ФКН ПИ

Курсовая работа

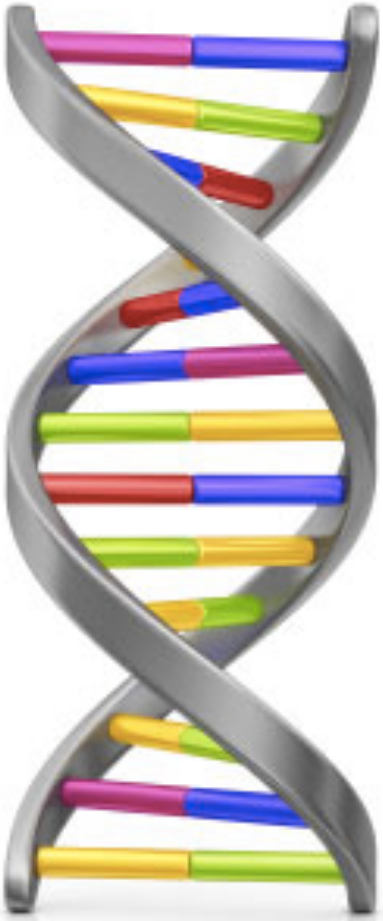
Исследование алгоритмов реконструкции слов в целях восстановления кода ДНК

Выполнил:
студент группы 202ПИ
Дробинин В. Д.

Научный руководитель:
д. т. н.
Ульянов М. В.

Введение

Объект исследования



…ATTGTACCTAT…

Актуальность темы

- 1988 год — проект «Геном человека»
  - 3 000 000 000 $ на секвенирование *
- 2001 год — черновые результаты
  - 6 000 000 000 $ расходов *
- 2014 год — «Генотек»
  - 375 000 руб. за анализ ДНК **

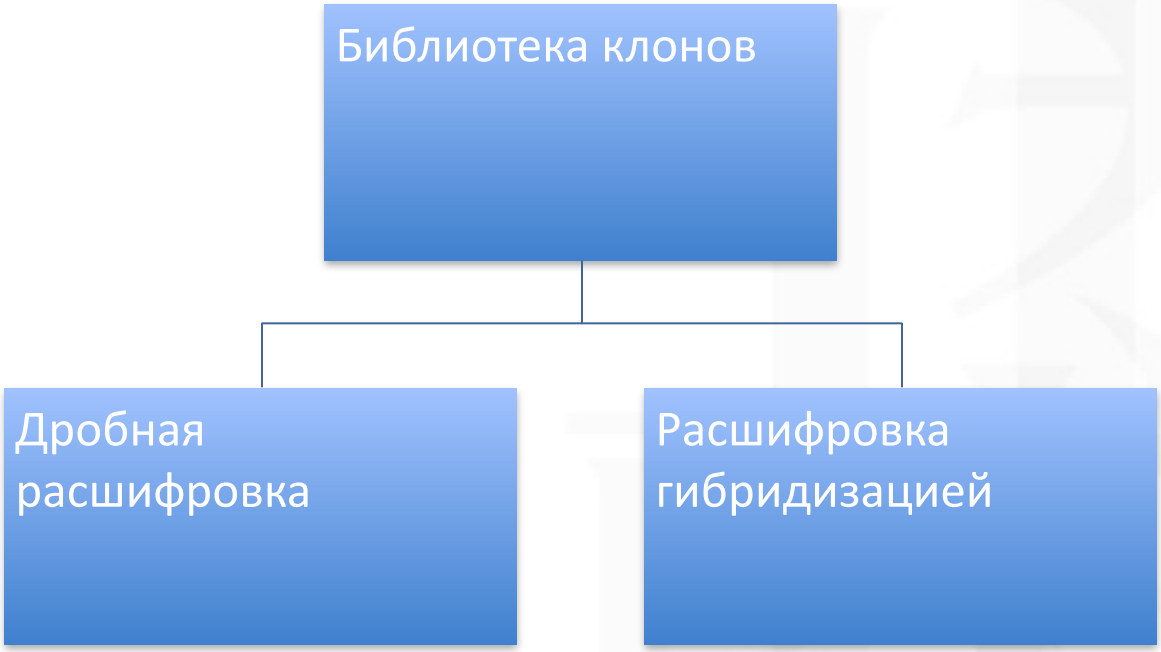
* http://mygenome.su/articles/89/

** https://www.genotek.ru/

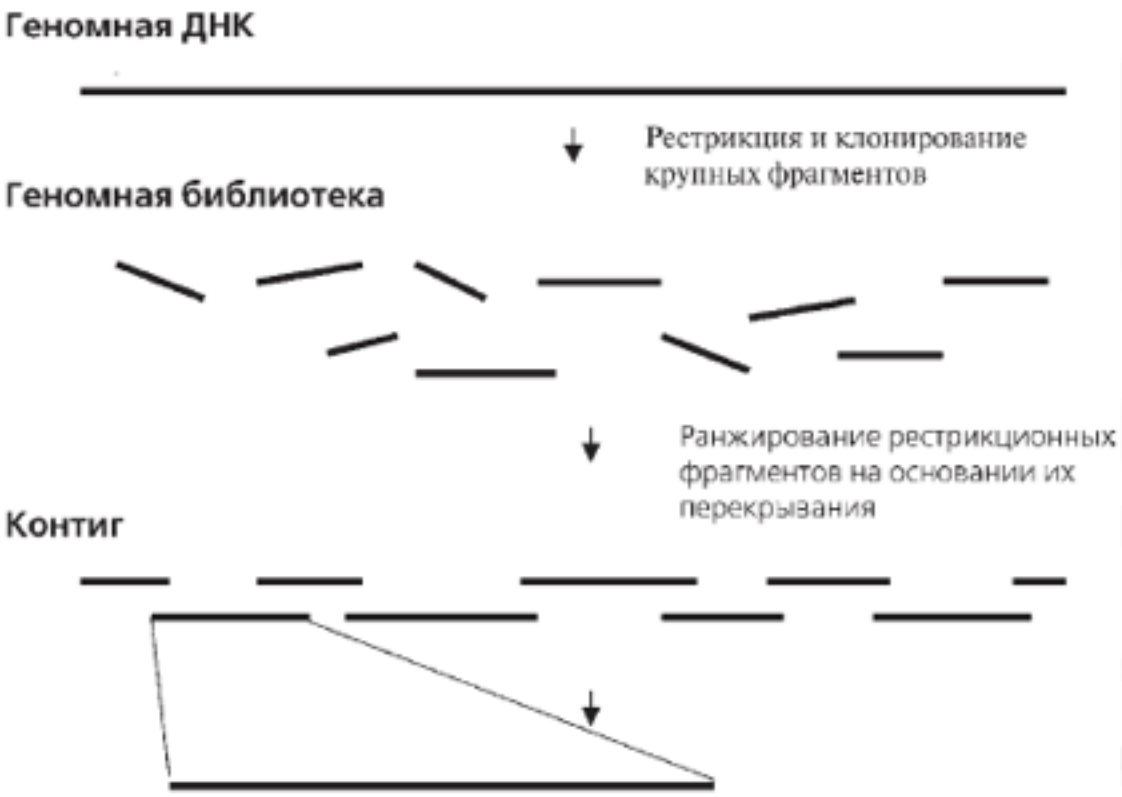
Цели и задачи

- Анализ литературы;
- Достоинства и недостатки алгоритмов;
- Проверка корректности;
- Эксперименты;
- Способы оптимизации.

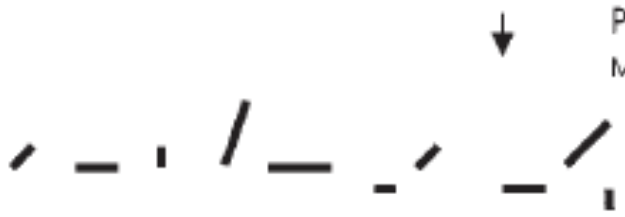
Методы восстановления ДНК



Дробная расшифровка



Дробная расшифровка



Секвенирование

Последовательности

фрагментов

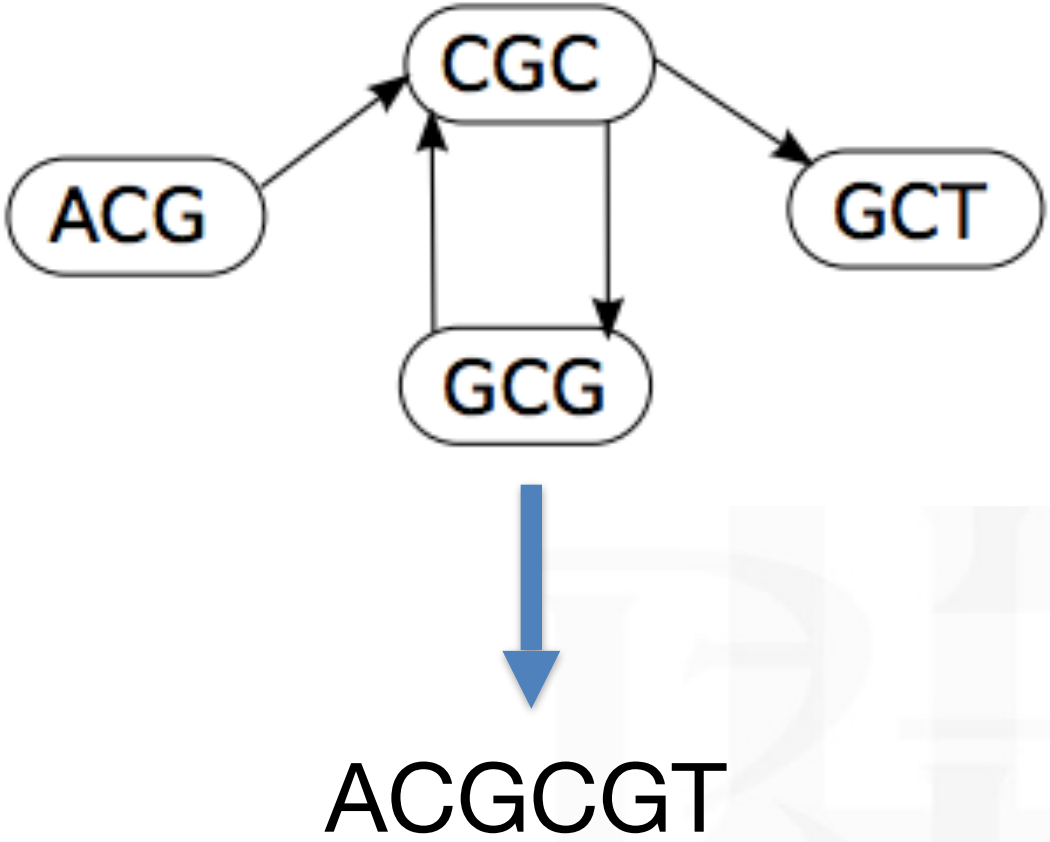
.....AATГГЦАЦЦГТААГГГГТЦЦГЦЦАТААЦЦГТТГЦ

ТАТТГГЦААЦГААТТААЦГГАЦГГАТ.....

Результирующая нуклеотидная последовательность

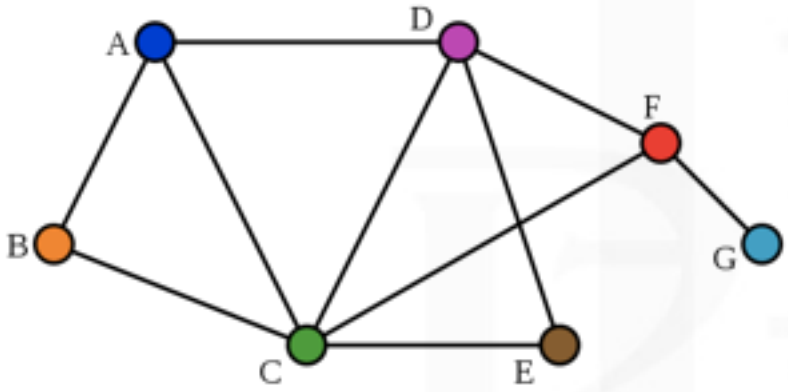
.....AATГГЦАЦЦГТААГГГГТЦЦГЦЦАТААЦЦГТТГЦААТТААЦГГАЦГГАТ.....

Расшифровка гибридизацией

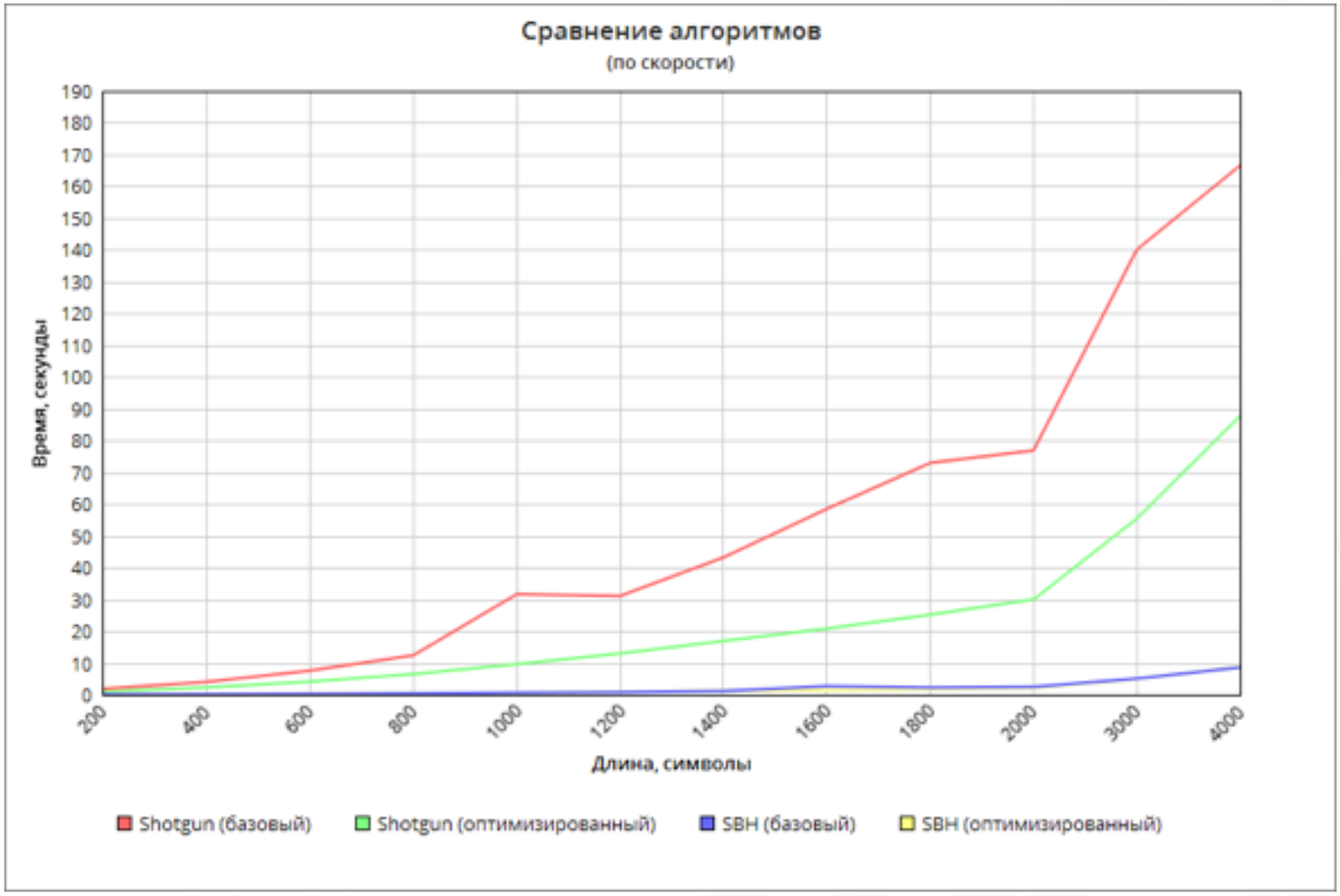


Идеи оптимизации

- Сжатие входных данных;
- Группировка;
- Интервальный граф.
- Эвристическое ускорение;
- Побочная информация;
- Программные средства.



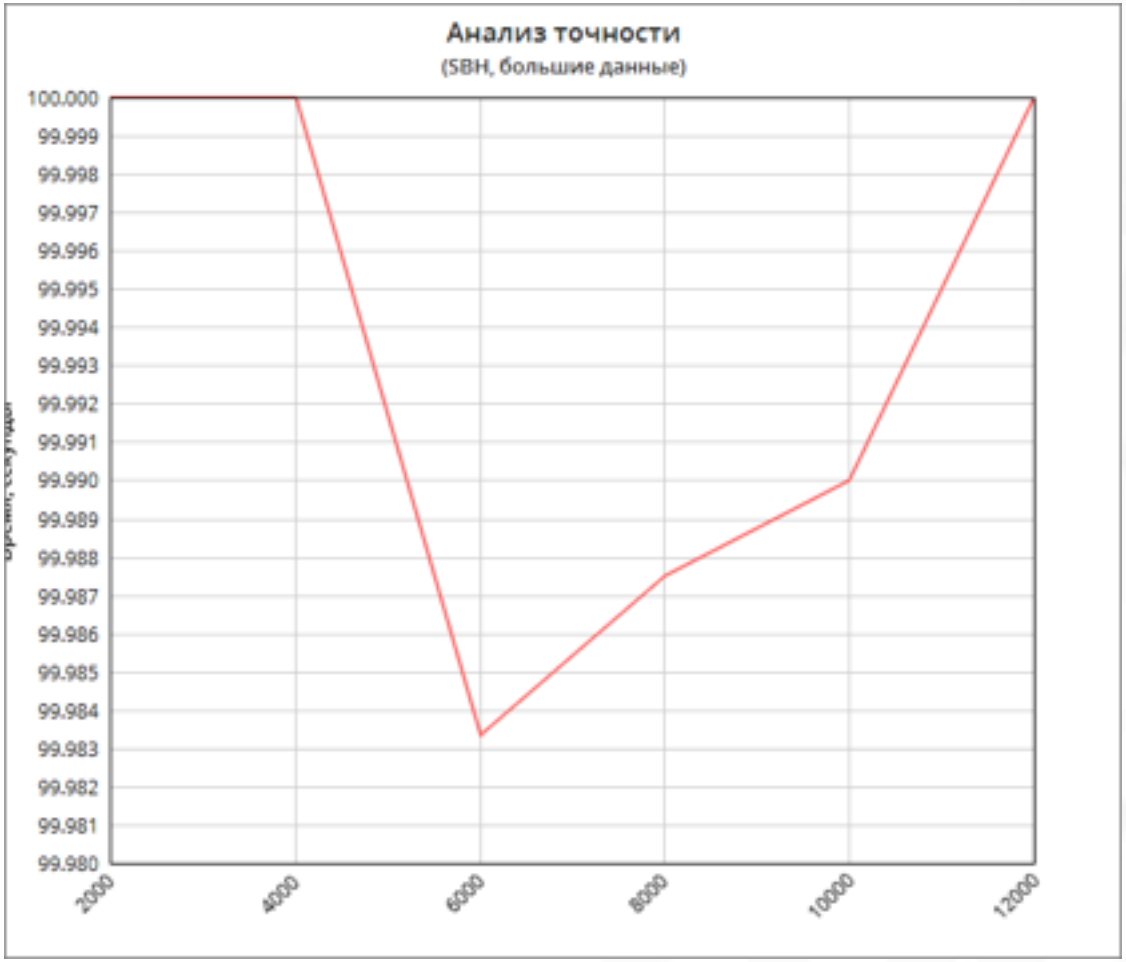
Эксперименты



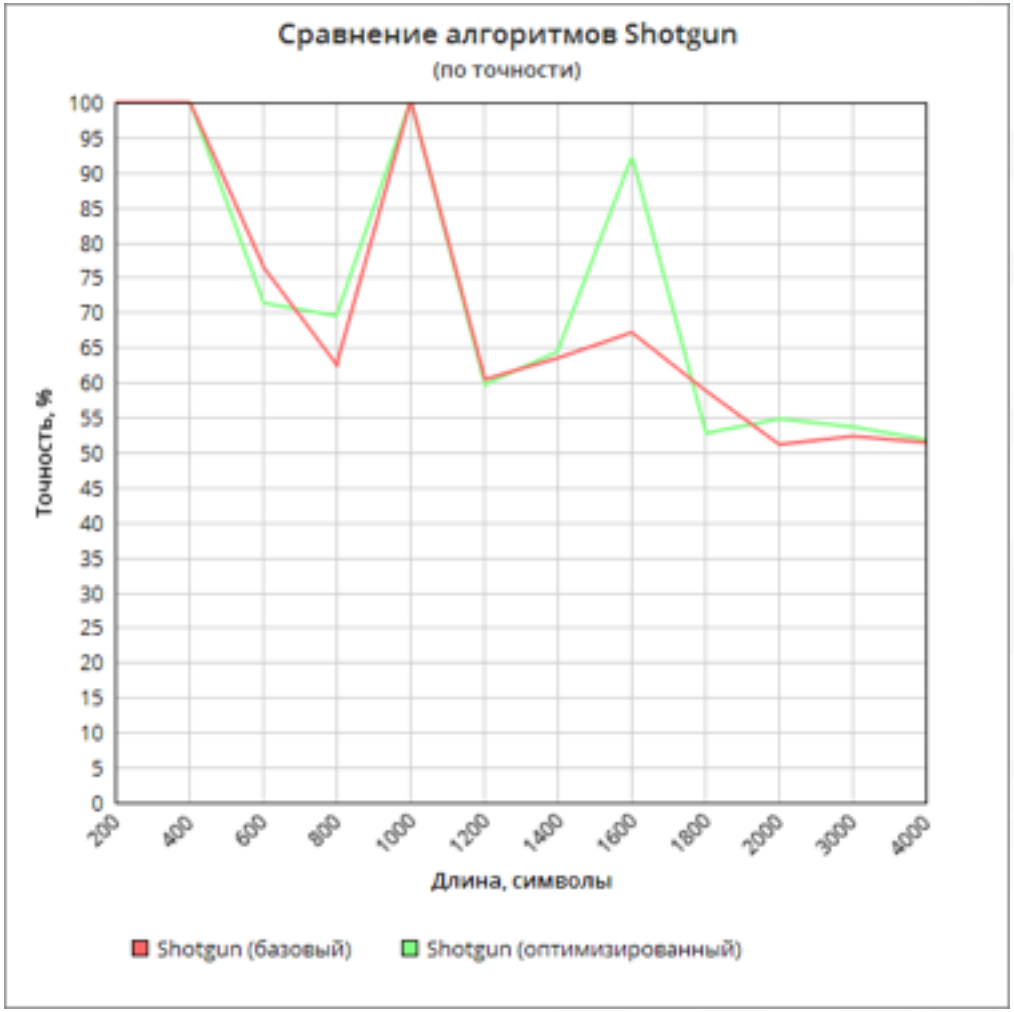
Определение корректности

- Метрика Левенштейна;
- Расстояние Дамерау-Левенштейна.

Эксперименты



Эксперименты



Выявленные проблемы

- Проблема с задачей о поиске гамильтонового цикла;
- Проблемы, связанные с дробным секвенирование;
- Проблема с поиском кратчайшей надстроки;
- Проблемы, связанные с SBH.

Дальнейшие перспективы

- Параллельные вычисления;
- Локализация генов;
- Оптимизации.

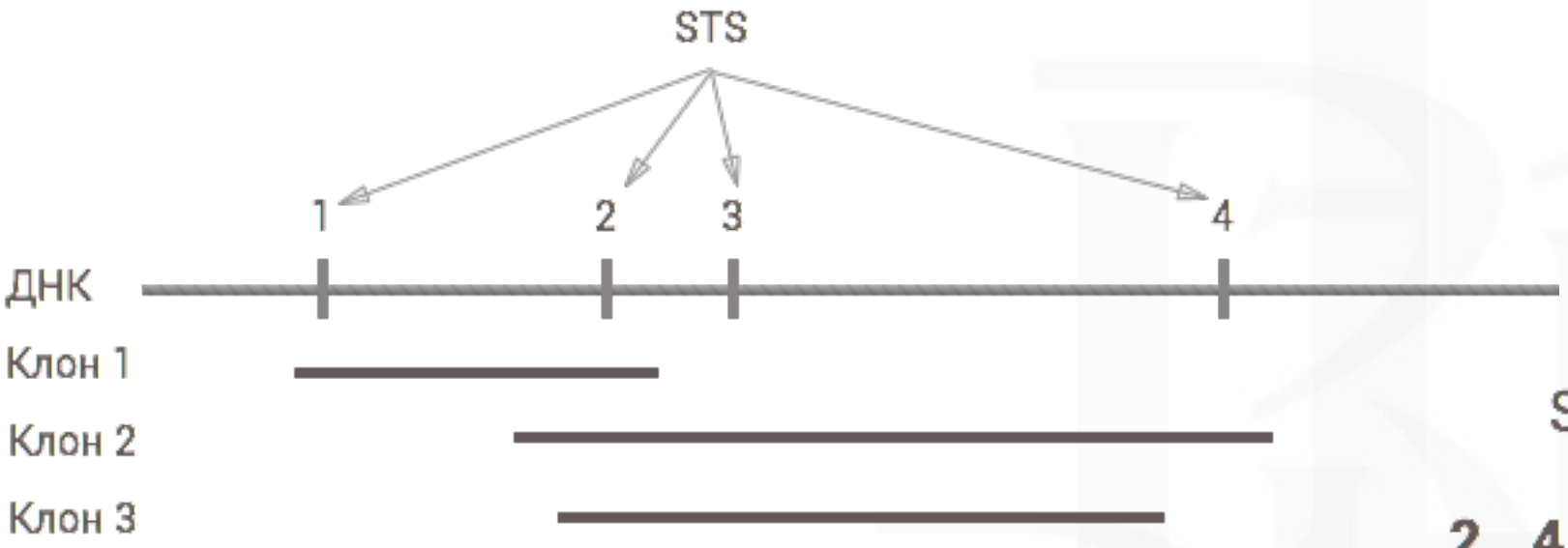
Спасибо за внимание!

Дробинин Вадим Дмитриевич,
vadim@drobinin.com

Москва - 2015

Вопросы

Библиотеки клонов



Клоны	STS 2	4	1	3
1	1	0	1	0
2	1	1	0	1
3	1	0	0	1

Поиск гамильтонового пути

Дано множество строк x_i. Рассмотрим следующий ориентированный граф G: каждая его вершина v_i соответствует строке x_i, каждое ребро (v_i, v_j) — перекрытию между строками x_i и x_j. В графе G требуется найти гамильтонов путь.

Задача *NP*-полная.

Задача о наименьшей надстроке

Дано множество строк x_i. Требуется найти минимальную по длине строку s такую, что все x_i входят в нее в качестве подстроки.

Задача *NP*-полная, но существует жадный алгоритм, дающий неплохие результаты на практике.

Поиск эйлерова цикла

procedure find_all_cycles (v)
var массив cycles

1. *пока есть цикл, проходящий через v, находим его*
    1. *добавляем все вершины найденного цикла в массив cycles (сохраняя порядок обхода)*
    2. *удаляем цикл из графа*
2. *идем по элементам массива cycles*
    1. *каждый элемент cycles[i] добавляем к ответу*
    2. *из каждого элемента рекурсивно вызываем себя: find_all_cycles (cycles[i])*

Алгоритм Вагнера-Фишера

$$d(S_1, S_2) = D(M, N), \text{ где}$$

$$D(i,j) = \begin{cases} 0 & ;\ i = 0,\ j = 0 \\ i & ;\ j = 0,\ i > 0 \\ j & ;\ i = 0,\ j > 0 \\ D(i-1, j-1) & ;\ S_1[i] = S_2[j] \\ \min\big(& \\ \quad D(i, j-1) + insertCost & \\ \quad D(i-1, j) + deleteCost & ;\ j > 0,\ i > 0,\ S_1[i] \neq S_2[j] \\ \quad D(i-1, j-1) + replaceCost & \\ \big) & \end{cases}$$

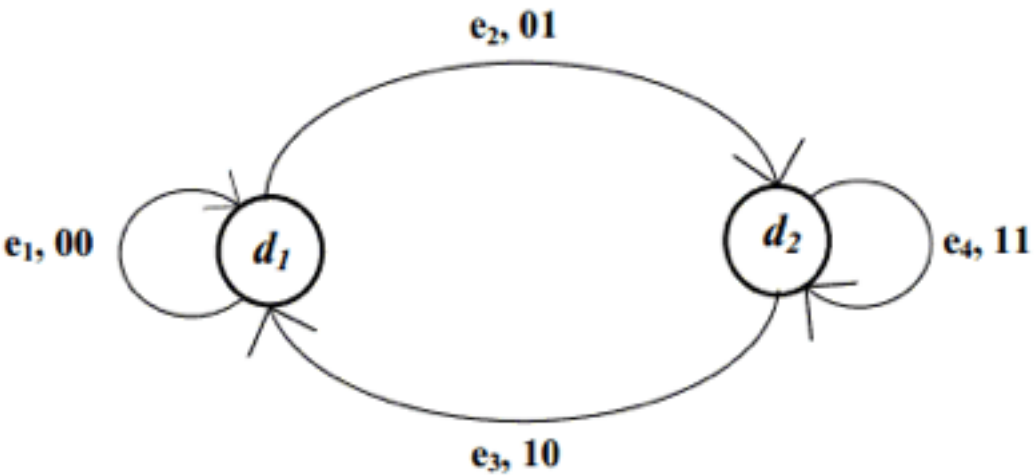
Расстояние Дамерау-Левенштейна

$$D(i,j) = \min\left(A, D(i',j') + (i - i' - 1) \cdot deleteCost + transposeCost + (j - j' - 1) \cdot insertCost\right) (*)$$

, где

$$A = \begin{cases} 0 & ;\ i = 0,\ j = 0 \\ i & ;\ j = 0,\ i > 0 \\ j & ;\ i = 0,\ j > 0 \\ D(i-1, j-1) & ;\ S[i] = T[j] \\ \min(& \\ \quad D(i, j-1) + insertCost & \\ \quad D(i-1, j) + deleteCost & ;\ j > 0,\ i > 0,\ S[i] \neq T[j] \\ \quad D(i-1, j-1) + replaceCost & \\ ) & \end{cases}$$

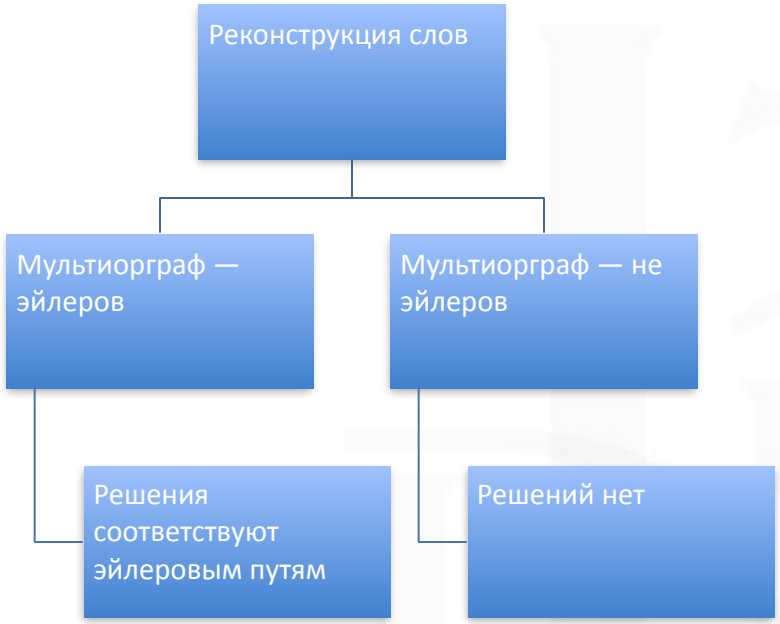
Мультиорграф де Брёйна



Мультиорграф
де Брёйна для множества
подслов
{00,01,10,11}

$$h_1 = (d_1, d_1, e_1, 1, 00),\ h_2 = (d_1, d_2, e_2, 1, 01),\ h_3 = (d_2, d_1, e_3, 1, 10),\ h_4 = (d_2, d_2, e_4, 1, 11).$$

Мультиорграф де Брёйна



Мультиорграф де Брёйна

Поиск и перечисление всех эйлеровых путей

1. Идея возведения в степень матрицы смежности графа на основе символьного умножения имен дуг;
2. Эйлеров цикл, при фиксации начальной вершины обхода цикла, является эйлеровым путем.

$$A = \begin{pmatrix} (e_1) & (e_2) \\ (e_3) & (e_4) \end{pmatrix},$$

$$A^2 = \begin{pmatrix} (e_2,e_3) & (e_1,e_2)\oplus(e_2,e_4) \\ (e_3,e_1)\oplus(e_4,e_3) & (e_3,e_2) \end{pmatrix},$$

$$A^3 = \begin{pmatrix} (e_2,e_3,e_1)\oplus(e_1,e_2,e_3)\oplus(e_2,e_4,e_3) & (e_1,e_2,e_4) \\ (e_4,e_3,e_1) & (e_3,e_1,e_2)\oplus(e_4,e_3,e_2)\oplus(e_3,e_2,e_4) \end{pmatrix},$$

$$A^4 = \begin{pmatrix} (e_2,e_4,e_3,e_1)\oplus(e_1,e_2,e_4,e_3) & \varnothing \\ \varnothing & (e_4,e_3,e_1,e_2)\oplus(e_3,e_1,e_2,e_4) \end{pmatrix}.$$

Поиск эйлеровых путей на примере графа, построенного по множеству подслов {00, 01, 10, 11}.

Мультиорграф де Брёйна

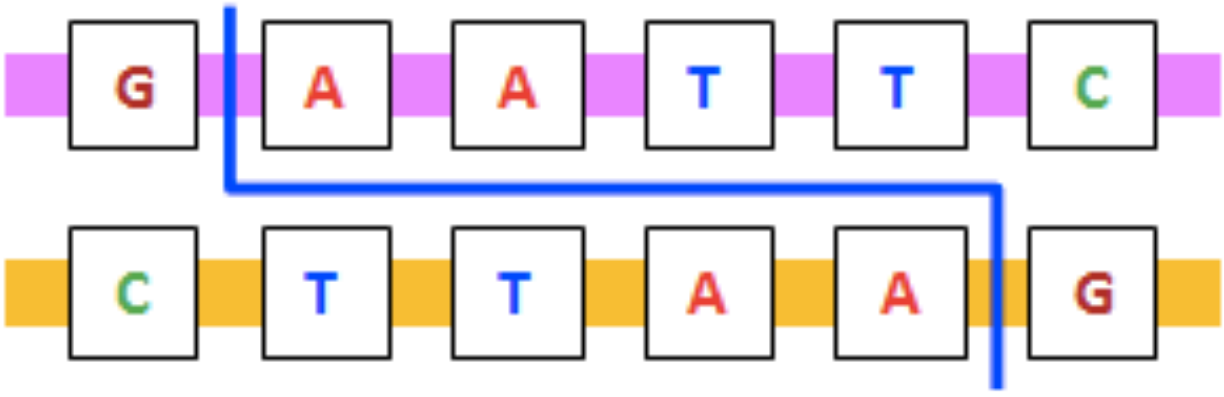
Реконструкция на основе эйлеровых путей

(e2, e4, e3, e1) => (e2 = 01, e4 = 11, e3 = 10, e1 = 00)
=> (01,11,10,00) => (0,1,1,0,0) => 01100

(e2, e4, e3, e1) => 01100
(e1, e2, e4, e3) => 00110
(e4, e3, e1, e2) => 11001
(e3, e1, e2, e4) => 10011

Реконструкция слов на примере графа, построенного по множеству подслов {00, 01, 10, 11}.

Рестриктазы



Рестриктаза EcoR I

(бактерия *Escherichia coli*)

GenBank

http://www.ncbi.nlm.nih.gov/genbank/