичное место. Оно пропитано бюрократией и неохотно принимает новое. В клинике программист не чувствует себя на вершине мира, куда его усердно возносят корпорации в духе Google: он помогает врачу, но не может его заменить. В клинике нет фиксированных рабочих часов: болезнь не спрашивает, когда прийти. Но на войне как на войне, ведь в клинике идёт непрекращающаяся борьба за чужие жизни, и твои решения могут напрямую спасать людей. Среди тех задач, которые решают программисты, эта — одна из наиболее благородных.