



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

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.



# Объект исследования



...ATTGTACSTAT...

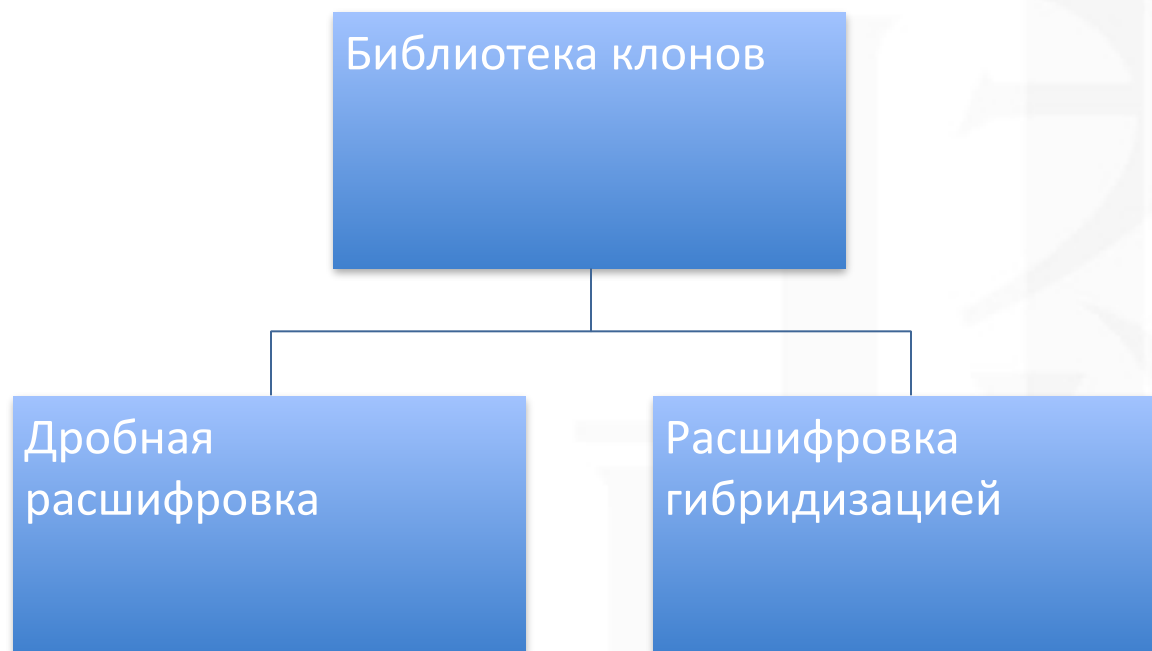
- 1988 год — проект «Геном человека»
  - 3 000 000 000 \$ на секвенирование \*
- 2001 год — черновые результаты
  - 6 000 000 000 \$ расходов \*
- 2014 год — «Генотек»
  - 375 000 руб. за анализ ДНК \*\*

\* <http://mygenome.su/articles/89/>

\*\* <https://www.genotek.ru/>

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

# Методы восстановления ДНК



# Дробная расшифровка

Геномная ДНК



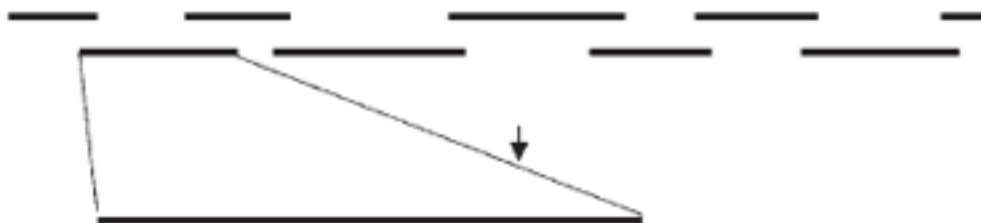
↓ Рестрикция и клонирование  
крупных фрагментов

Геномная библиотека



↓ Ранжирование рестриционных  
фрагментов на основании их  
перекрывания

Контиг



# Дробная расшифровка



**Последовательности  
фрагментов**

.....ААТГГЦАЦГТААГГГТЦЦГЦАТААЦГТТГЦ

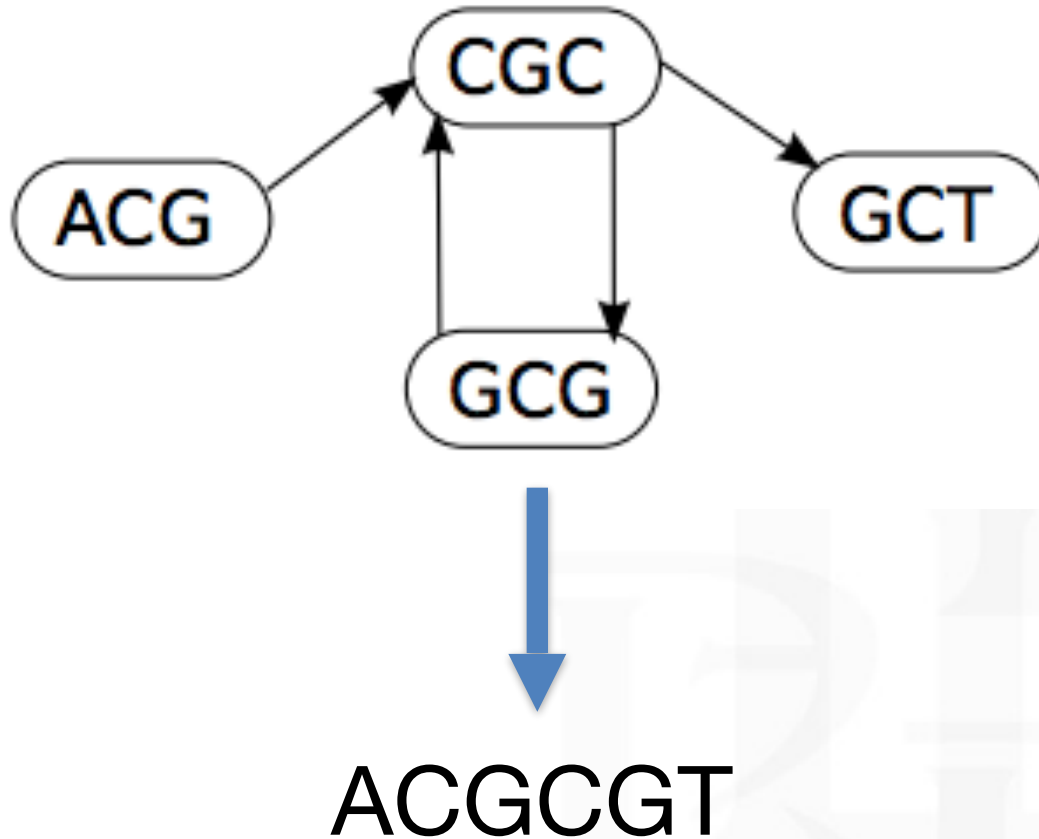
ТАТТГЦААЦГААТТААЦГГАЦГГАТ....

**Результирующая нуклеотидная последовательность**

.....ААТГГЦАЦГТААГГГТЦЦГЦАТААЦГТТГЦААТТААЦГГАЦГГАТ.....

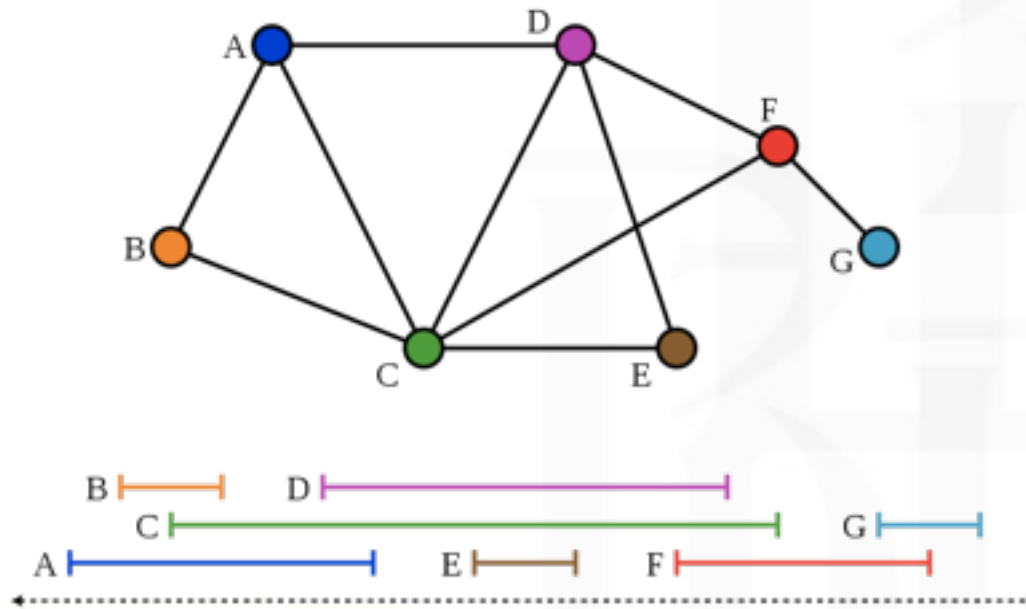


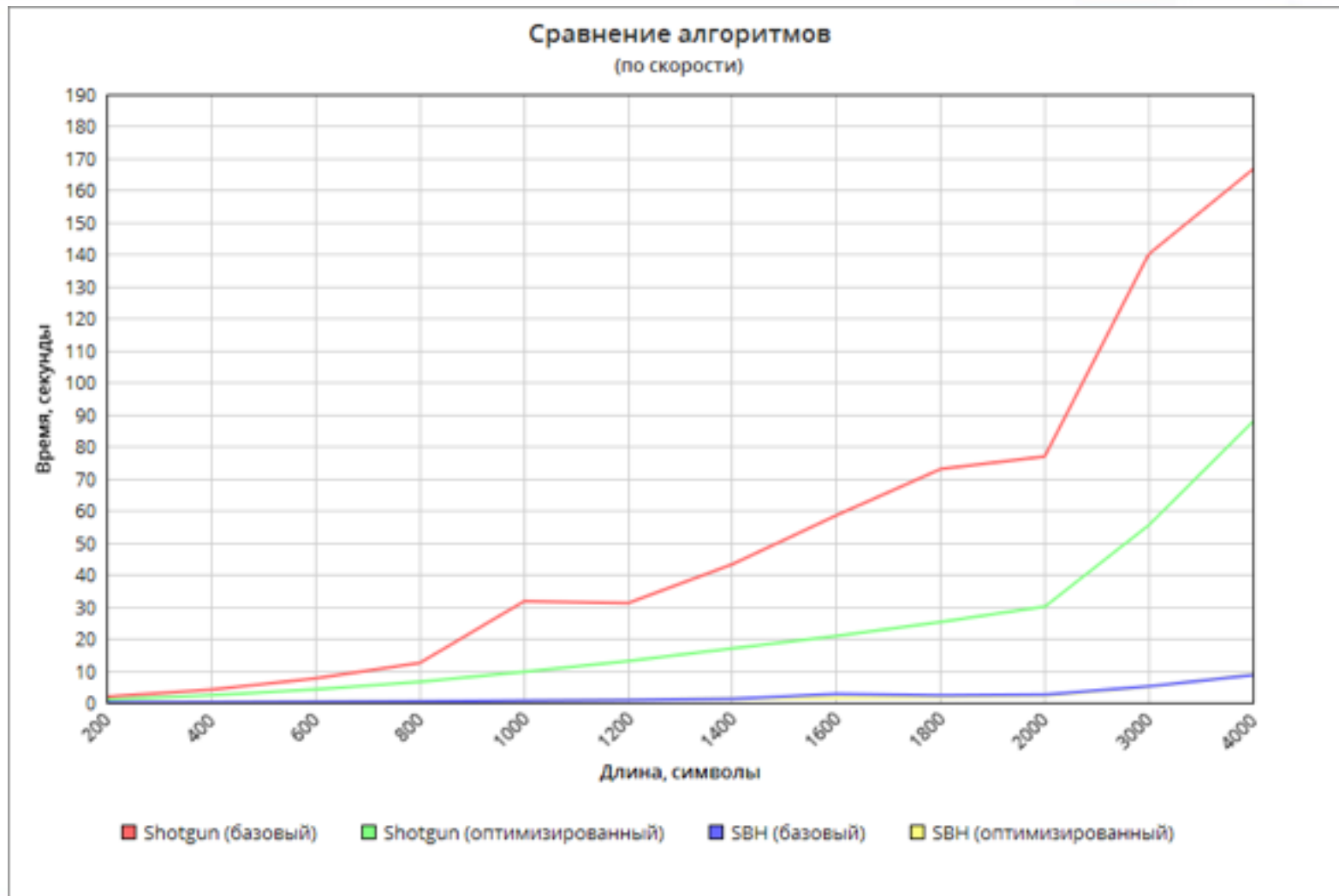
# Расшифровка гибридизацией



# Идеи оптимизации

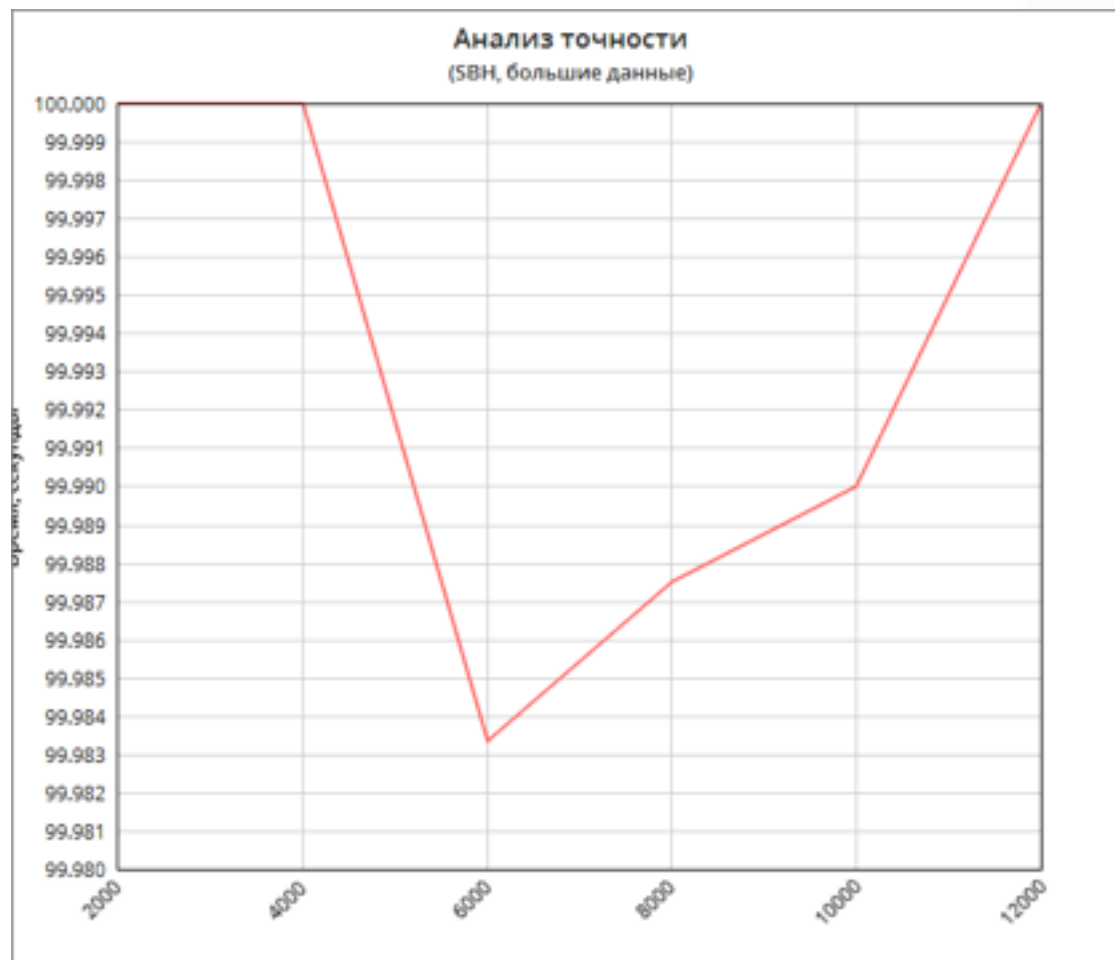
- Сжатие входных данных;
- Группировка;
- Интервальный граф.
- Эвристическое ускорение;
- Побочная информация;
- Программные средства.





# Определение корректности

- Метрика Левенштейна;
- Расстояние Дамерау-Левенштейна.





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

# Дальнейшие перспективы

- Параллельные вычисления;
- Локализация генов;
- Оптимизации.







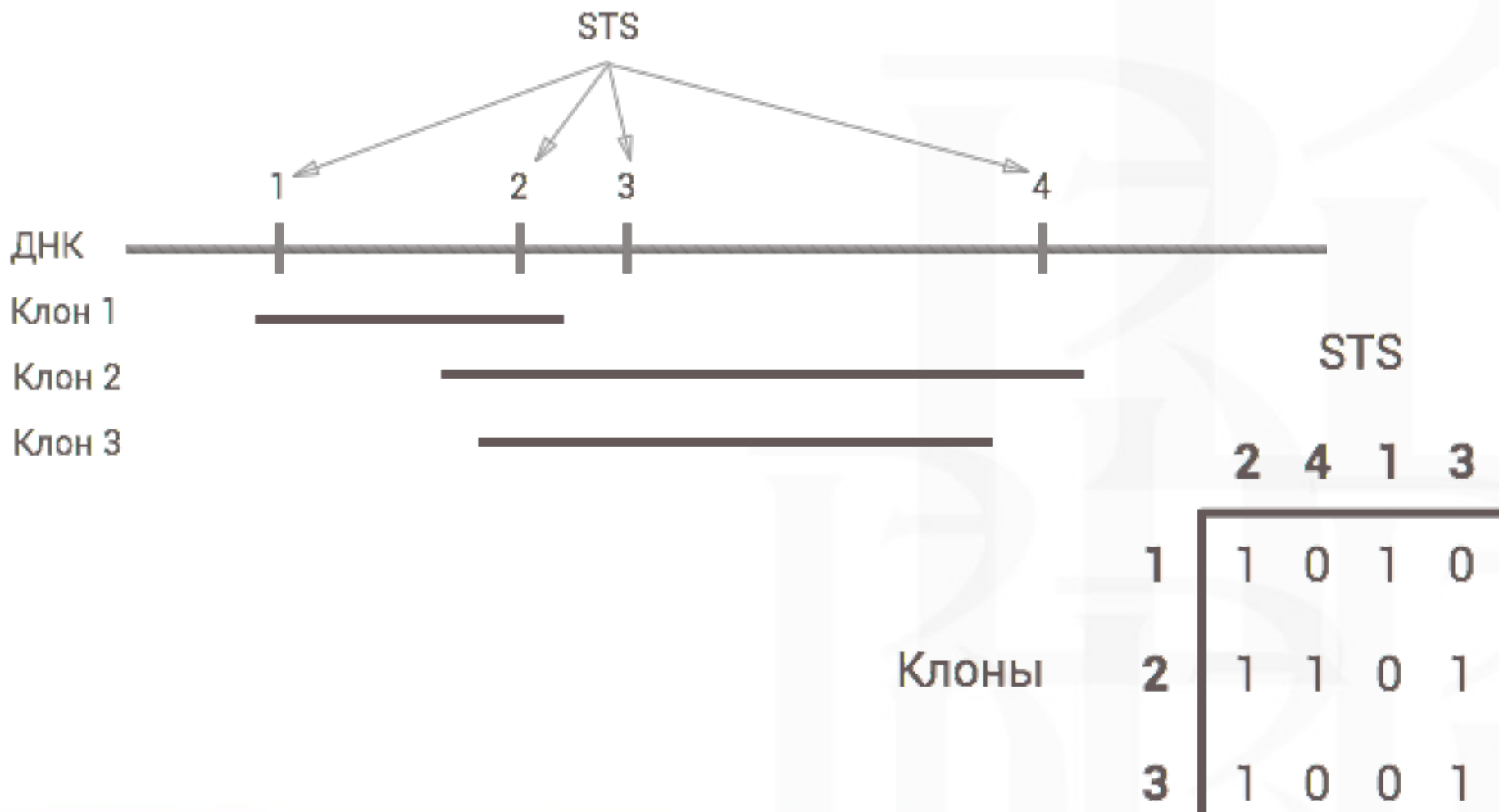
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ

# Спасибо за внимание!

Дробинин Вадим Дмитриевич,  
[vadim@drobinin.com](mailto:vadim@drobinin.com)

Москва - 2015





# 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

$d(S_1, S_2) = D(M, N)$ , где

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

# Расстояние Дамерау-Левенштейна

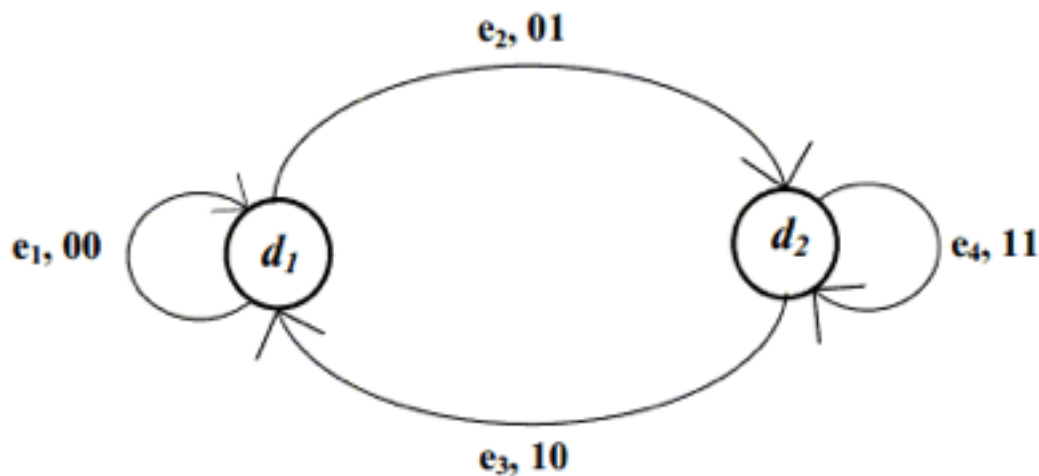
$$D(i, j) = \min (A, D(i', j') + (i - i' - 1) \cdot deleteCost + transposeCost + (j - j' - 1) \cdot insertCost) (*)$$

, где

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



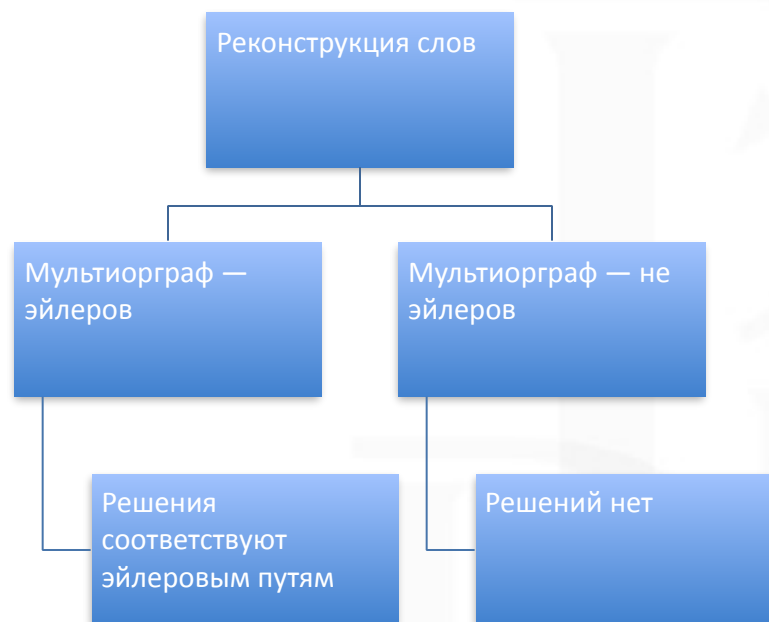
# Мультиорграф де Брёйна



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

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

# Мультиорграф де Брёйна



## Поиск и перечисление всех эйлеровых путей

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

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

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

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

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

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

## Реконструкция на основе эйлеровых путей

$(e_2, e_4, e_3, e_1) \Rightarrow (e_2 = 01, e_4 = 11, e_3 = 10, e_1 = 00)$   
 $\Rightarrow (01, 11, 10, 00) \Rightarrow (0, 1, 1, 0, 0) \Rightarrow 01100$

$(e_2, e_4, e_3, e_1) \Rightarrow 01100$

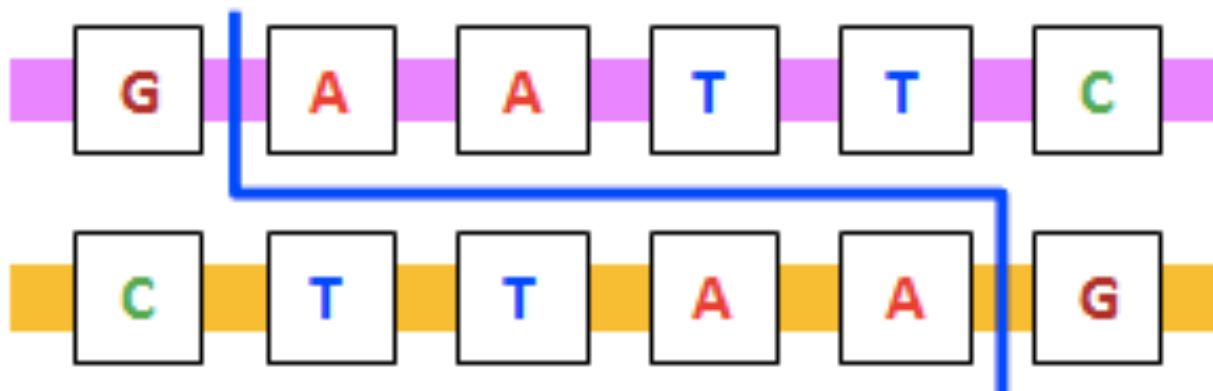
$(e_1, e_2, e_4, e_3) \Rightarrow 00110$

$(e_4, e_3, e_1, e_2) \Rightarrow 11001$

$(e_3, e_1, e_2, e_4) \Rightarrow 10011$

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/>

The screenshot shows the GenBank search interface. The search term '16000[SLEN]' is entered in the search bar. The results page displays two entries, both for 16,000 bp linear DNA. The first entry is for *Pseudomonas aeruginosa* strain Pae\_CF67.01, and the second is for *Humulus lupulus* var. *lupulus*. The interface includes navigation links like 'First', 'Prev', 'Page 1 of 17', 'Next', and 'Last'. There are also options for 'Display Settings' (Summary, 20 per page) and 'Send to'.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide 16000[SLEN] Search

Save search Advanced Help

Species  
Animals (223)  
Plants (21)  
Fungi (5)  
Protists (4)  
Bacteria (73)  
Customize ...

Molecule types  
genomic DNA/RNA (326)  
mRNA (3)  
Customize ...

Source databases  
GenBank (208)  
RefSeq (89)  
Customize ...

Display Settings: Summary, 20 per page, Sorted by Default order Send to: Filters: Manage Filters

Results: 1 to 20 of 331 << First < Prev Page 1 of 17 Next > Last >>

[Pseudomonas aeruginosa strain Pae\\_CF67.01 CF67.01 contig\\_13 whole genome shotgun sequence](#)  
16,000 bp linear DNA  
Accession: LCSS01000013.1 GI: 821229337  
[GenBank](#) [FASTA](#) [Graphics](#)

[Humulus lupulus var. lupulus DNA contig: SW\\_scaffold40250\\_size15995 whole genome shotgun sequence](#)  
2. [sequence](#)  
16,000 bp linear DNA  
Accession: LD172672.1 GI: 735955442  
[GenBank](#) [FASTA](#) [Graphics](#)

Results by taxon  
Top Organisms [Tree]  
Jaculus jaculus (9)  
Echinops telfairi (9)  
Latimeria chalumnae (8)  
Escherichia coli (6)  
Balaenoptera bonaerensis (5)  
All other taxa (294)  
More...

Find related data  
Database: Select