Зачем биологам синтаксический анализ

Автор: Артём Горохов

Санкт-Петербургский государственный университет
Лаборатория языковых инструментов JetBrains

15 октября 2016г.

GLL

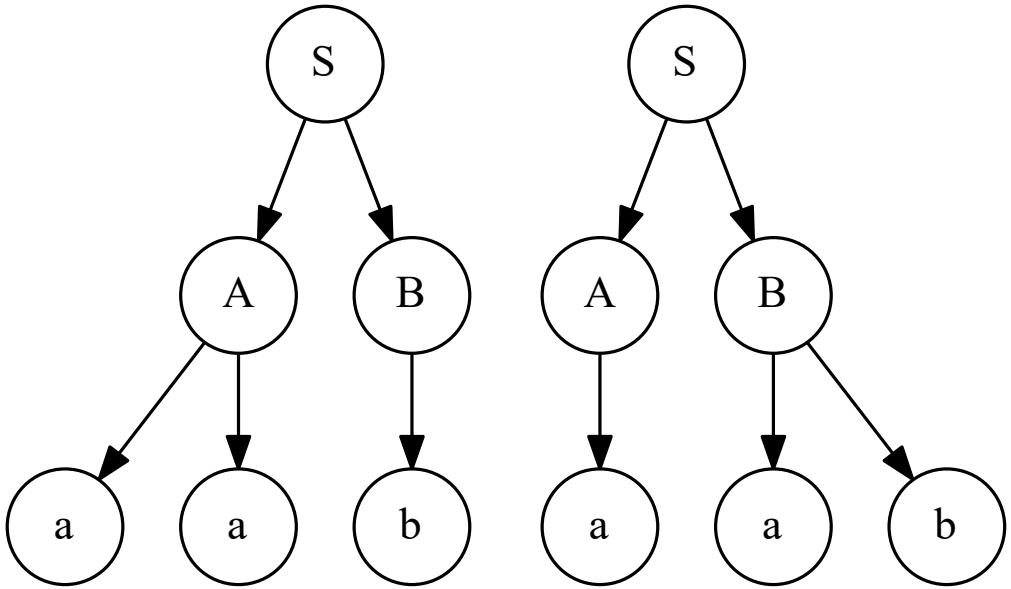
- Generalized LL
- Нисходящий синтаксический анализатор
- В лучшем случае работает за линейное время, в худшем - за $O(n^3)$
- Строит все возможные выводы цепочки

GLL

Вход: a a b

Грамматика:

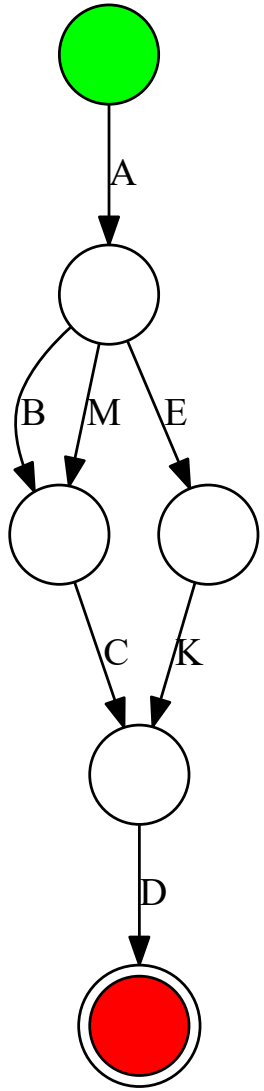
$$
\begin{aligned}
S &= A\ B \\
A &= a\ a \\
&\mid a \\
B &= b \\
&\mid a\ b
\end{aligned}
$$



GLL для графов

- На вход поступает граф, задающий все входные цепочки
- На рёбрах терминалы

$$\{ABCD; AMCD; AEKD\} =>$$



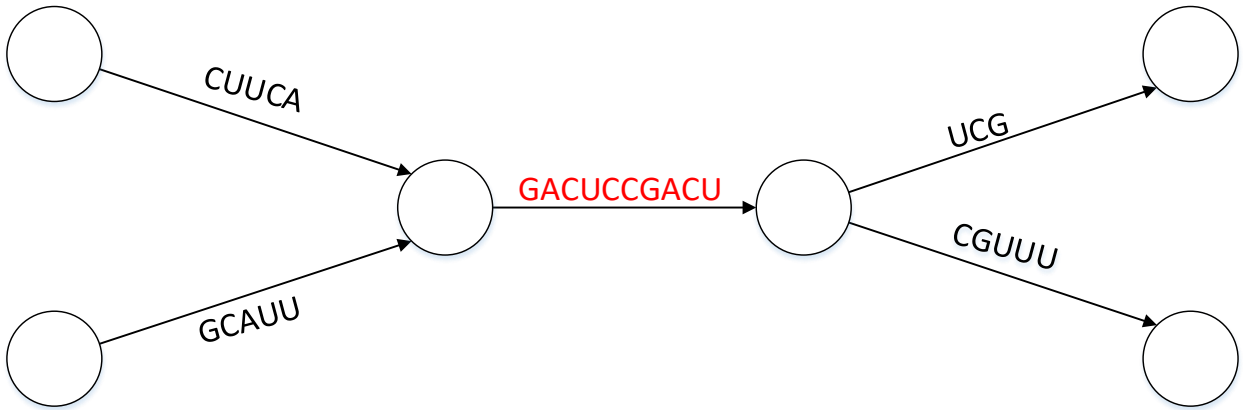
Метагеномная сборка

- Есть множество цепочек, подлежащих анализу
- Все объединяются в граф

CUUCAGACUCCGACUUCG

UCCGACUCGUUU

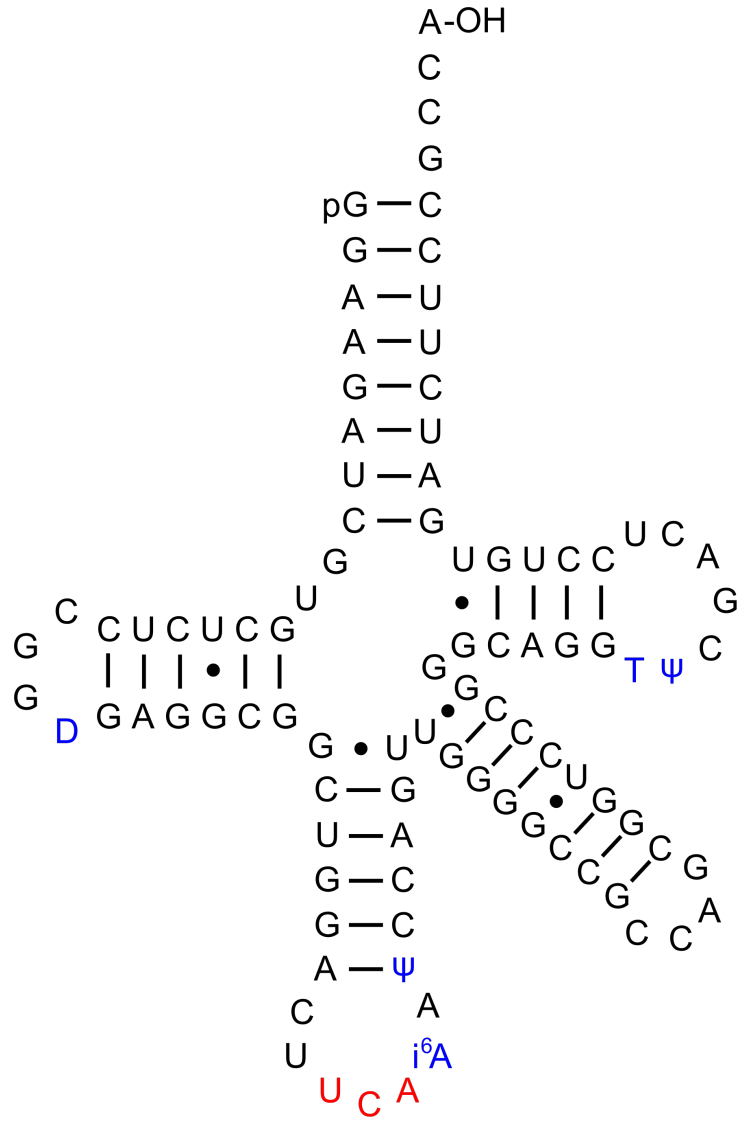
GCAUUGACUC

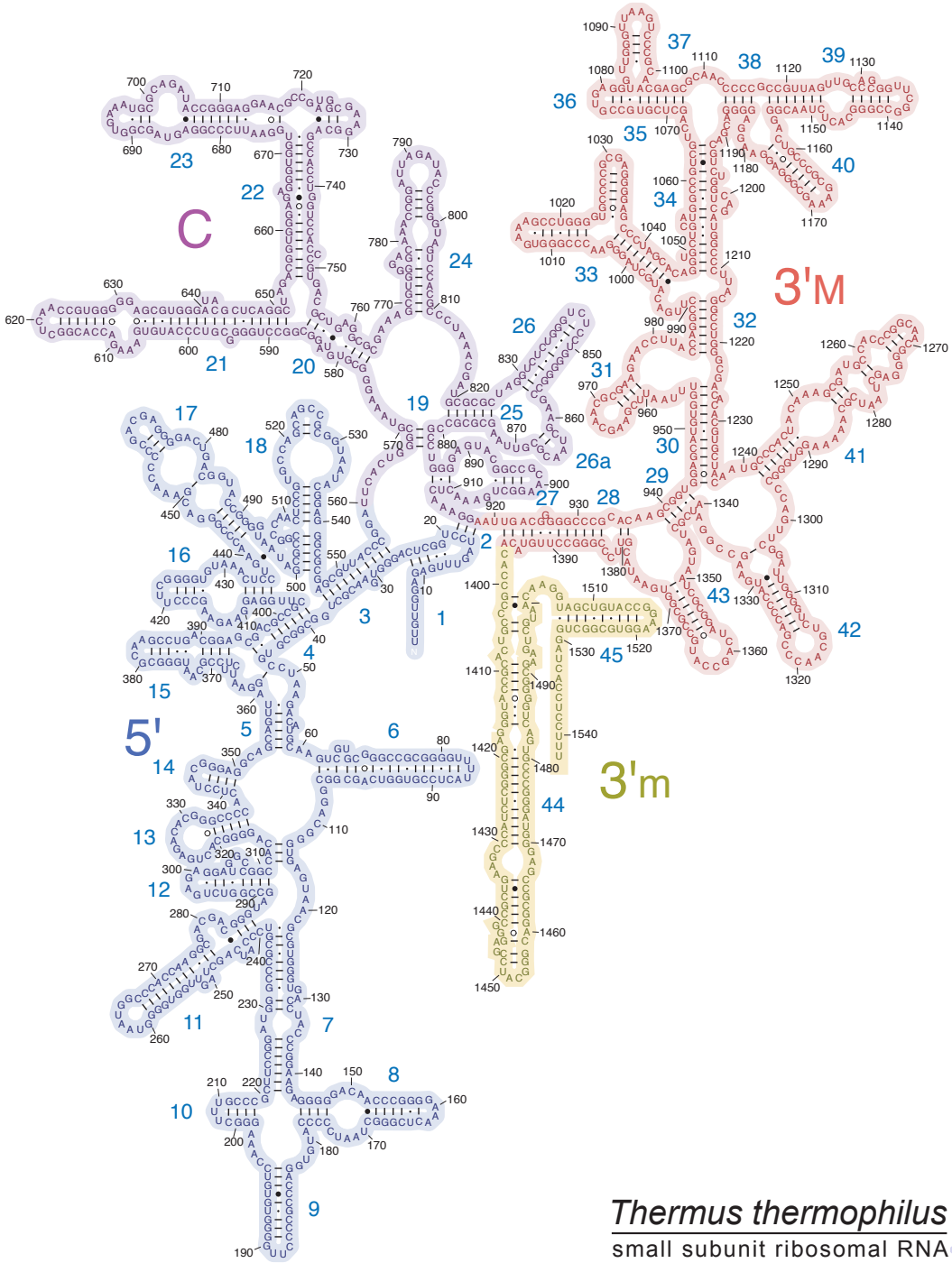


Структура цепочек

- KC грамматика может описать вторичную структуру

$$GGAAGAUCG...GCA... =>$$





Thermus thermophilus

small subunit ribosomal RNA

Увеличение производительности

- Полученные метагеномные сборки не поддаются анализу без предварительных преобразований
- Сам алгоритм нуждается в модернизации

Фильтрация рёбер

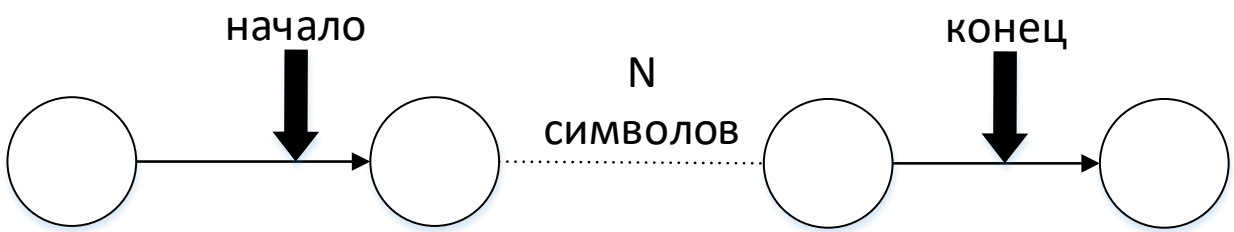
- Infernal позволяет распознавать структуры в линейном входе
- Рёбра, длиннее искомых структур можно делить на части и проверять infernal'ом

Разбиение на компоненты

- После фильтрации рёбер граф распадается на компоненты связности
- Можно запускать анализатор независимо на разных компонентах

Отказ от построения дерева

- Парсер возвращает лишь границы и длину найденой цепочки
- Восстановление цепочки идёт путём извлечения подграфа
- Ложные фильтруются infernal

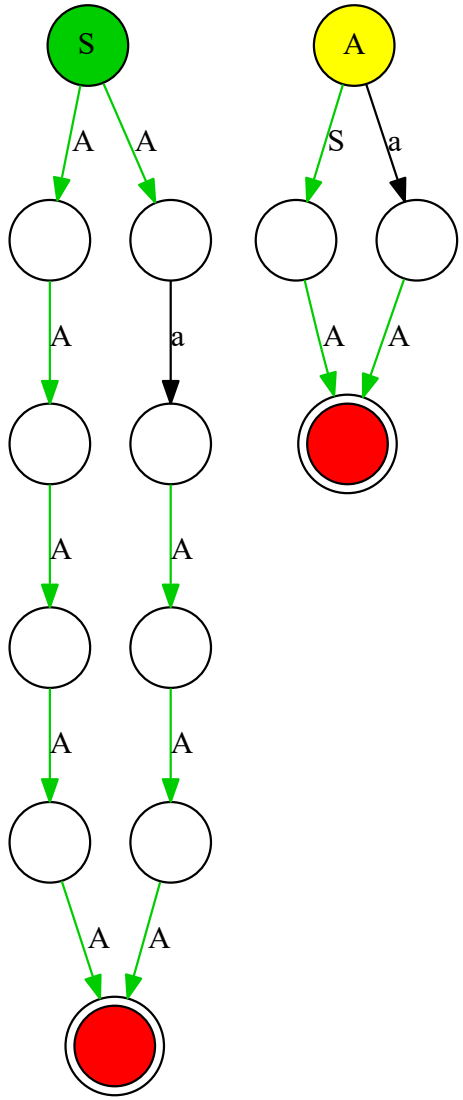


Преобразование грамматики к автомату

Грамматика

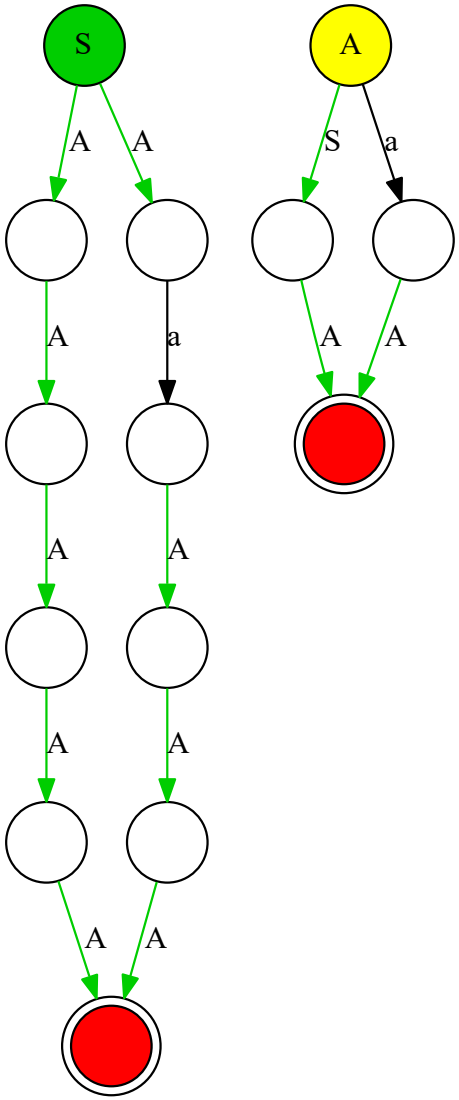
$$
\begin{aligned}
S &= A\ A\ A\ A\ A \\
&\mid A\ a\ A\ A\ A \\
A &= S\ A \\
&\mid a\ A \\
&\mid a
\end{aligned}
$$

Автомат

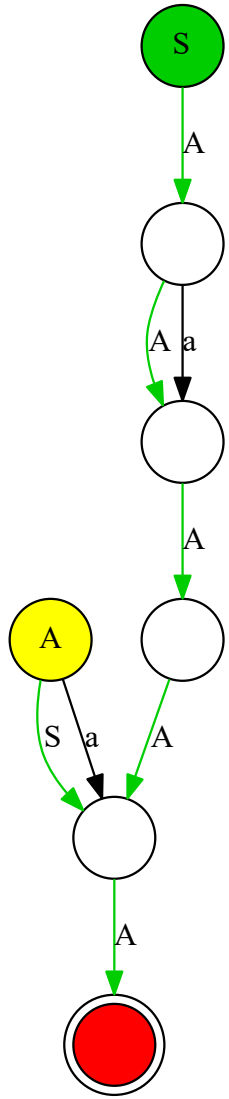


Минимизация автомата

Изначальный автомат



Минимизированый автомат



Эксперименты

	начальная грамматика	мин. автомат
Время работы	10 часов	3ч. 40 мин.

Направление работ

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