

Алгоритм выравнивания последовательностей ДНК

Проект по информатике за 4 семестр

Работу выполнили:
Олег Зубенко
Антон Анфиногентов
Антон Алашеев
Антон Тураев

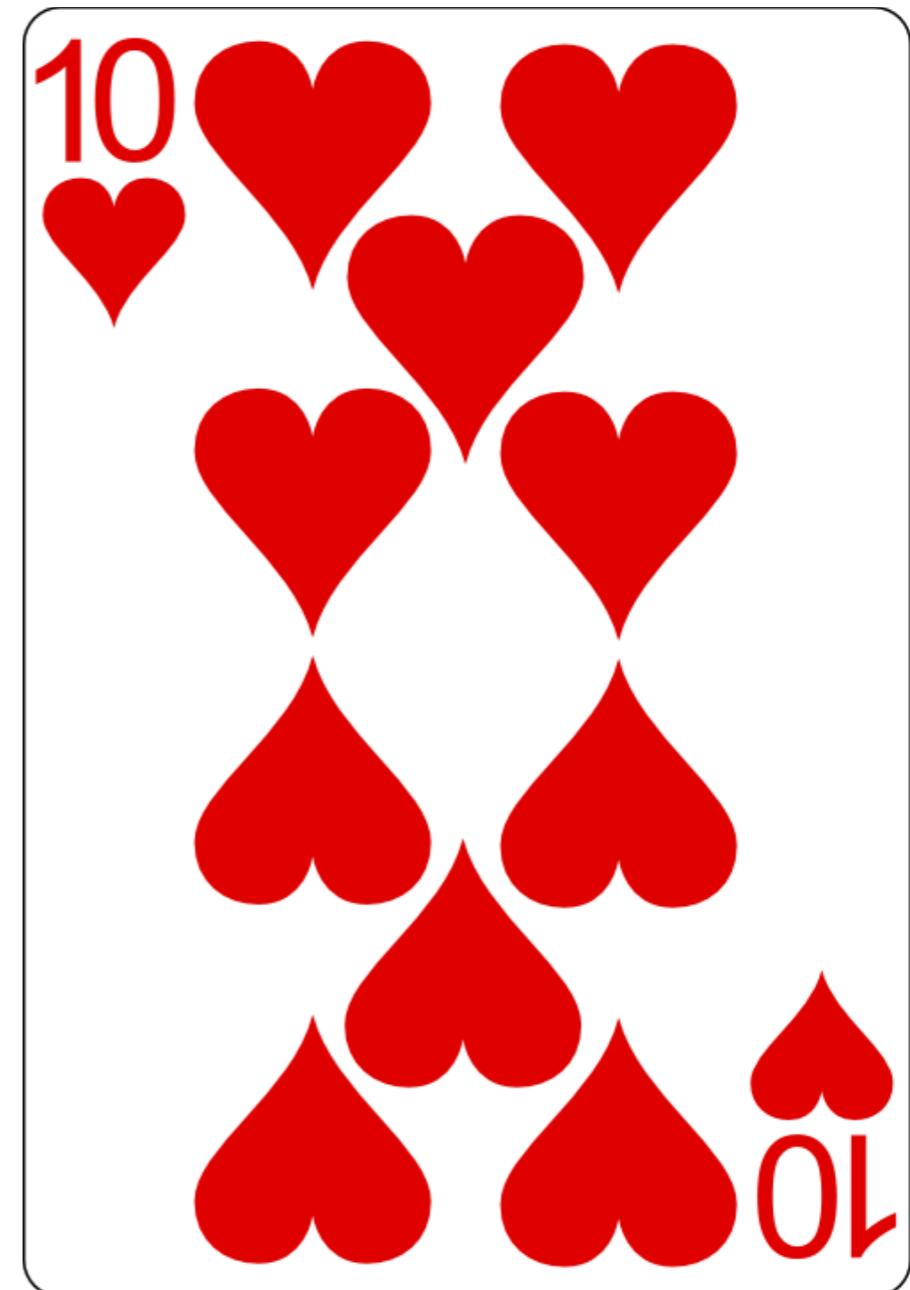
Описание проекта

Решение задачи выравнивания двух последовательностей нуклеотидов для определения степени схожести с использованием ООП С++

Применение

Работа предложена научной группой функциональной геномики человека Медико-генетического научного центра РАМН.

Если процент совпадения выровненных последовательностей больше 99%, организмы можно относить к одному виду.



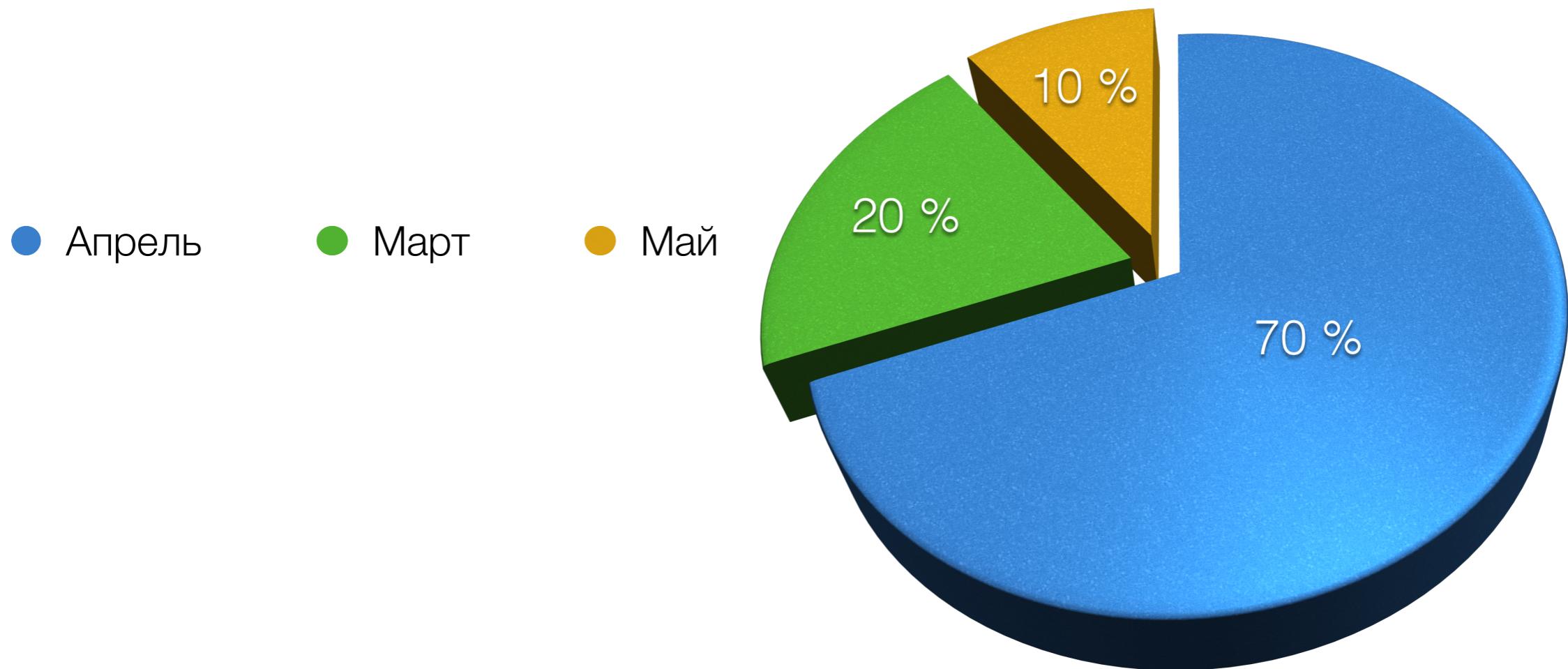
Задачи

- ООП
- Динамическое программирование
- Работа в команде. Git
- Отделение интерфейса от реализации
- Makefile

Оборудование

- ОС: Xubuntu 14.04, Mac OS X Yosemite 10.10.3
- Компиляторы: g++, c++
- Кодовые редакторы: Sublime Text 2
- Язык программирования: C++

Распределение времени



Моделирование

- Реализация алгоритма Нидлмана-Вунша
- Подключение ввода/вывода из файла
- Создание режимов работы программы

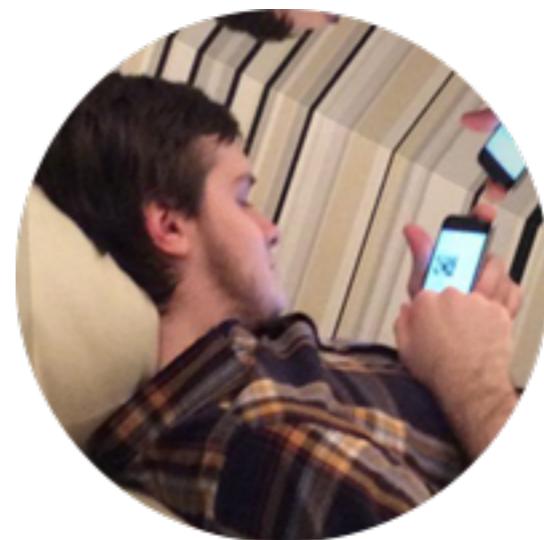
Ход работы программы

- Ввод последовательностей из файла
- Определение режима работы
- Выравнивание
- Вывод результата в файл / консоль

Роли



Олег Зубенко
Алгоритм



Антон Алашев
Ввод



Антон Анфиногентов
Ввод, вывод



Антон Тураев
Выход

Ввод

- Формат файла: .FASTA
- Выбор последовательностей через консоль
- Usage
- Режимы программы
- Защита от дурака

```
align.fasta ×  
1 >human  
2 ATGTCGCCGACCGTTGACTATTCTCTACAAACCACAAAGACATTGGAACACTATACCTATTATCGCGC.  
3 >chimp  
4 ATGTCACCGACCGCTGACTATTCTCTACAAACCACAAAGATATTGGAACACTATACCTACTATCGGTGC.  
5 >mouse  
6 ATGTCATTAATCGTTGATTATTCTAACCAATCACAAAGATATCGGAACCCCTTATCTACTATCGGAGC.  
7 >Henlea_nasuta  
8 TTAGGAGTATGAGCAGGAATAATAGGCGCAGCTATAAGCCTGCTAATCGAATTGAACTAAGACAACCAGG.  
9 >Henlea_ventriculosa  
10 GGAGTATGAGACGGAATAATAGGAGCAGCTATAAGTTATTAAATTGAACTAAGACAACCAGGTC.  
11 >Henlea_perpusilla  
12 ACACTATATTCTATTCTAGGCGTATGAGCCGGAATGATAGGAGCAGCCATAAGCCTTCTAATTGAAATTGA.  
13 >COI47  
14 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.  
15 >COI48  
16 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.  
17 >COI49  
18 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.  
19 >COI50  
20 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.  
21 >COI51  
22 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.  
23 >COI52  
24 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGTCACCACCTGCAGGGTC.
```

.FASTA

```
➔ dna-align git:(master) ✘ ./align align.fasta --file
Sequences found:
human chimp mouse Henlea_nasuta Henlea_ventriculosa Henlea_perpusilla COI47 COI48 COI49 COI50
Usage: <sequence.name> <sequence.name>
COI48 COI49
Aligned sequences printed into Align.txt
Traceback printed into Traceback.txt
Score printed into Score.txt
```

Выбор последовательностей

```
→ dna-align git:(master) ✘ ./align
```

Usage:

```
./align <file.fasta>
```

Functions:

```
--identity <number>
```

Number of decimal places for identity value

```
--penalty <number> <number>
```

Set penalties for sequences

```
--file
```

Print aligned sequences into TXT file

```
--output.path <file>
```

Sequence output TXT file

```
--traceback.path <file>
```

Traceback output TXT file

```
--score.path <file>
```

Score output TXT file

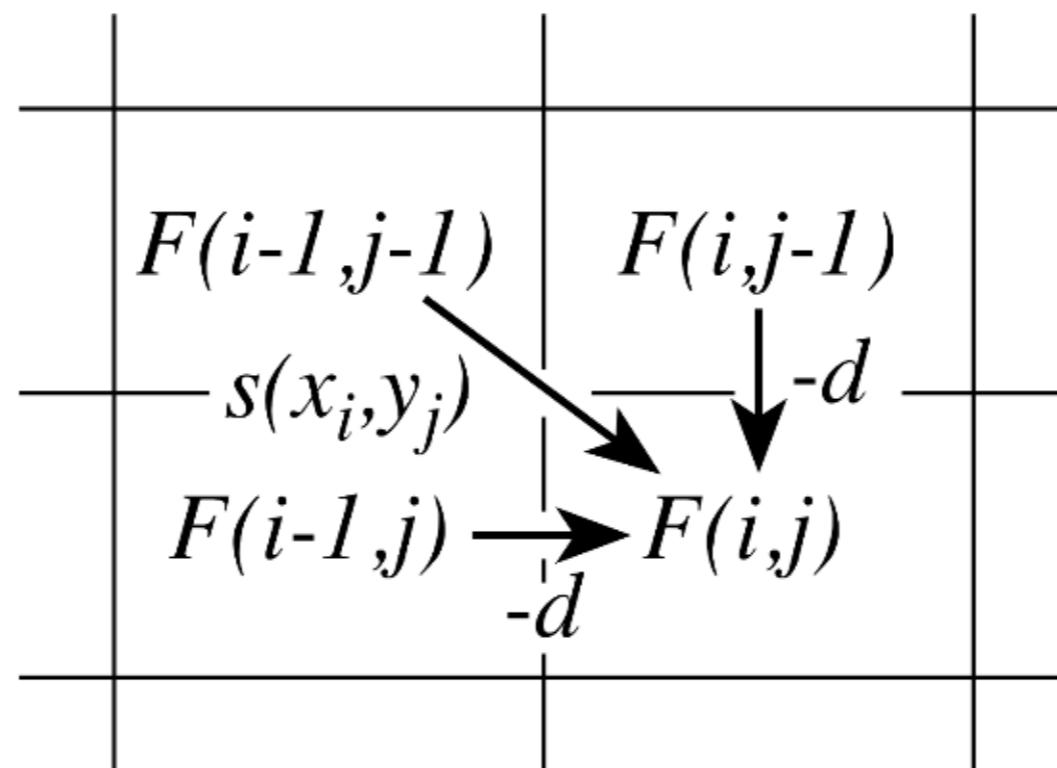
```
→ dna-align git:(master) ✘
```

Usage

Режимы

- Точность определения процента совпадения
- Штрафы за разрыв
- Вывод в файл/консоль
- Вывод матриц похожести и «обратного пути»
- Задание имён для файлов ввода/вывода

Алгоритм



$$F(i, j) = \max \begin{cases} F(i - 1, j - 1) + s(x_i, y_j), \\ F(i - 1, j) - d, \\ F(i, j - 1) - d. \end{cases}$$

Алгоритм

2.3 Alignment algorithms

21

	H	E	A	G	A	W	G	H	E	E	
P	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
A	-8	-2	-9	-17	-25	-33	-41	-49	-57	-65	-73
W	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
H	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
E	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
A	-40	-22	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-30	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-38	-24	-11	-6	-12	-14	-15	-12	-9	1

HEAGAWGHE-E

--P-AW-HEAE

test-1	AAACCTACGTACA
test-2	AAATACGTAA

FASTA format, sequence identity = 77%

test-1	AAACCTACGTACA [13]
test-2	AAA--TACGTA-A [10]
	*** ****** *

Пример работы алгоритма

Score.txt

1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
2	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
3	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
4	-3	-1	1	3	2	1	0	-1	-2	-3	-4
5	-4	-2	0	2	3	2	2	1	0	-1	-2
6	-5	-3	-1	1	2	3	3	2	1	0	-1
7	-6	-4	-2	0	2	2	3	3	3	2	1
8	-7	-5	-3	-1	1	3	2	3	3	4	3
9	-8	-6	-4	-2	0	2	4	3	3	3	4
10	-9	-7	-5	-3	-1	1	3	5	4	3	3
11	-10	-8	-6	-4	-2	0	2	4	6	5	4
12	-11	-9	-7	-5	-3	-1	1	3	5	7	6
13	-12	-10	-8	-6	-4	-2	0	2	4	6	7
14	-13	-11	-9	-7	-5	-3	-1	1	3	5	7
15											

Матрица очков

Traceback.txt	
1	0 ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
2	X X X ~ X ~ ~ ~ X X
3	X X X ~ X ~ ~ ~ X X
4	X X X ~ X ~ ~ ~ X X
5	X X X ~ ~ ~ ~
6	X X X X X X X
7	X X X X X ~ ~
8	X X X X X X X X X
9	X ~ X X X
10	X ~ X X
11	X X ~ ~
12	X X X X X X
13	X X
14	X X X X X X
15	

Матрица обратного пути

Вывод

- Файлы ТХТ
- Консоль
- Процент совпадения
- Индикатор совпадений «*»

Вывод в файл

```
align.txt x
1 FASTA format, sequence identity = 99%
2
3
4 COI47 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGGTACCCACCTGCAGGGTCAAAGAAAGATGTATTGAGATTCGGT
5 COI52 TATACTTCTGGGTGTCCGAAGAATCAAAAAAGATGCTGGTATAGAATTGGGTACCCACCTGCAGGGTCAAAGAAAGATGTATTGAGATTCGGT
6
7
```

score.txt	x
1	0 -1 -2 -3 -4 -5 -6 -7 -8 -9 -10
2	-1 1 0 -1 -2 -3 -4 -5 -6 -7 -8
3	-2 0 2 1 0 -1 -2 -3 -4 -5 -6
4	-3 -1 1 3 2 1 0 -1 -2 -3 -4
5	-4 -2 0 2 4 3 2 1 0 -1 -2
6	-5 -3 -1 1 3 5 4 3 2 1 0
7	-6 -4 -2 0 2 4 6 5 4 3 2
8	-7 -5 -3 -1 1 3 5 6 5 4 4
9	-8 -6 -4 -2 0 2 4 5 7 6 5
10	-9 -7 -5 -3 -1 1 3 4 6 8 7
11	-10 -8 -6 -4 -2 0 2 3 5 7 9
12	-11 -9 -7 -5 -3 -1 1 3 4 6 8
13	-12 -10 -8 -6 -4 -2 0 2 4 5 7
14	-13 -11 -9 -7 -5 -3 -1 1 3 5 6
15	-14 -12 -10 -8 -6 -4 -2 0 2 4 6
16	-15 -13 -11 -9 -7 -5 -3 -1 1 3 5
17	-16 -14 -12 -10 -8 -6 -4 -2 0 2 4
18	-17 -15 -13 -11 -9 -7 -5 -3 -1 1 3
19	-18 -16 -14 -12 -10 -8 -6 -4 -2 0 2
20	-19 -17 -15 -13 -11 -9 -7 -5 -3 -1 1
21	-20 -18 -16 -14 -12 -10 -8 -6 -4 -2 0
22	-21 -19 -17 -15 -13 -11 -9 -7 -5 -3 -1
23	-22 -20 -18 -16 -14 -12 -10 -8 -6 -4 -2
24	-23 -21 -19 -17 -15 -13 -11 -9 -7 -5 -3

Вывод в консоль

```
→ dna-align git:(master) ✘ ./align align.fasta
Sequences found:
human chimp mouse Henlea_nasuta Henlea_ventriculosa Henlea_perpusilla a b c
Usage: <sequence.name> <sequence.name>
a b
Terminal output, sequence identity = 78%
a      ATGTTGCCGACCGTTGACTATTCTCT----- 27
b      ATGTTCACCGACCGCTGACTATTCTCTACAAA 32
***** ***** *****
```

[!\[\]\(98db30c87fe7287784edd65062d473d3_img.jpg\) 3112-students / dna-align](#)[Watch 2](#)[Star 0](#)[Fork 0](#)

DNA Alignment via Needleman Wunsch algroithm

19 commits

1 branch

0 releases

2 contributors



branch: master ▾

[branch: master](#) / +

Final touches, shrinking & optimizing

 factac authored 25 minutes agolatest commit 02160a93ca [Input.cpp](#)

Final touches, shrinking & optimizing

25 minutes ago

[Input.h](#)

Final touches, shrinking & optimizing

25 minutes ago

[Matrix.cpp](#)

Final touches, shrinking & optimizing

25 minutes ago

[Matrix.h](#)

Final touches, shrinking & optimizing

25 minutes ago

[NeedlemanWunsch.cpp](#)

Final touches, shrinking & optimizing

25 minutes ago

[NeedlemanWunsch.h](#)

Final touches, shrinking & optimizing

25 minutes ago

[Output.cpp](#)

Final touches, shrinking & optimizing

25 minutes ago

[Output.h](#)

Final touches, shrinking & optimizing

25 minutes ago

[README.md](#)

Final Input & Output

13 days ago

[Score.cpp](#)

Final Input & Output

13 days ago

[Score.h](#)

Final Input & Output

13 days ago

[Code](#)[Issues 0](#)[Pull requests 0](#)[Pulse](#)[Graphs](#)

HTTPS clone URL

<https://github.com/> You can clone with [HTTPS](#) or [Subversion](#). [Clone in Desktop](#)[Download ZIP](#)<https://github.com/3112-students/dna-align>

Спасибо за внимание!