The biomolecular universe: inception

Петр

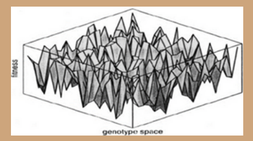
Илья

Состав участников

Лилия Богдан

Алексей Матяш

Ирина Малышева

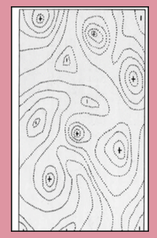


1 ОПИСАНИЕ ПРОЕКТА

«Ландшафт фитнеса» - ключевой концепт в эволюционной биологии, отображающий взаимосвязь между генотипом и репродуктивным успехом (фитнесом). В таком представлении, о сходных (по последовательностям) генотипах говорят, что они «близки» друг к другу, о сильно различающихся, что они «далеки» друг от друга, а по вертикали откладывают фитнес. Это удобная теоретическая конструкция - однако, несмотря на годы теории и экспериментов, мы по-прежнему мало знаем о реальном устройстве этого ландшафта, особенно на глобальном масштабе.

Рассматривая пространство последовательностей и «населяющих» его точек, вполне естественно перейти к наблюдению некоторых **областей повышенной и пониженной плотности точек**, реально встречающихся в природе – функциональных, «подходящих» последовательностей. Тогда можно оперировать геометрическими и топологическими категориями для изучения этого пространства. До недавнего времени количество доступных (секвенированных) биологических молекул (ДНК, РНК, белки) было недостаточно велико и просто **не позволяло веси вышепомянутый анализ** - но теперь аннотированных последовательностей - а значит, и точек в пространстве последовательностей – достаточно велико.

В нашем проекте мы использовали подходы из геометрии и топологии для изучения распределения белков в их «пространстве последовательностей» и на ландшафте фитнеса на разных - и больших и малых - расстояниях. Для отдельных семейств белков мы оценили «индивидуальные» свойства их пространства последовательности - и визуализировали таковое. Отдельно-интересно было проверить гипотезу, что в силу фундаментальных особенностей эволюционного процесса, параметры распределения последовательностей, в глобальном «пространстве» всех их возможных вариантов, остаются постоянными белковых семейств на протяжении миллиардов лет эволюции.



описание ДВИЖКА для визуализации Sequence Space

2 ИДЕИ для исследования пространства последовательностей - как реализованные, так и для будущего развития проекта

- Работа с данными происходит в среде Python 3 в инстанции Jupyter Notebook, который лежит на бесплатном сервере Kaggle
- Ноутбук считывает загруженные на сервер FASTA файлы выравнивания и выделяет из них сами выровненные последовательности, которые нужны для построения sequence space.
- Затем идет расчет матрицы относительных попарных расстояний Хемминга между всеми выделенными из выравнивания последовательностей.
- На основании этой матрицы считается среднее попарное расстояние, а также рассчитывается корреляционная размерность пространства этих последовательностей.
- Затем с помощью многомерного шкалирования изначальное многомерное sequence space этих последовательностей проецируется на трехмерное и затем визуализируется с помощью интерактивных scatter plots.

Ключевые модули:

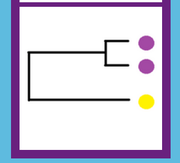
загрузка последовательностей и/или готовых выравниваний → “проектор” точек на 3D-пространство

3

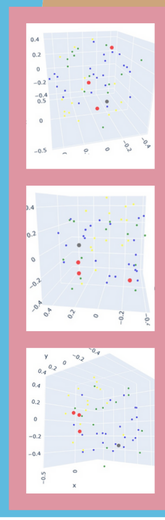
- сравнение топологий "натурального" набора точек (взятых из природы) и "стохастического" (сгенерированных случайным распределением мутаций)
- восстановление филогении (для конкретного семейства) и последующая визуализация Sequence Space'а с учётом распределения точек (последовательностей) на эволюционных ветках
- сравнение распределения точек для работающей и не-работающей копии ОДНОГО и ТОГО ЖЕ гена (т.к. тогда мы можем их наблюдать/изображать в одном "секторе" Sequence Space'а) - и, таким образом, оценка влияния отбора (и его отсутствия) на формирование топологии
- составление списков "экстремальных" видов (бактерий) с необычными свойствами - термофилы, быстро-эволюционирующие, и пр. - и написание АВТОМАТИЧЕСКОГО ПАРСЕРА, который мог бы сопоставлять названия видов из таких списков с названиями в большом датасете, чтобы "пометить особые точки"
- визуализация и анализ SequenceSpace'а генов, которые "олицетворяют" какой-то важный межвидовой переход... например: гены, которые были под сильным движущим отбор во время отделения человека и приматов (предполагается, что эти гены как-то связаны с формированием интеллекта)

4

Эту диаграмму получили проекцией упорядоченного по убыванию времени “последней модификации” белка, начиная от наиболее старых (желтый цвет) и заканчивая фиолетовым через весь градиент viridis.



филогенетическое построение: начиная с желтого, заканчивая фиолетовым (палитра viridis)



$$\sqrt{\frac{\sum(d_{ij}-\delta_{ij})^2}{\sum d_{ij}^2}}$$

Оценка качества проекции называется **стресс**.

Стресс определяется по формуле", где dij - изначальные попарные расстояния между точками, а δij - соответствующие попарные расстояния в пространстве новой размерности.

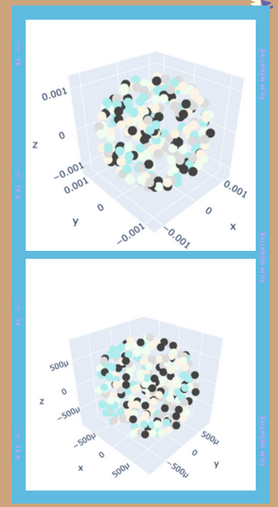
При плохом “стрессе” каждый раз получаются разные паттерны. Обратите внимание на красные точки, если хотите в этом убедиться.

(стресс примерно 0.4)

наборы стохастических (случайных) точек

5

Было решено поставить **эксперимент** (который подверг серьезному испытанию нашу веру в применимость используемых в проекте методик) с генерированием **случайных последовательностей аминокислот** на основе консенсуса и среднего расстояния от последовательностей, на основании которых он и был рассчитан как "средняя" строка, до консенсуса же. Удивительно было развенчать наши ожидания четырех концентрических (с центром в консенсусе) сфер рисунками, показанными выше.



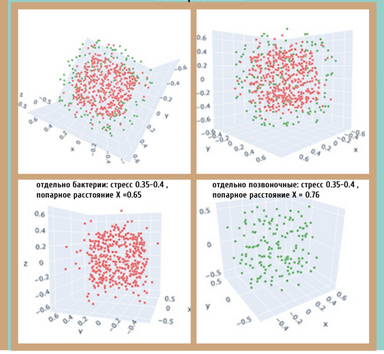
визуализация сиквенсов двух очень разных эволюционных групп - **позвоночных и бактерий**

зеленые - позвоночные ,
красные - бактерии

6

Семейство 1

стресс = 0.35-0.4, среднее попарное расстояние X = 0.8
Бактерии поместились внутри сферы, а позвоночные на ее оболочке в основном. Точек позвоночных чуть меньше, чем бактериальных



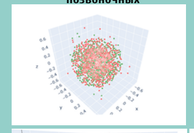
отдельно Бактерии: стресс 0.35-0.4 , попарное расстояние X =0.65

отдельно позвоночные: стресс 0.35-0.4 , попарное расстояние X = 0.76

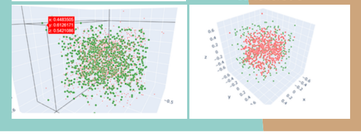
Мы выбрали белки, последовательности которых очень похожи И в бактериях, И в позвоночных.Таким образом, мы хотели посмотреть на Sequence Space для семейств, где точки (последовательности) прошли очень разный эволюционный отбор, и наверняка сформировали очень разные паттерны в пространстве последовательностей, НО сохранили заметное сходство **последовательностей = структур = функций.**

Семейство 2

Видна такая же закономерность при одинаковом количестве точек бактерий и позвоночных



Семейство 3



7 РЕЗУЛЬТАТЫ

- Создан удобный практичный инструмент анализа и визуализации пространства последовательностей
- Визуализированы пространства последовательностей для белковых семейств, последовательности которых сильно схожи для бактерий и позвоночных (а значит, имеют и единую эволюционную историю). Это позволяет "взглянуть" на SequenceSpace в ситуации, когда точки (гены) прошли очень разный эволюционным отбор (таковая разница очевидна для ситуации bacteria VS vertebrates)
- Неожиданными - с точки зрения интуитивных ожиданий - являются визуальные паттерны распределения множеств точек, отстоящих от заданного центра на различные дискретные “шаги”: можно было бы предположить, что таковые множества образуют сферы различного диаметра, вложенные друг в друга по принципу “матрёшки” - НО в действительности точки этих множеств перемешаны друг с другом, образуя некоторый единый сферический слой

ZIMIN FOUNDATION

SMTB

hhmi | Howard Hughes Medical Institute