УЛУЧШЕНИЕ БИНИНГА КОНТИГОВ В МЕТАГЕНОМНЫХ СБОРКАХ С ИСПОЛЬЗОВАНИЕМ ГРАФА СБОРКИ

Иван Дмитриевский, рук. Сергей Нурк — мнс лаборатории ЦАБ СПБГУ

11 декабря 2015 г.

Институт биоинформатики

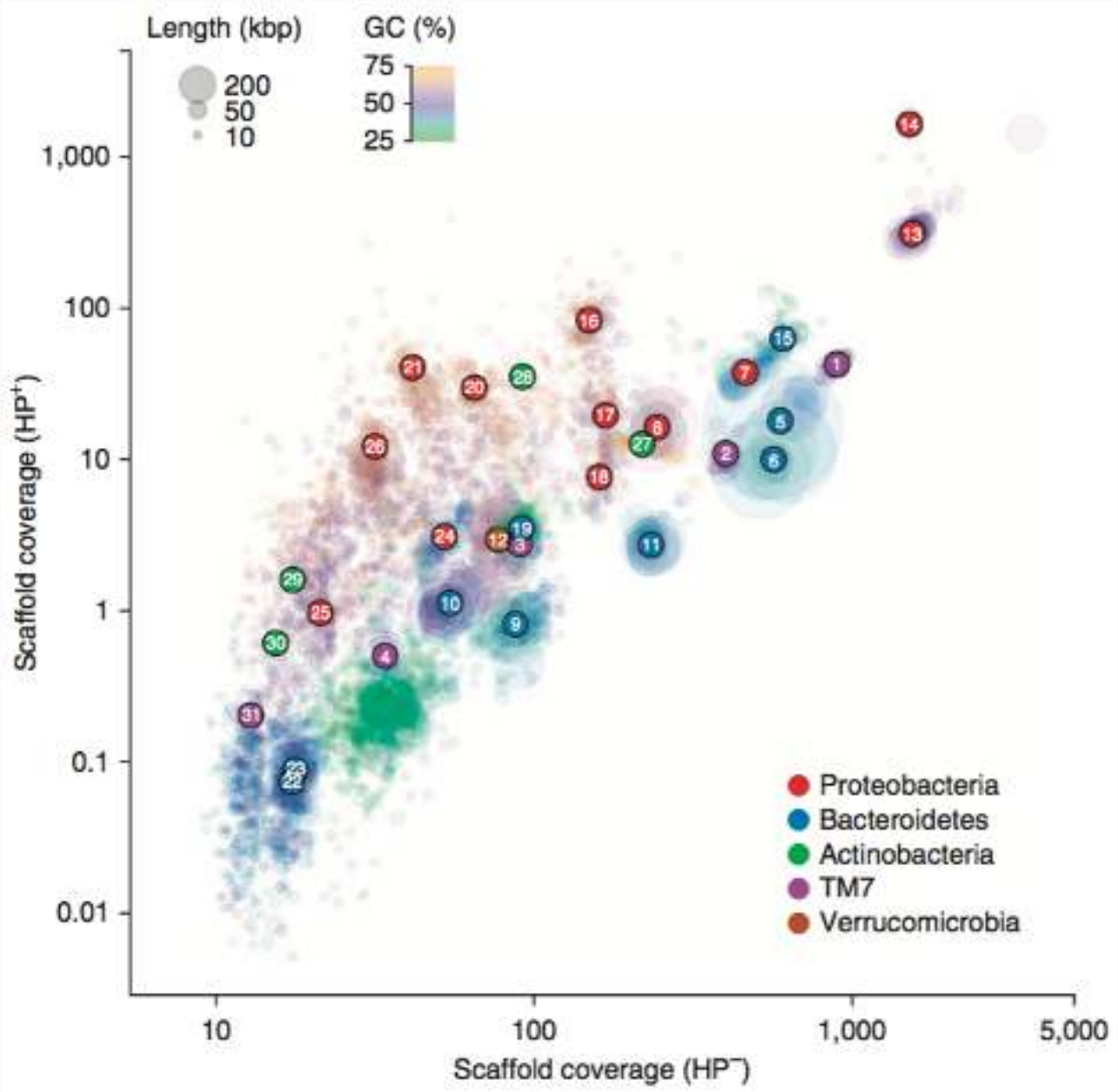
БИНИНГ

В метагеноме организмы представлены в различных пропорциях.

Контиги можно кластеризовать по среднему покрытию.

В случае, если дано несколько образцов (метагеномный ряд), то кластеризовать можно по вектору среднего покрытия.

ИЛЛЮСТРАЦИЯ ДЛЯ ДВУХ ОБРАЗЦОВ



ПРОБЛЕМЫ

- консервативные участки, встречающиеся в разных организмах
- контиги, специфичные для конкретных штаммов, могут оказаться в отдельных кластерах
- обычно из анализа исключаются короткие контиги

ЗАДАЧА

Дано:

- сжатый граф сборки
- разбиение подмножества контигов на кластеры

Требуется разработать стратегию улучшения бининга с использованием графа сборки (различные штаммы должны попадать в один кластер).

Основная идея

Для простоты изложения контиги — это рёбра.

Рис. 1: Предварительный бининг

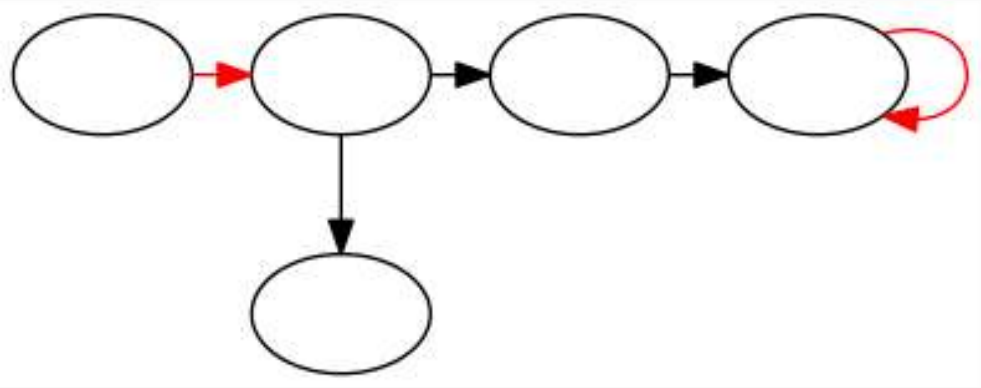
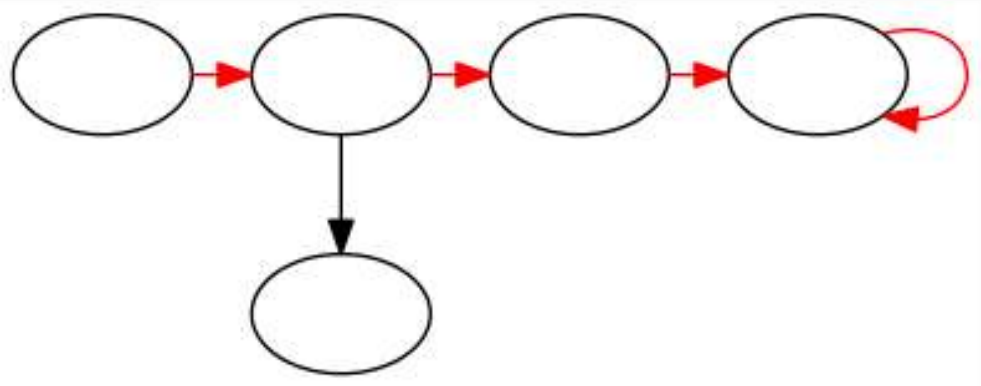


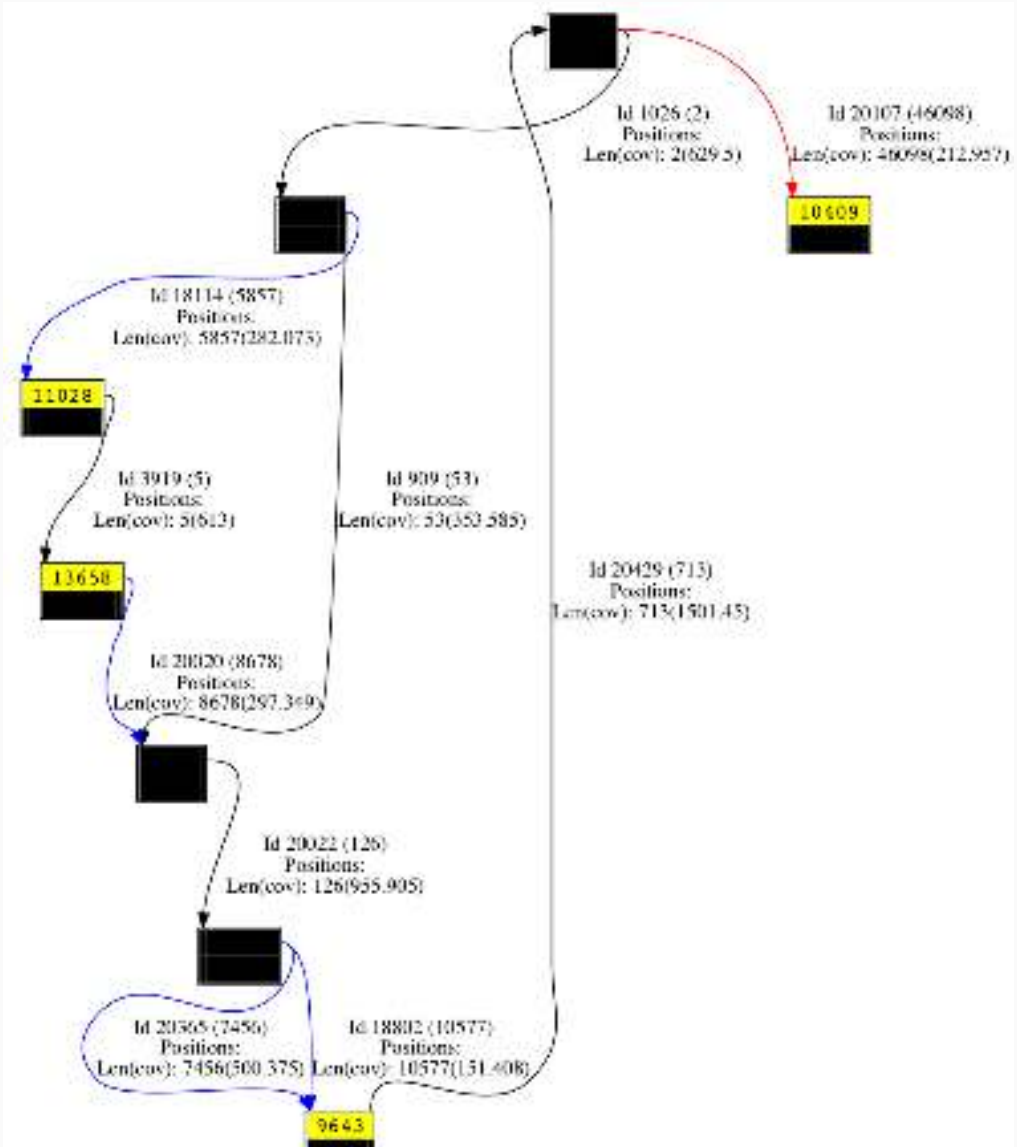
Рис. 2: Результат работы алгоритма



ОПИСАНИЕ АЛГОРИТМА

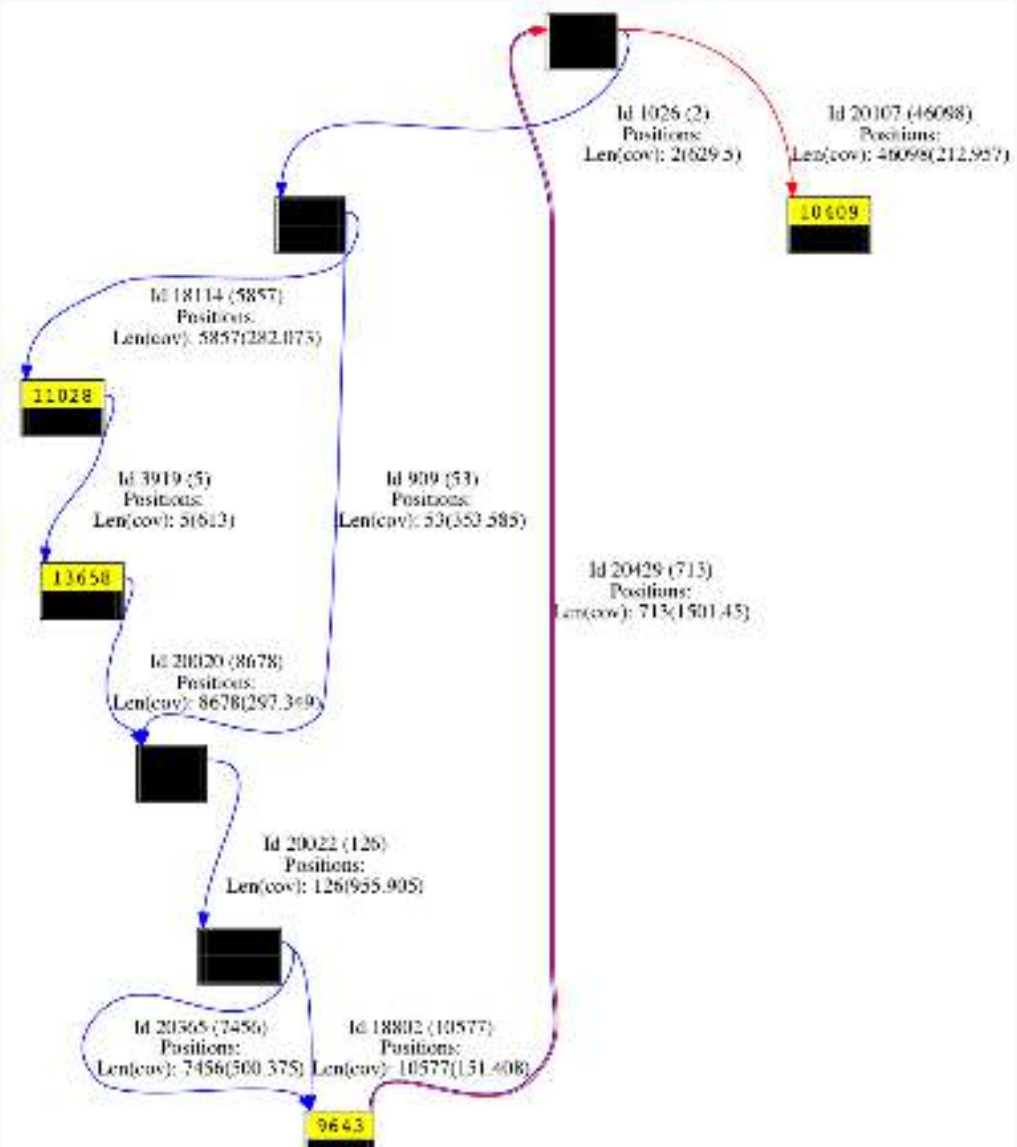
- рассматривается каждый кластер K из множества кластеров
- для каждого $(u, v) \in K$ из конца v запускается ограниченный алгоритм Дейкстры
- если была посещена вершина x такая, что $(x, y) \in K$, то находятся все рёберные пути $\langle v, x \rangle$ короче 5kb, и их рёбра добавляются в K (распространение аннотаций).

Рис. 3: Предварительный бининг



ПРИМЕР РАБОТЫ

Рис. 4: Результат работы алгоритма



ОЦЕНКА КАЧЕСТВА

- известные геномы связываются с найденными кластерами
- в графе находится рёберный путь P_g, соответствующий геному

Проходом по P_g подсчитывается количество

- $e \in P_g$, которые не были кластеризованы
- $e \in P_g$, которые были кластеризованы в результате распространения аннотаций

РЕЗУЛЬТАТЫ

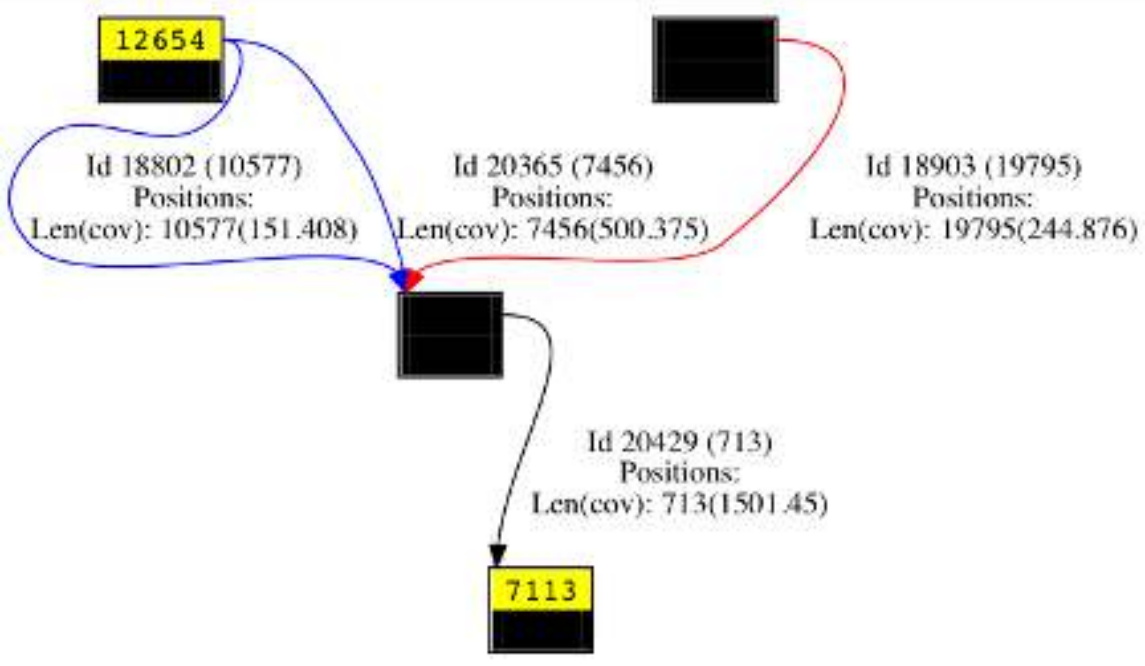
Для получения осмысленных оценок качества требуется запустить алгоритм на различных метагеномах и сравнить, что получится.

ПЛАНЫ НА БУДУЩЕЕ

- тестирование на симулированных метагеномах
- определение недостатков описанной стратегии на основе результатов тестирования
- реализация дополнительных стратегий распространения аннотаций
- тестирование на реальных данных (иногда известны отдельные компоненты метагенома)

Спасибо за внимание

Рис. 5: Предварительный бининг



ПРИМЕР РАБОТЫ

Рис. 6: Результат работы алгоритма