Детальное описание финального проекта

Цели проекта:

- определить происхождение вируса SARS-CoV-2 человека;
- оценить, как выбор конкретных филогенетических методов влияет на конечный результат анализа.

Этапы анализа:

1. Найдите последовательности генов, кодирующих (1) белок S (spike), (2) любой другой белок (НЕ советуем брать ORF1a/b) у разных коронавирусов. Необходимо найти:

1.1) минимум 10 последовательностей SARS-CoV-2 человека, взятых в разные годы и/или в разных странах;

1.2) минимум 10 последовательностей для любых других вирусов рода Betacoronavirus, имеющих не менее 5 разных организмов-хозяев;

1.3) минимум по 2 различные последовательности вирусов MERS и SARS-CoV человека;

1.4) минимум 10 последовательностей вирусов рода Alphacoronavirus, имеющих не менее 5 разных организмов-хозяев.

2. Постройте множественные выравнивания найденных последовательностей (с помощью алгоритма MUSCLE или другого на ваш выбор). Сделайте выравнивание отдельно для ДНК-последовательностей и для транслированных аминокислотных последовательностей (создайте их до выравнивания). Выравнивания для разных генов/белков необходимо проводить отдельно! Всего должно получиться 4 выравнивания.

3. Проанализируйте полученные выравнивания и при необходимости обрежьте вручную или с помощью Gblocks/другого понравившегося вам алгоритма. Обрезка необходима, если послеледовательности сильно отличаются по длине, например, вы не везде нашли полные кодирующие последовательности. Если вы проводите обрезку выравниваний, используйте для дальнейшего анализа уже обрезанные (в работе остается 4 выравнивания).

4. Выберите подходящие модели эволюции нуклеотидов или частот замен аминокислот для полученных выравниваний. Бонус для желающих: в зависимости от результата подбора моделей решите, будете ли вы объединять выравнивания двух генов для дальнейшей филогенетической реконструкции.

5. Построение и укоренение филогенетических деревьев:

5.1) методом NJ (ДНК и белок, 4 дерева, 500 bootstrap),

5.2) методом ML (только для ДНК, 2 дерева, 100 bootstrap) с подобранной моделью эволюции нуклеотидов. Сравнените полученные результаты. Внимание: ML будет считать долго, ML с bootstrap в 100 раз дольше! Учтите это при планировании. Если по техническим причинам не

получится рассчитать bootstrap поддержку для ML дерева, то используйте меньшее число повторностей.

6. Анализ деревьев.

6.1) Сравните деревья полученные одним методом, но на основании разных наборов данных (нуклеотидная и аминокислотная последовательности ОДНОГО гена/белка, РАЗНЫЕ гены/белки). Укажите, есть ли между ними различия.

6.2) Сравните деревья, построенные на основании последовательностей ДНК разными методами (NJ, ML).

Отчет по заданию должен быть представлен в формате формате pdf и включать в себя:

1) перечень последовательностей, которые были взяты в анализ (обязательно нужно указать организм-хозяин и идентификаторы последовательностей в базе, идентификаторы достаточно привести в прилагаемом fasta файле);

2) описание и трактовку результатов промежуточных этапов анализа (необходимость обрезки выравнивания, подбор оптимальной модели эволюции и др.);

3) рисунки с дендрограммами и их описание и сравнение;

4) заключение (трактовка полученных результатов).

К отчету необходимо приложить наборы взятых в анализ выровненных последовательностей в формате fasta.

Дополнительные ссылки для тех, кто хочет повозиться с нюансами

Другие способы набрать вирусные последовательности:

- NextStrain https://nextstrain.org/ - база вирусных последовательностей NextStrain
- NCBI Virus portal https://www.ncbi.nlm.nih.gov/labs/virus/vssi/ - база вирусных последовательностей NCBI
- NCBI Viral genomes https://www.ncbi.nlm.nih.gov/genome/viruses/ - раздел геномной базы NCBI с полными геномами вирусов

Открытый сервер для филогенетического анализа Института Пастера/Университета Монпелье: http://phylogeny.lirmm.fr/ (старый), https://ngphylogeny.fr/ (новый)