Дворкин М. Э.

Методы минимизации необходимого числа цепей для секвенирования ДНК

Научный руководитель:
Корнеев Г. А., к. т. н., доцент КТ СПбГУ ИТМО

Задача секвенирования ДНК

- ДНК содержит всю наследственную информацию организмов-эукариотов
- Двойная спираль из нуклеотидов: {A, C, G, T}
- Комплементарные пары: A—T, C—G
- Миллионы нуклеотидов (max≈2.2×10^8)
- Необходимо для диагностики, биотехнологий, судебной медицины, биологической систематики.

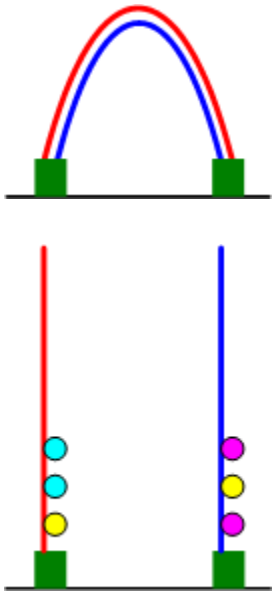
© www.pacificu.edu

Задача секвенирования ДНК (2)

- Последовательное чтение участка ДНК
  - ≤ 7000 пар оснований
- «Метод дробовика»
  - ДНК разбивается на участки (10—100 пар оснований)
  - Каждый участок читается последовательно
- Возможны ошибки, про каждый нуклеотид известно phred-качество: $Q = -10 \cdot \log_{10} P_{ошибки}$

Специфика метода Solexa

- Участки длиной ≈200 пар оснований крепятся адаптерами, образуя «мосты»
- «Мосты» расщепляются, оставаясь прикрепленными одним концом
- Нуклеотиды каждой половины считываются по одному с помощью флуоресцентных маркеров — азотистых оснований



Специфика метода Solexa (2)

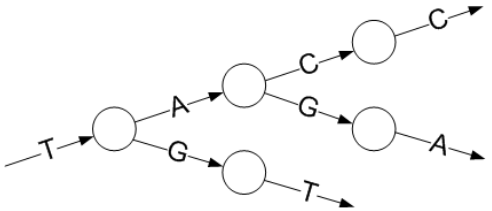
- Пары цепей длиной 36 нуклеотидов
- Пара (*s*, *t*) => в искомой строке встречается либо *s*??..??$\overleftarrow{t}$, либо *t*??..??$\overleftarrow{s}$.
  - $\overleftarrow{s}$ означает строку, обратно-комплементарную к *s*
- Расстояние между цепями в искомой строке распределено по Гауссу
- Стандартное отклонение < длины цепи
  - Выполняется во многих, но не во всех биологических экспериментах

Схема предлагаемого алгоритма

1. Построение графа возможных продолжений G_c
2. Вычисление весов ребер и дополнительной информации в вершинах и ребрах графа G_c
3. Обнаружение ошибочно считанных нуклеотидов во входных данных и их исправление;
4. Наращивание подстрок до достижения развилок и тупиков;
5. Построение графа развилок G_b и создание в нем кратных ребер;
6. Нахождение эйлерова пути в полученном мультиграфе.

1. Граф возможных продолжений

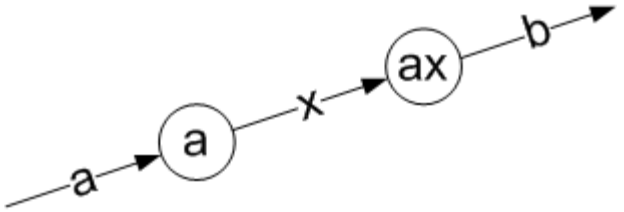
- Граф возможных продолжений G_c — аналог суффиксного дерева для множества цепей и строк, обратно-комплементарных к ним
- Строки длины $< \log_4|G|$ неинтересны, рассмотрим вершины слоя номер t и ниже.



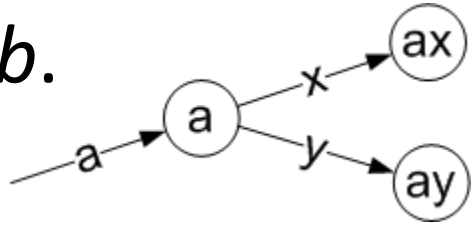
- Граф G_c распадается на части, в которых $\approx(|G|+\epsilon\cdot n)\cdot L/4^t$ вершин
  - умещается в оперативную память
- Можно строить части параллельно.

2. Пометки на ребрах графа G_c

- Цепь *axb* добавлет к ребру, соответствующему *x*, вес $P_{\text{ошибки в } a} \cdot P_{\text{ошибки в } x}$



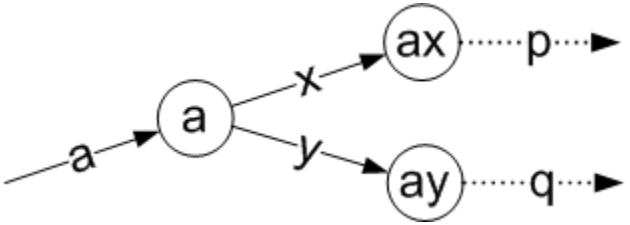
- Также хранится цепь, парная к *axb*.
- Из вершины, соответствующей a, известны (в первом приближении) вероятности продолжений по {A, C, G, T}



- Вес меньше порогового — ошибка в данных

3. Исправление ошибок в данных

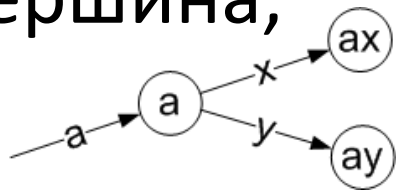
- В графе G_c для каждой вершины найдем наиболее вероятное продолжение (динамическое программирование на дереве)



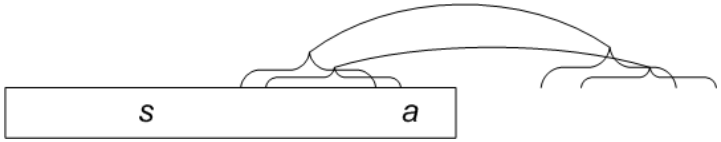
- Ребро *y* с малым весом попробуем заменить тремя другими вариантами
- Если его продолжение *q* «хорошо укладывается» в поддерево, соответствующее *x*, то *y* — ошибочный нуклеотид, соответствующие цепи следует исправить.

4. Наращивание подстроки (1)

- Текущее состояние: подстрока *sa*, вершина, соответствующая *a* в графе G_c.
  - Инициализация — любая цепь с max качеством



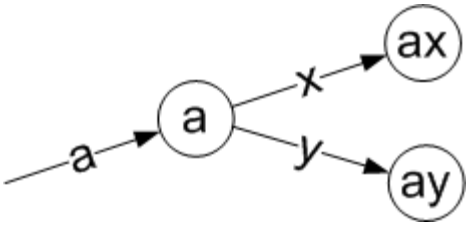
- Также набор парных цепей, ожидаемых «в будущем»

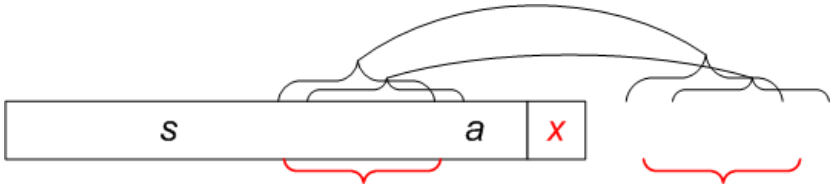


- Если все ребра из *a* имеют малый вес, перейдем от *a* к суффиксу *a*.
- Если ровно одно ребро с большим весом, ОК

4. Наращивание подстроки (2)

- Если ребер с большим весом > 1, соответствующие парные цепи «прикладываются» к строке *sa* и к «будущим» цепям

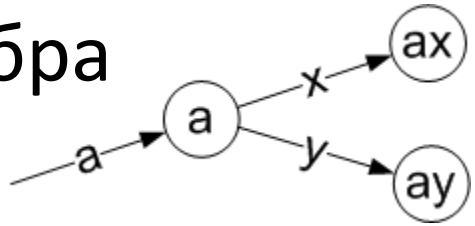




- Наличие пересечений (малое расстояние Хэмминга, с учетом сдвига) повышает вероятность перехода по символу *x*.

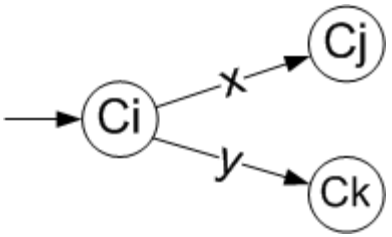
4. Наращивание подстроки (3)

- Если подходящих ребер все еще > 1, «запоминаем» развилку B_{sa}, перебираем (рекурсивно) все подходящие ребра
- Для развилки запоминаем веса ребер, они характеризуют частоту подстрок *ax* и *ay* в искомой строке.



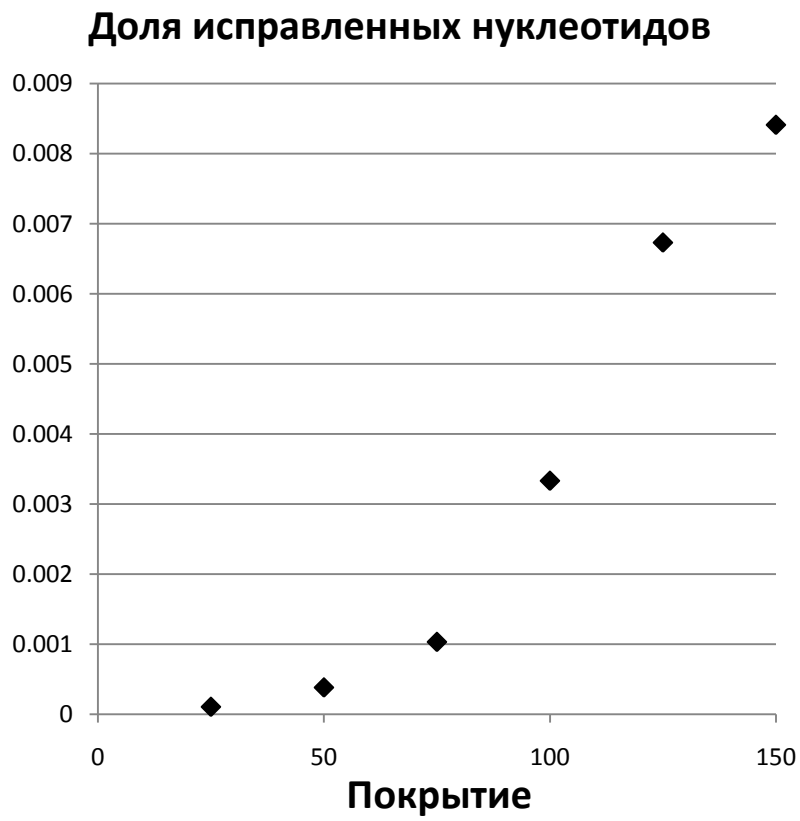
5. Граф развилок G_b

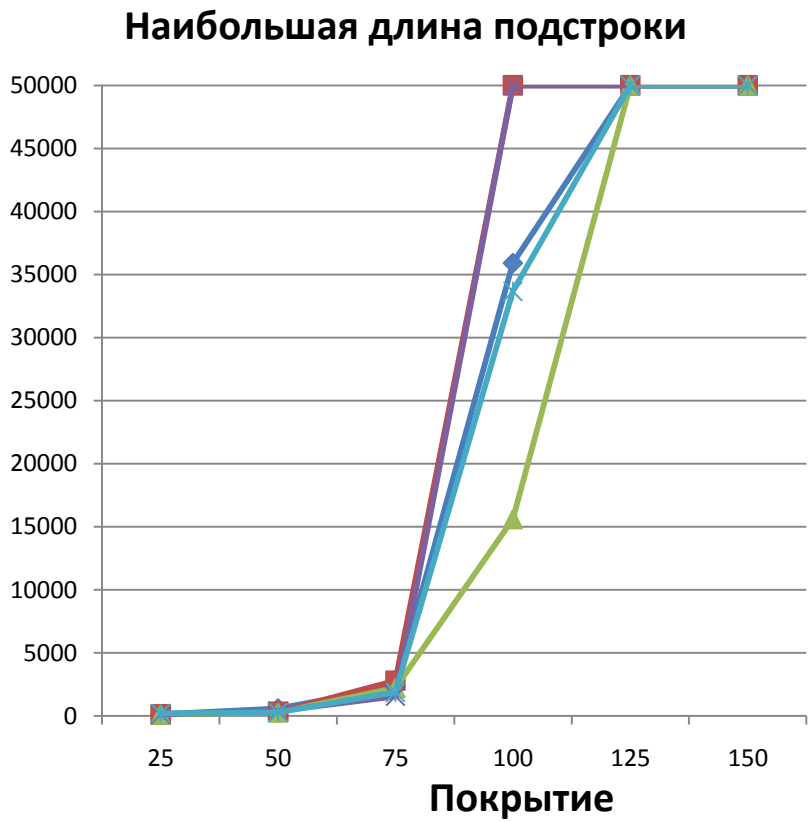
- Искомая строка — обход графа G_b
- Вес ребра —среднее число цепей, покрывающих один нуклеотид
- Для каждой развилки B_i известны веса соответствующих ребер в G_b
- Отношение весов соответствует частоте выбора ребра
- Веса ребер нормируются и округляются
- В полученном графе с кратными ребрами ищется эйлеров путь



Бактерия Escherichia coli K-12

Результаты работы программы:





Выводы

- Предложен алгоритм секвенирования ДНК организмов-эукариотов
- Предложены эффективные методы обнаружения и исправления ошибок во входных данных
- Этапы алгоритма поддаются параллелизации

Спасибо за внимание!