Зачем биологам синтаксический анализ

Автор: Артём Горохов

Санкт-Петербургский государственный университет
Лаборатория языковых инструментов JetBrains

15 октября 2016г.

Биоинформатика

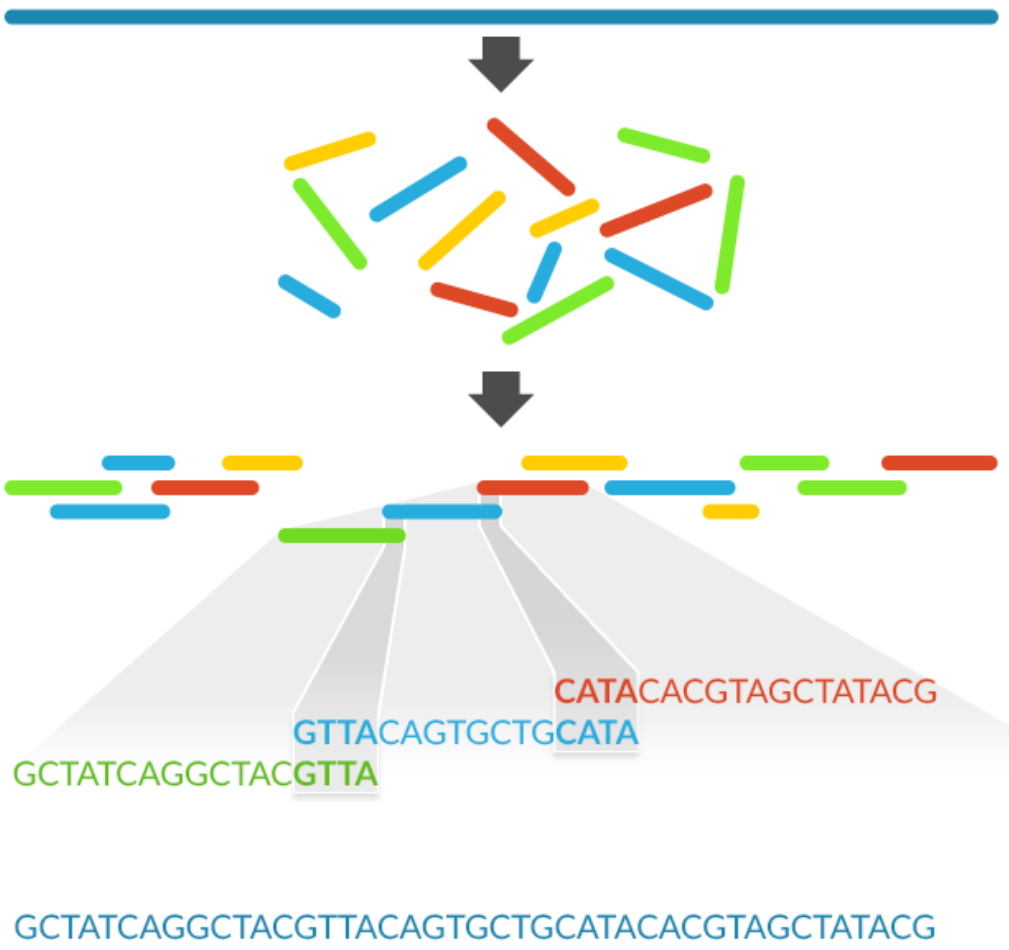
- Множество задач, связанных с обработкой и пониманием биологических данных
- Одна из задач — поиск организмов в метагеномных сборках

Геном

- Геном — длинная последовательность нуклеотидов
- На деле строка над алфавитом {A, C, G, U}

Получение данных

- Из биологического материала получаются последовательности строчек длиной около 100 символов
- Эти кусочки склеиваются в более длинные строки
- Множество строчек — сборка
- Данных очень много, так что строится граф, задающий множество полученных строк



Сборки

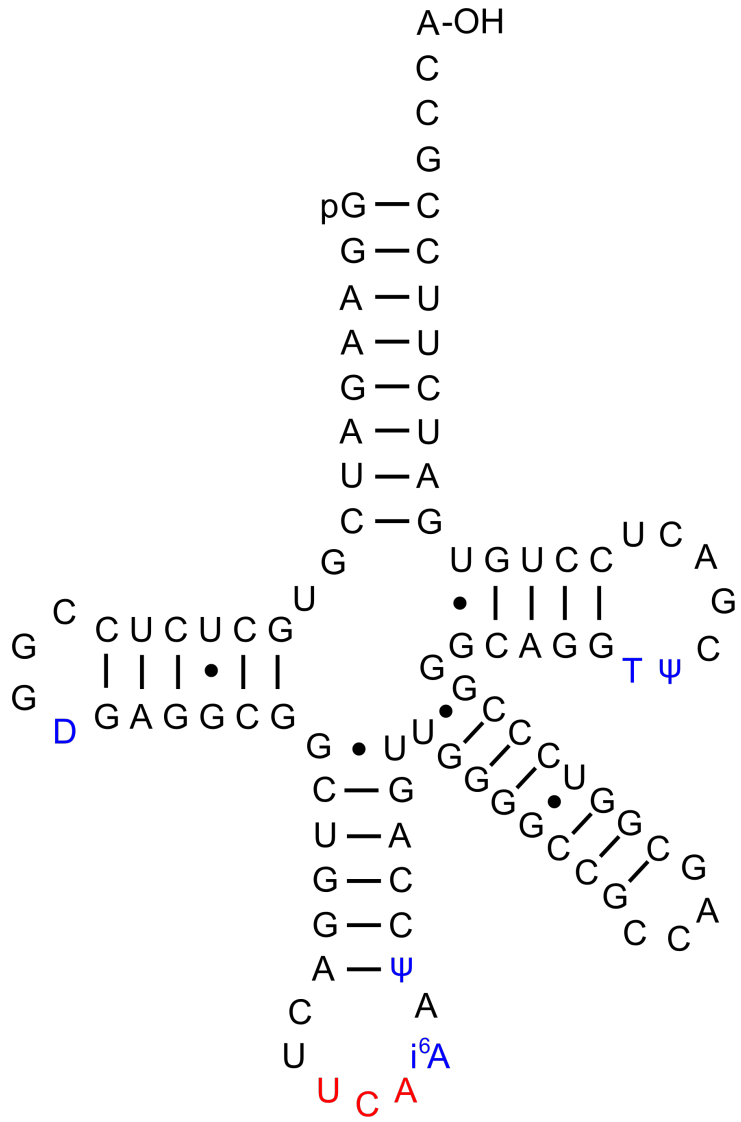
В зависимости от биологического материала, из которого получаются данные, сборки бывают:

- single-cell, multi-cell: клетки одного штамма
- **метагеномные**: данные взяты из среды обитания целевой колонии, в которой были как её представители, так и соседствующих

Что ищем

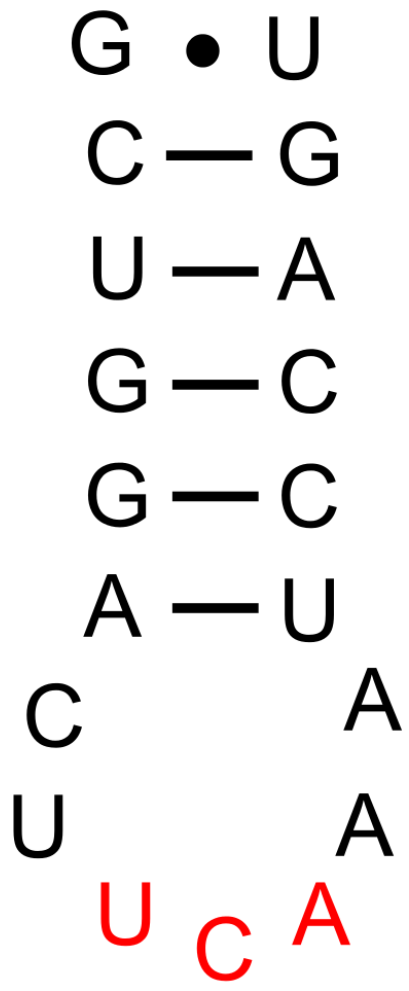
- Хочется понять что у нас в сборке
- Такие последовательности как тРНК, рРНК хорошо характеризуют организм которому они принадлежат
- У этих последовательностей есть вторичная структура, которая может быть описана КС-грамматикой

GGAAGAUCG...GCA... =>

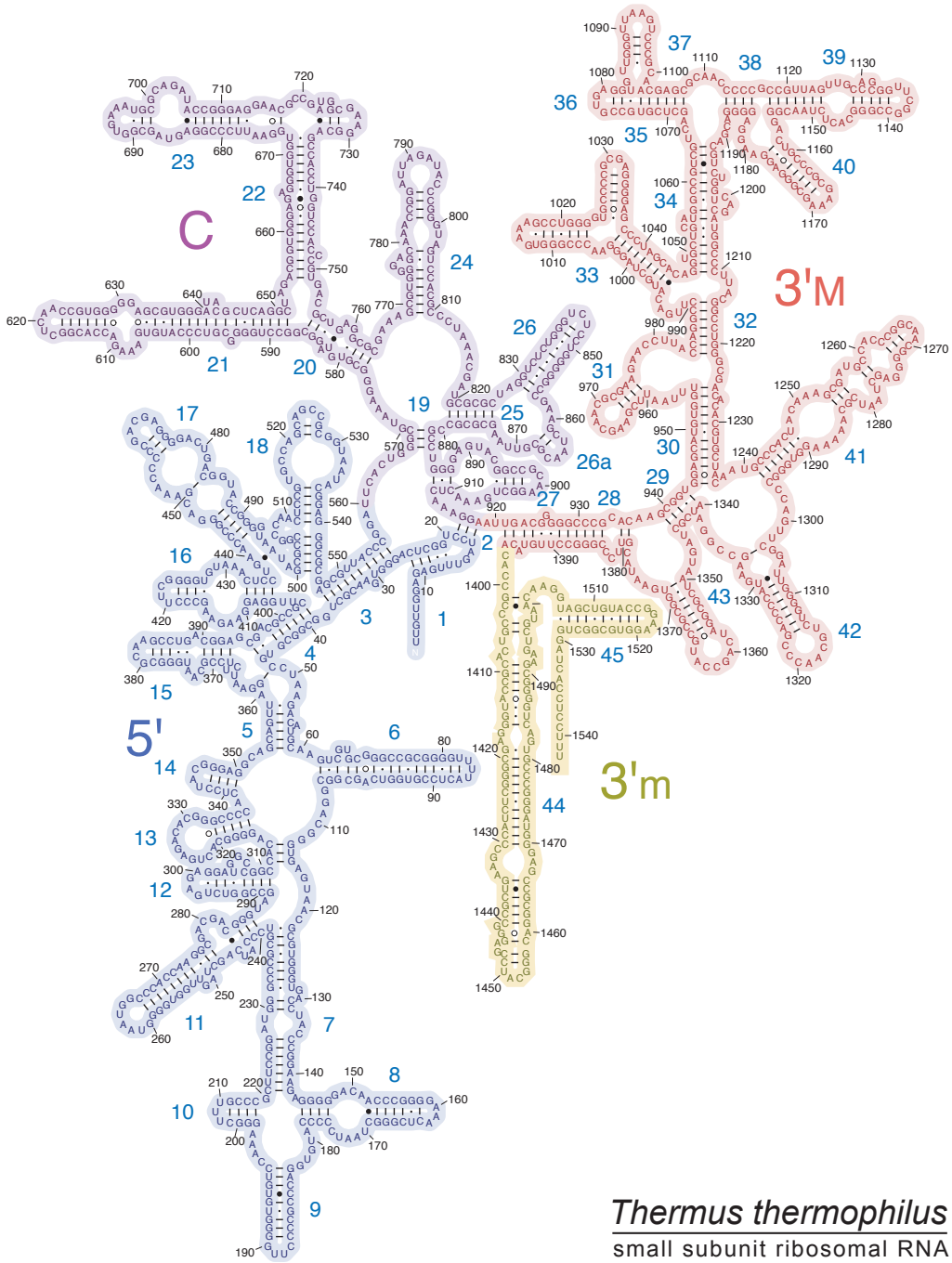


Грамматика для кусочка тРНК

$$
\begin{aligned}
START &= STEM \\
STEM &= a\ STEM\ u \\
&\mid u\ STEM\ a \\
&\mid c\ STEM\ g \\
&\mid g\ STEM\ c \\
&\mid g\ STEM\ u \\
&\mid u\ STEM\ g \\
&\mid ANY^*[4..7] \\
ANY &= a \mid u \mid g \mid c
\end{aligned}
$$



Вторичная структура 16s рРНК



Цель и задачи

Целью работы является с помощью синтаксического анализа находить в метагеномной сборке структуры, принадлежащие организмам
Задачи:

- Адаптировать существующий алгоритм под специфику задачи
- Провести экспериментальные исследования работы алгоритма

Graph parsing

- В лаборатории созданы алгоритмы
- Реализован инструмент, основанный на алгоритме GLL
- Умеет решать задачу поиска линейных цепочек в графе, удовлетворяющих KC-грамматике

GLL

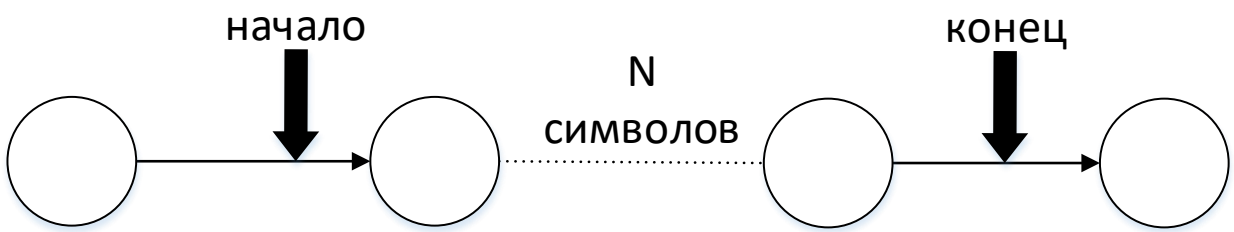
- Разбор осуществляется при помощи дескрипторов
- Дескриптор — четвёрка (слот, позиция во входе, дерево, вершина стека)
- На каждом шаге достаём дескриптор из очереди и разобрав очередной символ создаём новые дескрипторы

Особенности графа

- Полученные метагеномные сборки не поддаются анализу без предварительных преобразований
- Infernal позволяет распознавать структуры в линейном входе
- Рёбра, длиннее искомых структур можно делить на части и проверять с помощью Infernal
- После фильтрации рёбер граф распадается на компоненты связности, на которых алгоритм можно запускать анализатор независимо

Отказ от построения дерева

- Синтаксический анализатор возвращает лишь границы и длину найденных цепочек
- Восстановление цепочки идёт путём извлечения подграфа, состоящего из путей заданной длины
- Ложные фильтруются с помощью Infernal



Преобразование грамматики

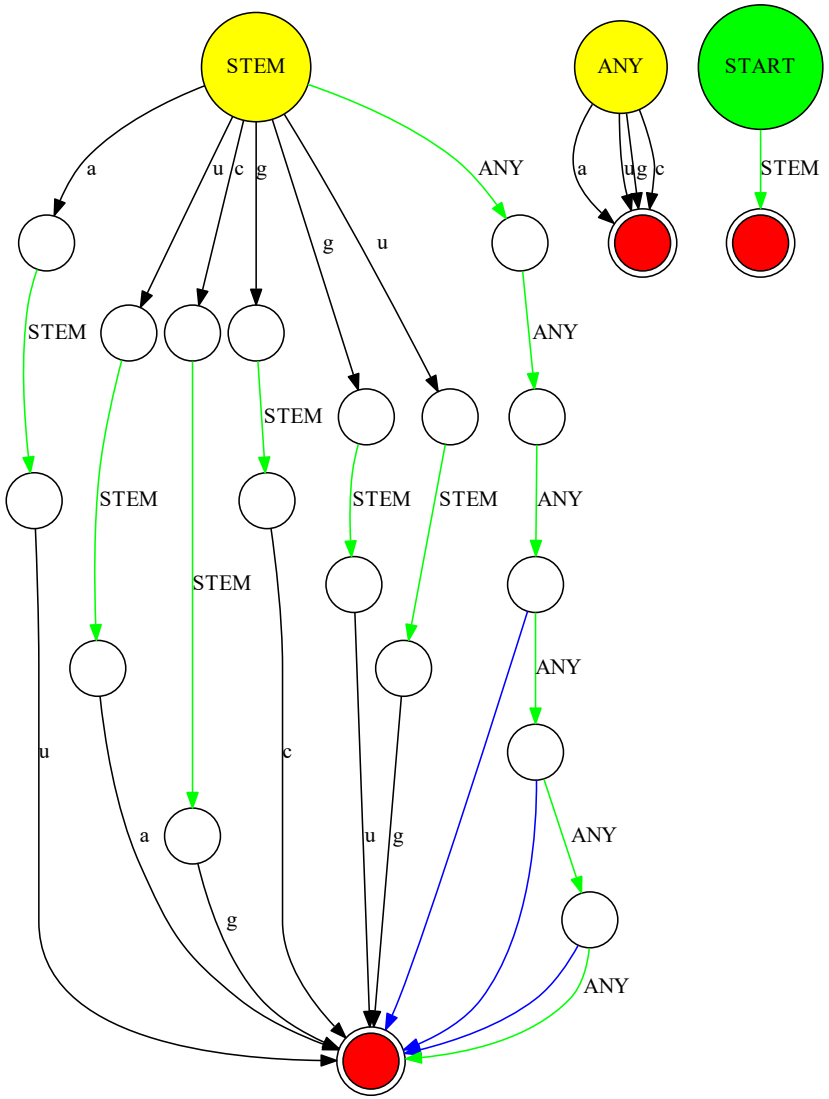
- Грамматика для 16s сильно неоднозначная и довольно большая
- Из-за этого количество слотов в грамматике избыточно
- В процессе разбора создаётся большое количество ненужных дескрипторов

Преобразование грамматики к автомату

Грамматика

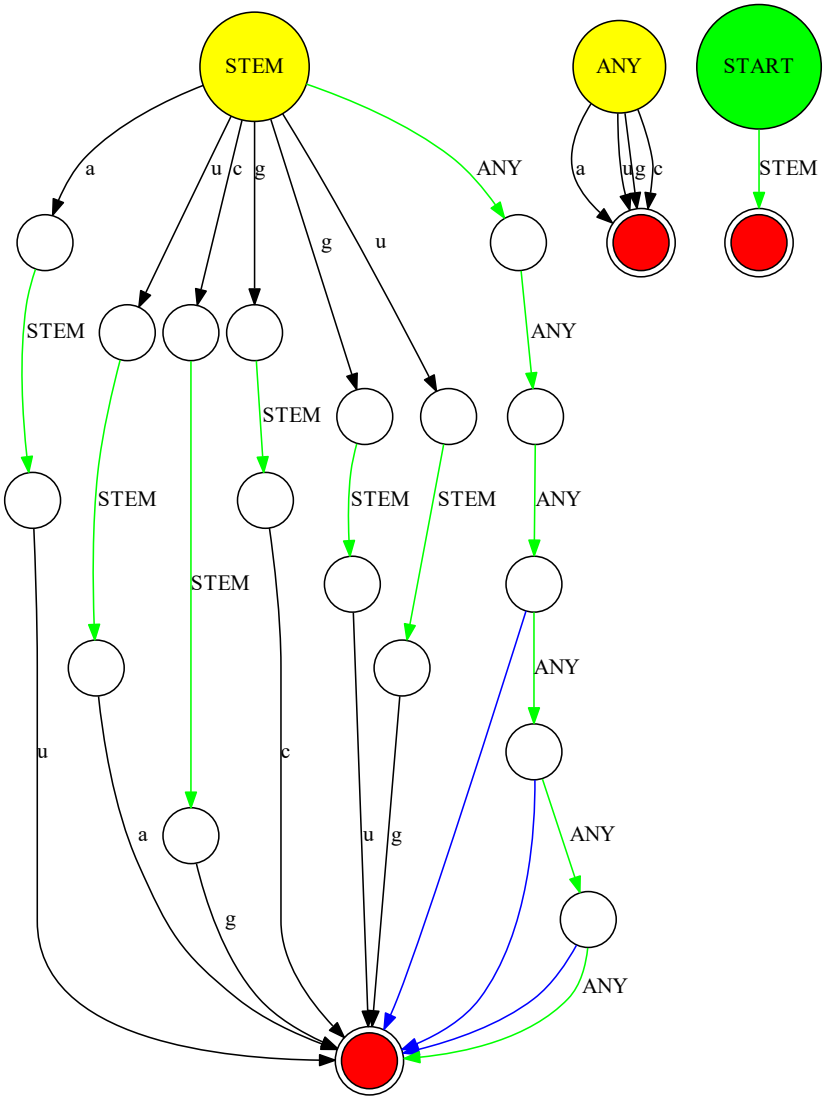
$$
\begin{aligned}
START &= STEM \\
STEM &= a\ STEM\ u \\
&\mid u\ STEM\ a \\
&\mid c\ STEM\ g \\
&\mid g\ STEM\ c \\
&\mid g\ STEM\ u \\
&\mid u\ STEM\ g \\
&\mid ANY^*[3..6] \\
ANY &= a \mid u \mid g \mid c
\end{aligned}
$$

Автомат

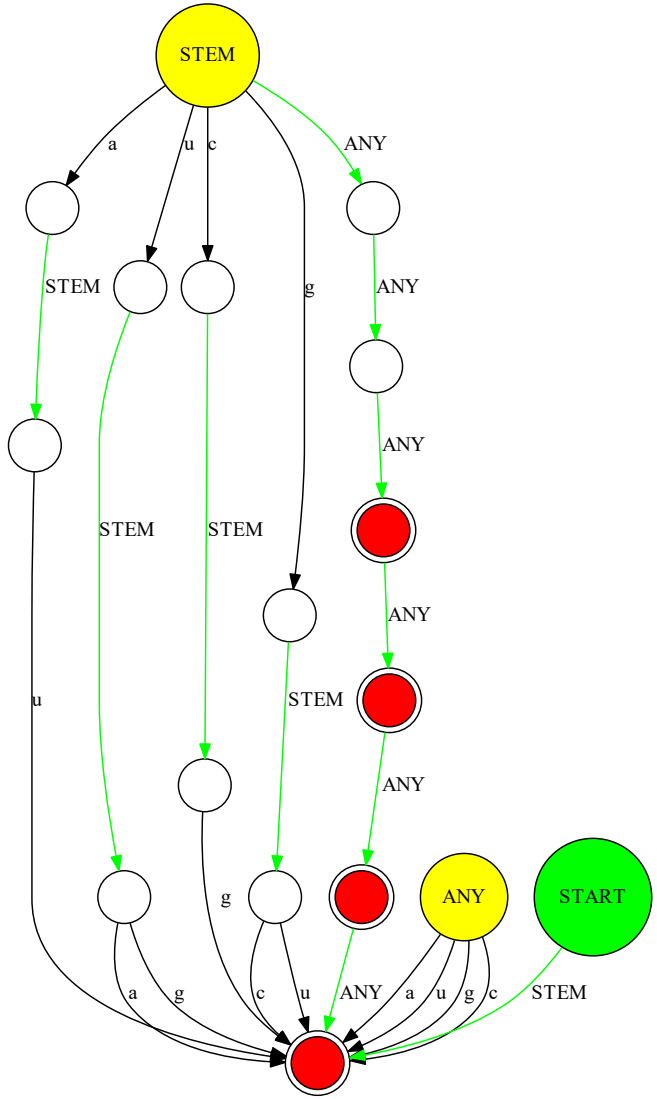


Минимизация автомата

Изначальный автомат



Минимизированый автомат



Эксперименты

Результаты работы на сборке, состоящей из 59000 вершин и 87000 рёбер

	начальная грамматика	мин. автомат
Время работы	10 ч.	3 ч. 40 мин.
Кол-во слотов /состояний	41	17

Эксперименты проводились на машине с 32 ГБ RAM и CPU core i7-4790

Результаты

- Разработан механизм подготовки сборок к синтаксическому анализу
- GLL адаптирован под распознавание по грамматике в форме EBNF
- Проведены эксперименты на части грамматики 16s PHK

Направление работ

- Детальный анализ качества результата
- Возможно, можно сильнее фильтровать граф, например применяя infernal
- Поиск полноразмерных 16s