Московский Государственный Университет им. М. В. Ломоносова
Факультет Вычислительной Математики и Кибернетики
Кафедра Суперкомпьютеров и Квантовой Информатики

Тойгильдин Владислав

Разработка и исследование параллельного алгоритма поиска неточных повторов в геноме.

Научный руководитель:
к.ф-м.н., доцент
Попова Нина Николаевна

апрель 2015

Цель работы

Дипломная работа посвящена параллельным методам анализа и обработки биологических последовательностей на суперкомпьютерах с целью сокращения времени обработки и увеличения объёма обрабатываемых данных.

Актуальность работы

В работе рассматривается задача поиска неточных протяженных повторов в биологических последовательностях. Важность поиска повторяющихся элементов обусловлена биологической ролью повторов в функционировании организмов.

Нахождение повторов позволяет решать многие задачи:

- Определение родства групп организмов на генном уровне.
- Диагностика генетических заболеваний.
- и другие

Мотивация постановки задачи

Мотивация постановки задачи связана с растущими объемами генетических данных и необходимостью проведения многократных быстрых вычислительных экспериментов.

В данной работе предлагается и исследуется возможность суперкомпьютерной реализации метода поиска повторов в биологических последовательностях.

Рассматриваемый метод

В дипломной работе рассматривается метод поиска протяженных размытых повторов, разработанный коллективом сотрудников кафедры математических методов прогнозирования ВМК МГУ и института математических проблем в биологии Российской академии наук (Пущино).

Дипломная работа выполнялась в тесном сотрудничестве с авторами метода. [1]

[1] ВСТАВЬ ССЫЛКУ НА РАБОТЫ

Задачи дипломной работы

- Разработка и реализация параллельного алгоритма поиска повторов, ориентированного на суперкомпьютерную реализацию: большой объем обрабатываемых данных (размер данных порядка 1 ГБ), время обработки в пределах 1 часа.
- Исследование эффективности использования графических процессоров для решения поставленной задачи.
- Анализ масштабируемости разработанного алгоритма на примере задачи сравнения конкретных геномов
- Разработка графического интерфейса для работы пользователя на локальной и удаленных системах.

Вычислительные платформы

Вычислительные платформы для решения поставленных задач:

- суперкомпьютеры Ломоносов, BlueGene/P
- 2-х процессорная рабочая станция на базе 4-х ядерного процессора Intel Xeon E5630 и 4-х графических ускорителей Tesla K40, Tesla K20c, 2xTesla C2075

Предложенные платформы обладают различной архитектурой и позволяют исследовать разрабатываемый алгоритм с учетом их особенностей.

Исходные данные

Объектом исследования является биологические последовательности, формально представляемые в виде последовательности символов A, T, G, C. Длины исследуемых последовательностей достигают порядка $10^6 - 10^9$ символов.

Источником используемых данных являются всемирные банки генетической информации, напримет NCBI. [2]

[2] ВСТАВЬ ССЫЛКУ НА РАБОТЫ

Математическая модель

Формально биологическая последовательность:

$$X = (x_n)_{n=1}^{N},\ N \in \mathbb{N},\ x_n \in \{A, T, G, C\},$$

где N - длина последовательности, A, T, G, C - озбоначения нуклеотидов.

Для последовательности X под $X|_i^k$ при условии $i + k < |X|$ будем понимать **часть последовательности** X с элемента с номером i длиной k:

$$X|_i^k = \{x_i, x_{i+1}, ..., x_{i+k-1}\}$$

Математическая модель

Под повтором будем понимать пару последовательностей (X_1,X_2), для которых справедливо неравенство:

$$\rho(X_1, X_2) \leq \varepsilon, \ |X_2| = |X_2| = K$$

где:

- K - длина последовательностей,
- $\rho(X_1, X_2)$ - т.н. расстояние редактирования, оценка близости последовательностей, [конкретный вид функции расстояния редактирования $\rho(X_1, X_2)$ **определяется алгоритмом**],
- ε - задаваемая точность поиска, значение которой будет **зависеть от задаваемой функции растояния**.

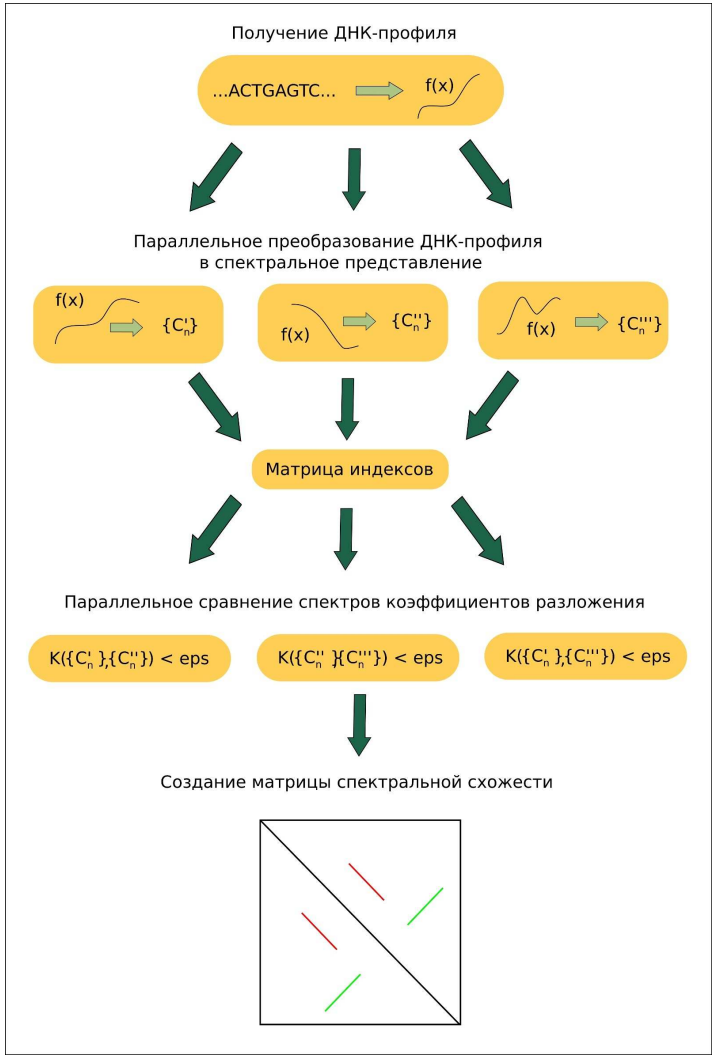
Математическая модель

Пусть есть две биологические последовательности
$X = (x_n)_{n=1}^{N_x}$ и $Y = (y_n)_{n=1}^{N_y}$

Под **задачей поиска повторов** будем понимать нахождение всех троек $\{i_x, i_y, k\}, \quad i_x, i_y, k \in \mathbb{N}$, таких что:
$i_x + k \leq N_x, i_y + k \leq N_y$ и $(X|_{i_x}^k, Y|_{i_y}^k)$ - повтор длины k, т.е.:

$$\rho(X|_{i_x}^k, Y|_{i_y}^k) \leq \varepsilon$$

Спектрально-аналитический метод



Спектрально-аналитический метод

```
char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA
```

Спектрально-аналитический метод

```
char sequence[N]; // A, T, G, C
```

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

L1 - длина окна; параметр алгоритма

Спектрально-аналитический метод

char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

считаем в окне некую
статистическую величину и получаем
неотрицательное число

4

Спектрально-аналитический метод

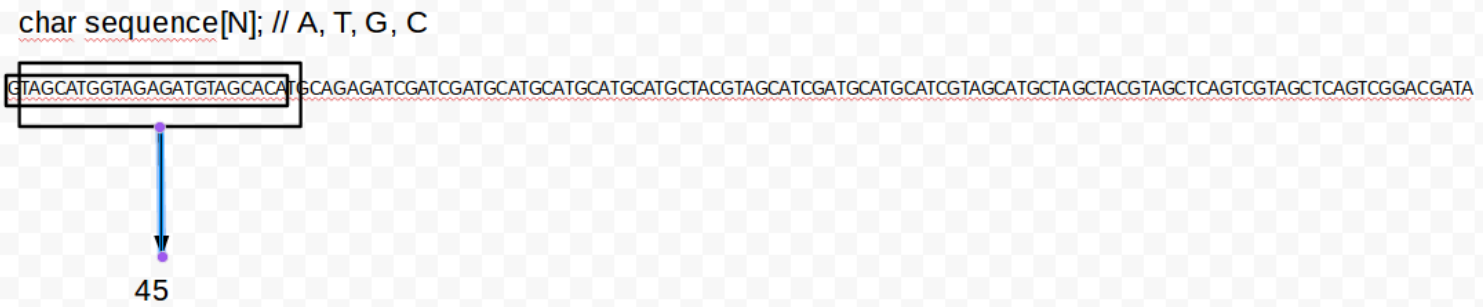
char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

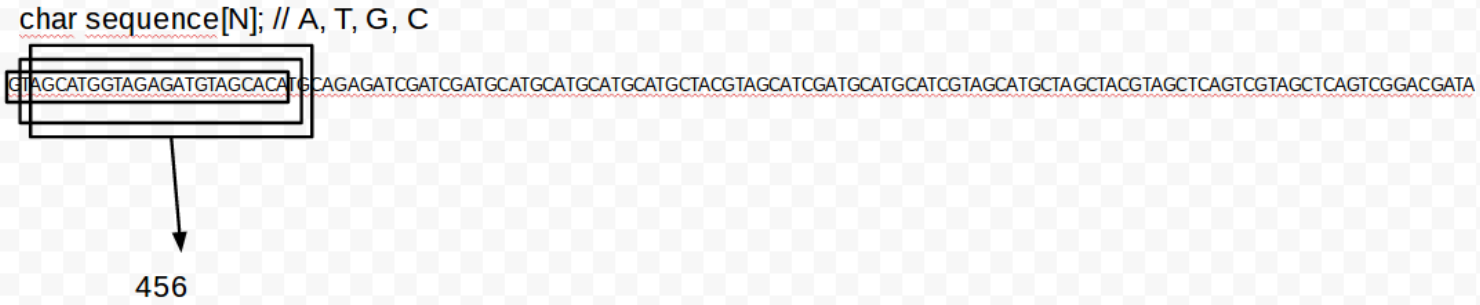
сдвигаем окно на 1 символ

4

Спектрально-аналитический метод



Спектрально-аналитический метод



Спектрально-аналитический метод

char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

из последовательности символов получаем дискретную числовую функцию

456765678987898878765456546543212101010123454543456765678987656567656767 87

Спектрально-аналитический метод

```
char sequence[N]; // A, T, G, C
```

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

```
unsigned int profile[N - L1 + 1];
```

4567656789878988787654565465432121010101234545434567656789876565676567678 7

L2 - длина окна; параметр алгоритма

Спектрально-аналитический метод

char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

unsigned int profile[N - L1 + 1];

4567656789878988787654565465432121010101234545434567656789876565676567678 7

считаем спектры - первые m коэффициентов разложения

1.42
3.23
2.12
6.23
4.76
2.74
4.02
1.63
6.35
4.14
7.52
0.25
1.25
0.67
2.52
1.25

m - параметр алгоритма

Спектрально-аналитический метод

char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

unsigned int profile[N - L1 + 1];

4567656789878988787654565465432121010101234545434567656789876565676567 6787

сдвигаем окно на s элементов; s - параметр алгоритма

1.42
3.23
2.12
6.23
4.76
2.74
4.02
1.63
6.35
4.14
7.52
0.25
1.25
0.67
2.52
1.25

Спектрально-аналитический метод

```
char sequence[N]; // A, T, G, C
```

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

```
unsigned int profile[N - L1 + 1];
```

4567656789878988787654565465432121010101234545434567656789876565676567678 7

1.42	6.35
3.23	4.14
2.12	7.52
6.23	0.25
4.76	1.25
2.74	0.67
4.02	2.52
1.63	1.25
6.35	1.42
4.14	3.23
7.52	2.12
0.25	6.23
1.25	4.76
0.67	2.74
2.52	4.02
1.25	1.63

Спектрально-аналитический метод

```
char sequence[N]; // A, T, G, C
```

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

```
unsigned int profile[N - L1 + 1];
```

4567656789878988787654565465432121010101234545434567656789876565676567678 7

1.42	6.35	0.67
3.23	4.14	2.52
2.12	7.52	1.25
6.23	0.25	1.42
4.76	1.25	3.23
2.74	0.67	2.12
4.02	2.52	6.23
1.63	1.25	4.76
6.35	1.42	2.74
4.14	3.23	4.02
7.52	2.12	1.63
0.25	6.23	6.35
1.25	4.76	4.14
0.67	2.74	7.52
2.52	4.02	0.25
1.25	1.63	1.25

Спектрально-аналитический метод

char sequence[N]; // A, T, G, C

GTAGCATGGTAGAGATGTAGCACATGCAGAGATCGATCGATGCATGCATGCATGCATGCTACGTAGCATCGATGCATGCATCGTAGCATGCTAGCTACGTAGCTCAGTCGTAGCTCAGTCGGACGATA

unsigned int profile[N - L1 + 1];

4567656789878988787654565465432121010101234545434567656789876565676567678 7

1.42	6.35	0.67
3.23	4.14	2.52
2.12	7.52	1.25
6.23	0.25	1.42
4.76	1.25	3.23
2.74	0.67	2.12
4.02	2.52	6.23
1.63	1.25	4.76
6.35	1.42	2.74
4.14	3.23	4.02
7.52	2.12	1.63
0.25	6.23	6.35
1.25	4.76	4.14
0.67	2.74	7.52
2.52	4.02	0.25
1.25	1.63	1.25

Проделываем тоже самое
со второй последовательностью

Спектрально-аналитический метод

Набор спектров первой последовательности

m - количество коэффициентов

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

N1 - количество спектров из первого набора

Спектрально-аналитический метод

Набор спектров второй последовательности

m

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

N1

m

8.23	7.62	0.71	9.18	4.66
3.46	4.12	3.35	7.27	2.92
5.61	0.60	5.21	1.78	5.48
2.32	8.66	0.73	4.07	5.07
4.56	5.56	8.58	6.57	1.75
3.52	1.97	7.18	6.95	8.57
2.88	6.15	0.83	0.90	6.78
3.64	4.49	0.02	8.10	8.93
6.09	7.22	6.57	5.25	3.35
5.95	4.70	7.77	8.98	0.44
4.63	7.10	3.27	2.20	3.72
1.22	3.58	7.88	2.83	3.67
6.94	2.42	1.72	7.18	8.85
6.46	1.70	2.61	0.69	6.34
5.94	4.22	6.01	9.68	5.23
5.03	2.97	3.82	7.63	5.39
8.63	9.96	5.41	5.30	7.38
0.50	9.21	2.84	6.82	4.06
5.86	4.22	6.54	3.52	9.04
0.64	0.75	8.00	3.18	2.25
4.95	3.87	7.47	9.38	9.23
6.35	2.53	3.93	8.61	3.23
6.35	5.32	4.48	3.26	6.63
8.21	8.26	5.74	8.80	0.03
8.57	3.13	6.95	0.17	4.51

N2 - количество спектров из второго набора

Спектрально-аналитический метод

m

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

N2

m

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

N1

Для удобства транспонируем второй набор

Спектрально-аналитический метод

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

Сравним каждый спектр из первого набора
с каждым спектром из второго набора
и построим гомологическую матрицу,
отражающую результат сравнения двух наборов

```
bool matrixGomology[N1][N2];
```

Спектрально-аналитический метод

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

Сравнение спектров проведем с помощью среднеквадратичного отклонения

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

$\sigma < \varepsilon$, где ε - параметр алгоритма

Спектрально-аналитический метод

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

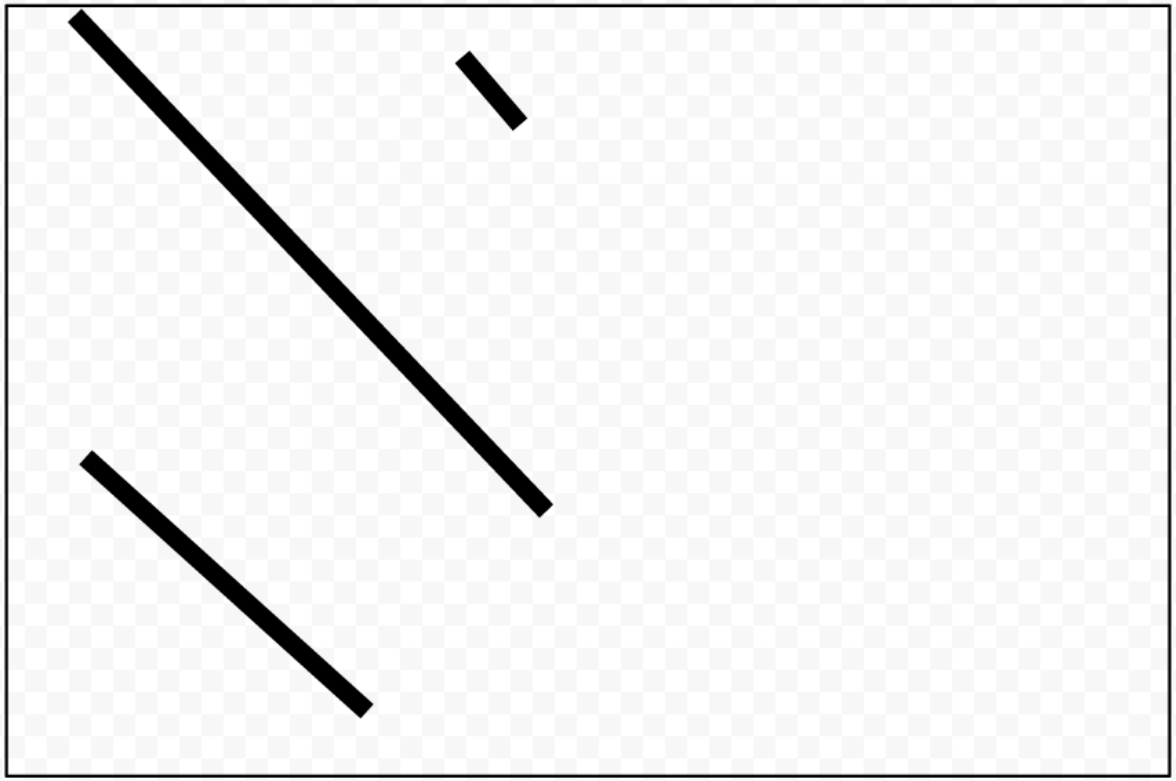
6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51

В случае выполнения $\sigma < \varepsilon$ в ячейку матрицы гомологии заносится true, иначе false. Для удобства графического представления будем закрашивать ячейку в черный цвет.

Спектрально-аналитический метод

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

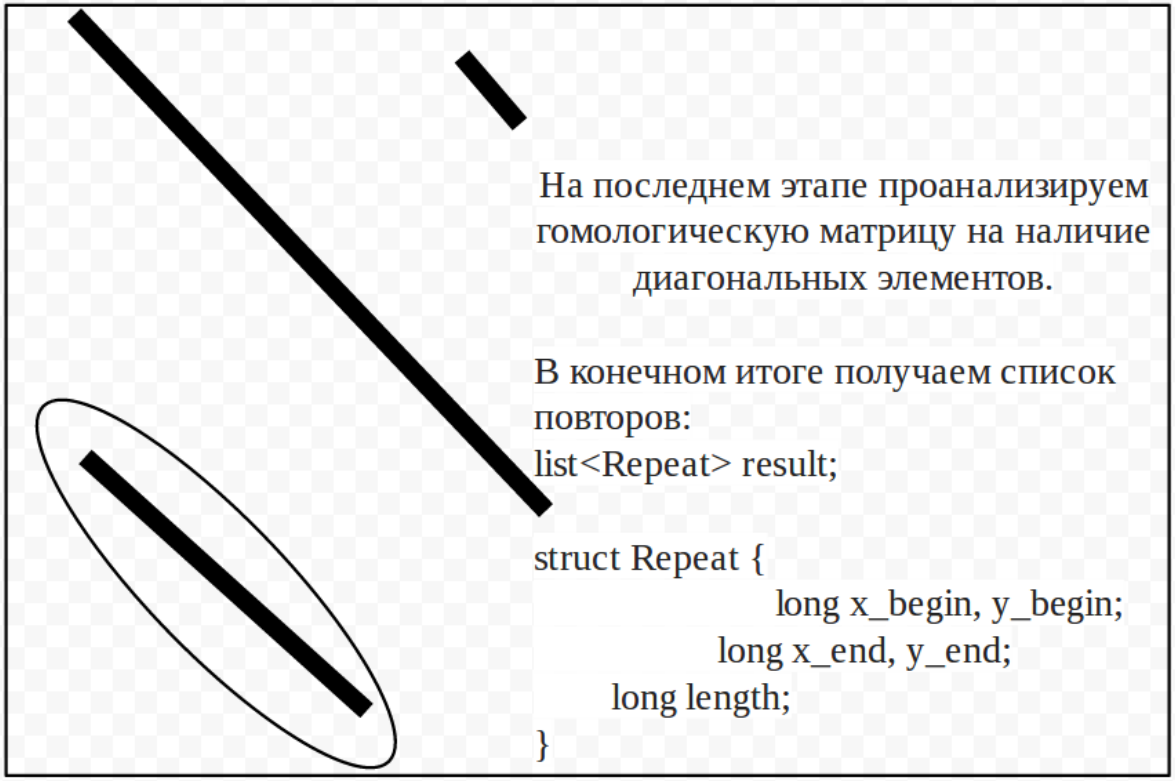
6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51



Спектрально-аналитический метод

8.23	3.46	5.61	2.32	4.56	3.52	2.88	3.64	6.09	5.95	4.63	1.22	6.94	6.46	5.94	5.03	8.63	0.50	5.86	0.64	4.95	6.35	6.35	8.21	8.57
7.62	4.12	0.60	8.66	5.56	1.97	6.15	4.49	7.22	4.70	7.10	3.58	2.42	1.70	4.22	2.97	9.96	9.21	4.22	0.75	3.87	2.53	5.32	8.26	3.13
0.71	3.35	5.21	0.73	8.58	7.18	0.83	0.02	6.57	7.77	3.27	7.88	1.72	2.61	6.01	3.82	5.41	2.84	6.54	8.00	7.47	3.93	4.48	5.74	6.95
9.18	7.27	1.78	4.07	6.57	6.95	0.90	8.10	5.25	8.98	2.20	2.83	7.18	0.69	9.68	7.63	5.30	6.82	3.52	3.18	9.38	8.61	3.26	8.80	0.17
4.66	2.92	5.48	5.07	1.75	8.57	6.78	8.93	3.35	0.44	3.72	3.67	8.85	6.34	5.23	5.39	7.38	4.06	9.04	2.25	9.23	3.23	6.63	0.03	4.51

6.27	9.87	7.55	9.63	8.67
7.72	2.00	9.83	0.60	3.84
9.76	5.04	8.77	7.18	9.64
5.59	4.32	1.35	5.91	2.34
8.87	8.58	4.14	5.24	9.31
8.52	0.74	5.75	1.69	5.16
6.73	7.25	6.21	2.93	9.58
6.70	5.33	9.32	3.69	4.22
3.63	2.07	6.30	3.89	0.75
9.33	9.80	8.53	5.18	8.12
6.72	3.21	0.11	1.53	7.08
6.79	8.49	5.09	6.68	2.78
8.04	4.90	8.69	3.84	4.89
8.62	6.79	3.67	4.36	0.69
0.96	2.16	6.33	4.25	6.23
0.43	9.14	2.11	8.77	1.69
7.91	5.28	2.14	1.59	4.16
7.94	9.75	4.58	7.35	9.09
5.85	7.86	4.87	2.81	5.91
7.30	4.81	8.52	1.53	3.46
2.38	0.82	1.51	8.22	4.22
2.45	6.85	5.05	0.27	5.35
2.36	10.00	5.23	4.12	8.62
2.46	8.98	3.14	8.67	9.35
8.68	2.95	2.10	7.24	5.87
8.55	5.24	5.40	7.07	7.24
7.10	6.70	2.97	6.58	1.21
4.51	7.72	5.42	5.47	2.19
5.25	9.57	9.30	0.97	4.49
2.06	7.91	7.68	8.60	2.51



На последнем этапе проанализируем гомологическую матрицу на наличие диагональных элементов.

В конечном итоге получаем список повторов:

```
list<Repeat> result;

struct Repeat {
        long x_begin, y_begin;
        long x_end, y_end;
    long length;
}
```

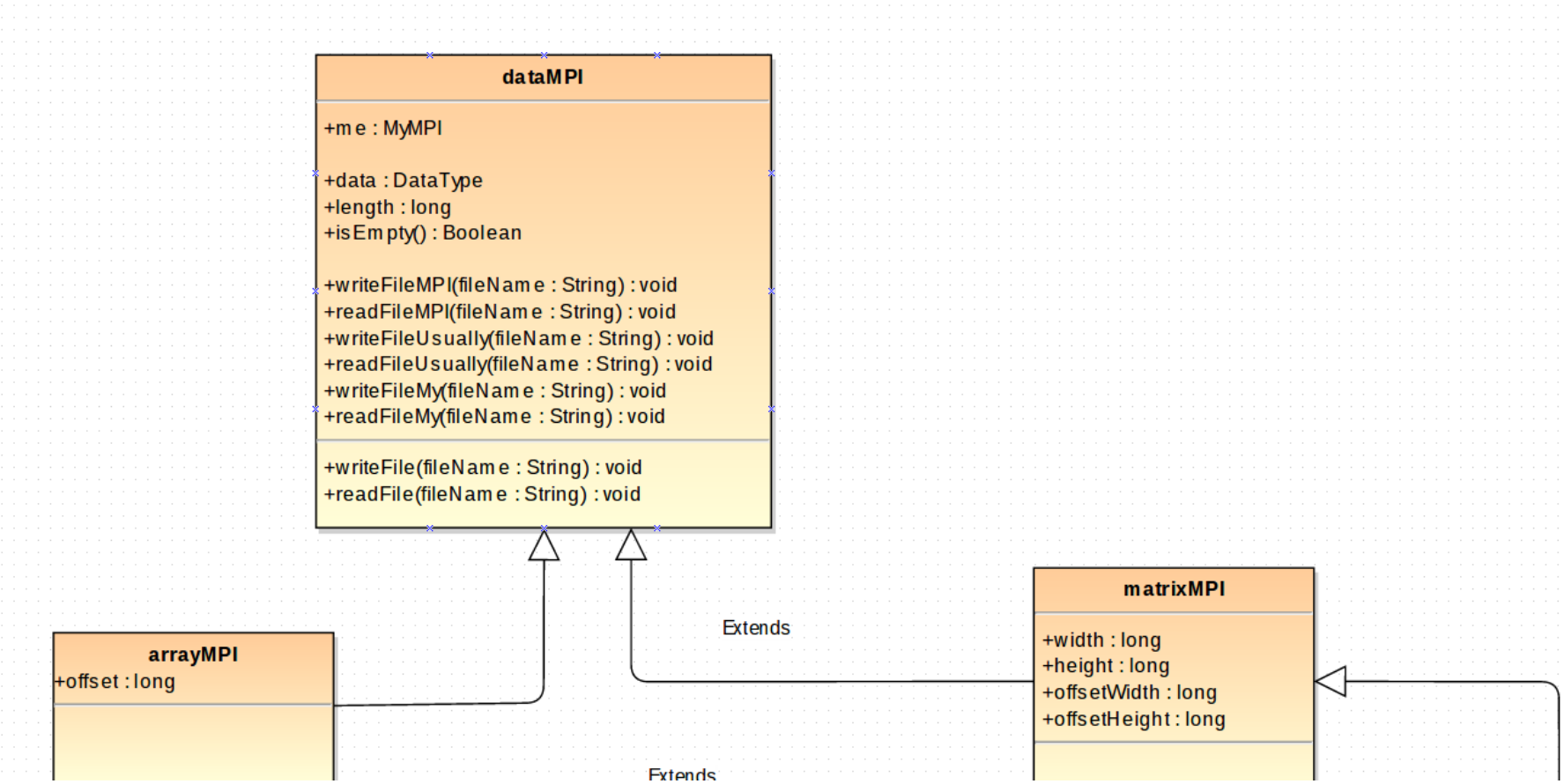
Структура программной реализации параллельного алгоритма

Параллельная реализация программы выполнена на C++ с использованием технологии параллельного программирования MPI и технологии CUDA-C для использования графических ускорителей Nvidia.
Взаимодействие с MPI и CUDA происходит с помощью отдельных классов, что позволяет при желании откомпилировать программу без использования этих технологий.
В стуктуре классов можно выделить следующие группы:

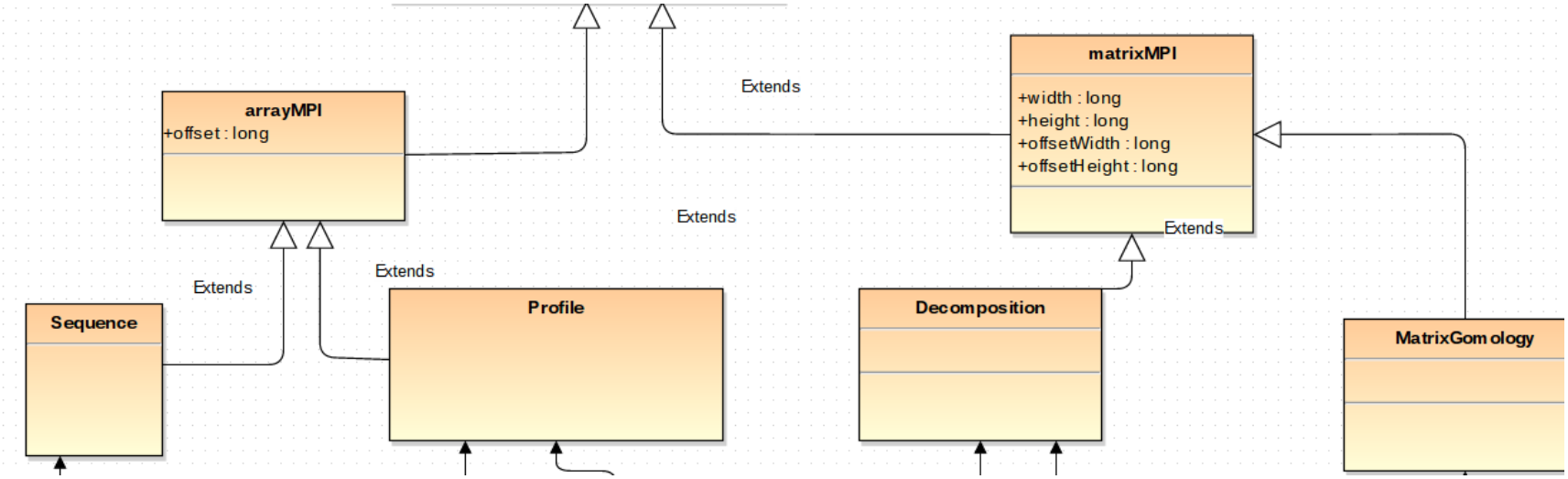
- Классы сущности
- Классы вычислений
- Классы управления

Структура алгоритма позволяет выделить каждый этап в отдельный модуль. Таким образом работу программы можно начать с любого этапа.

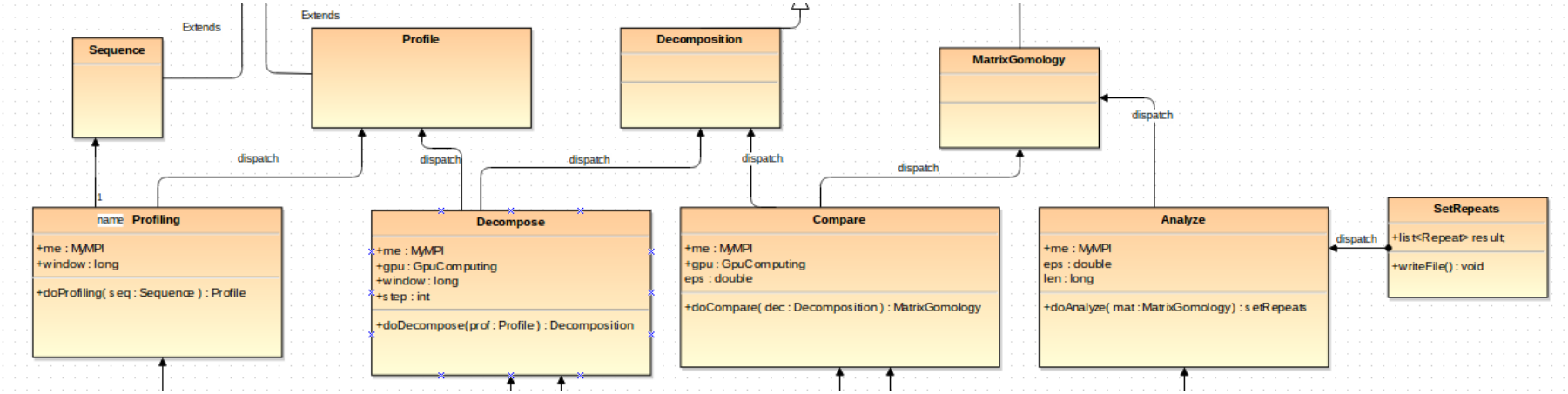
Параллельная версия алгоритма



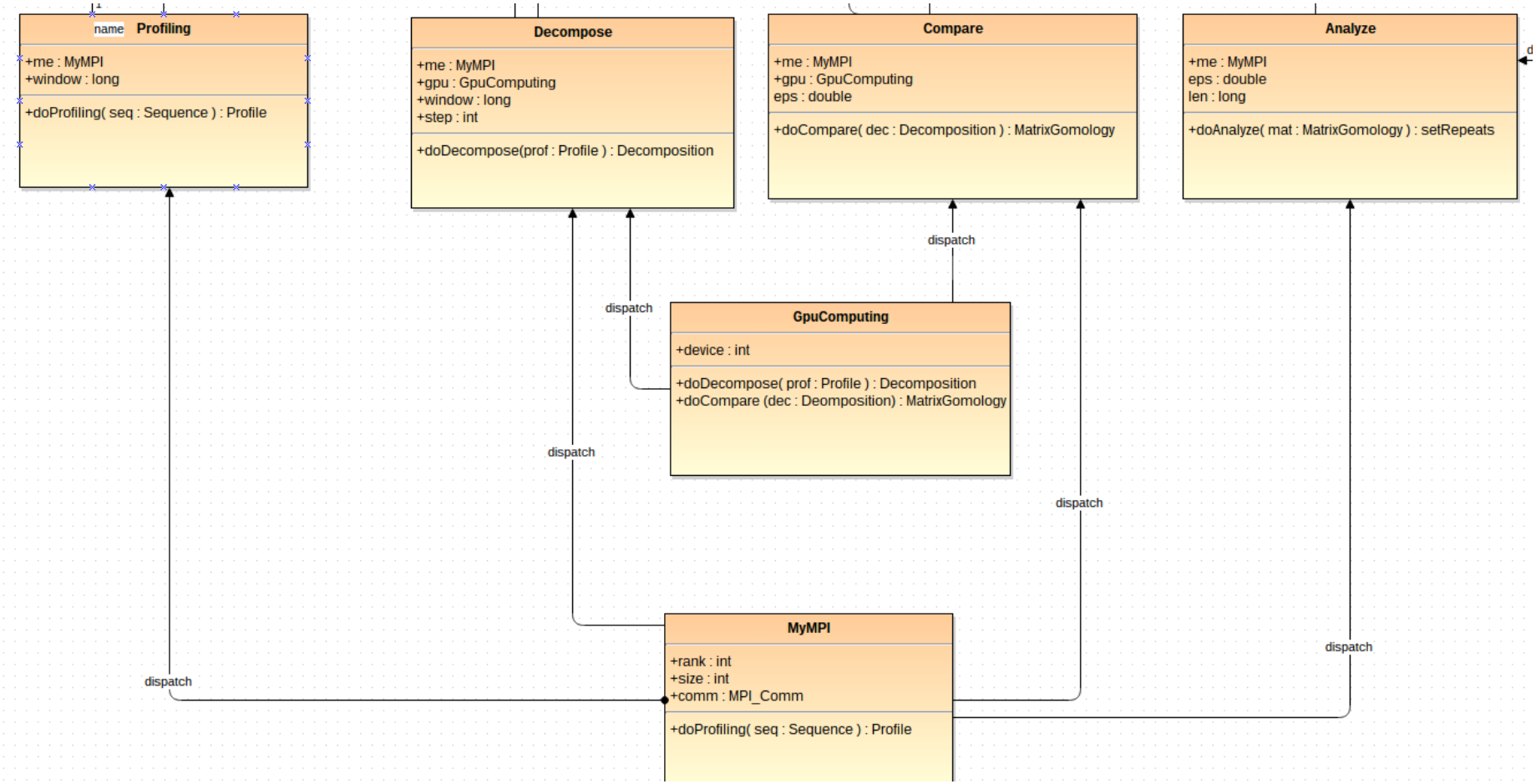
Параллельная версия алгоритма



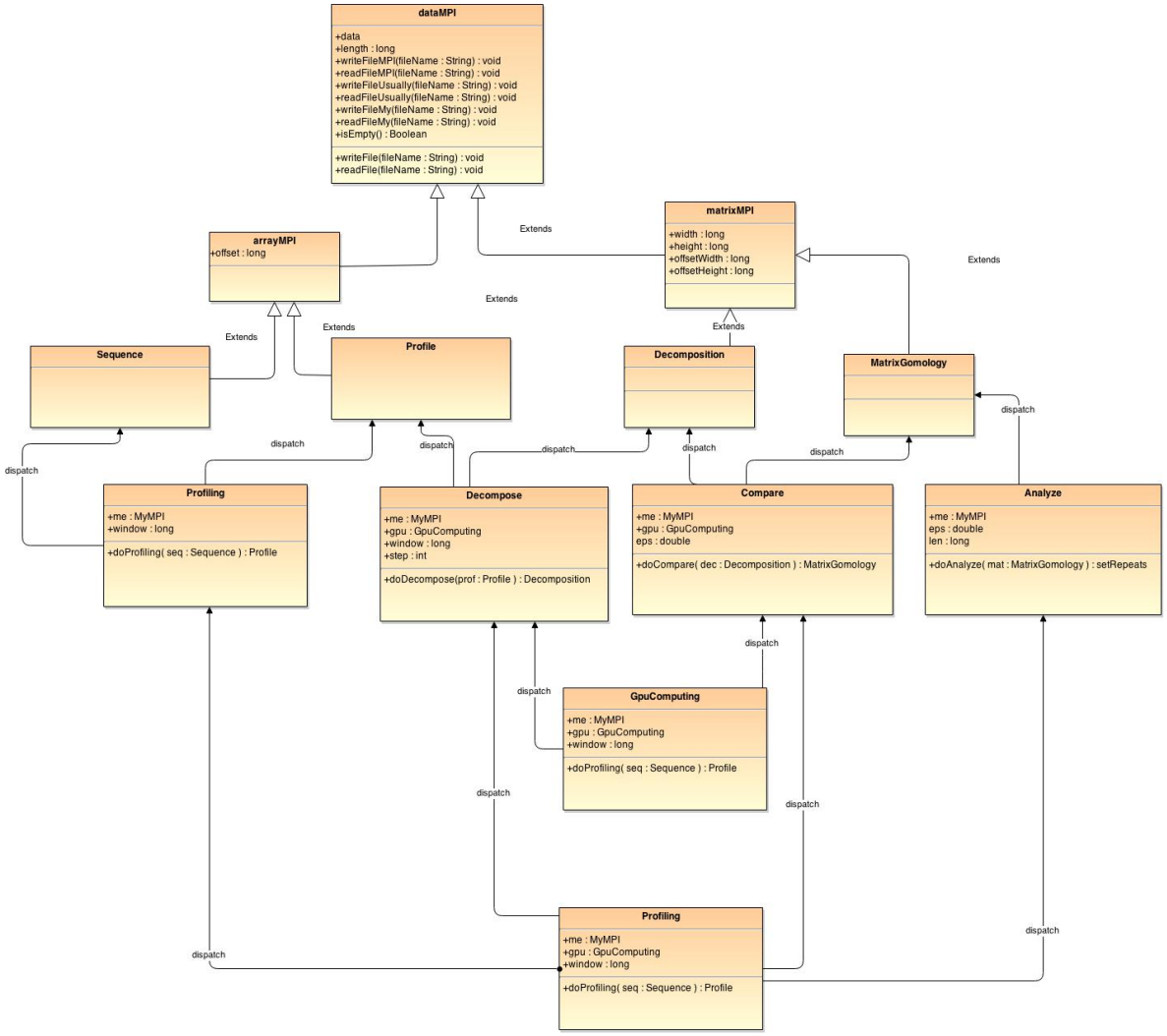
Параллельная версия алгоритма



Параллельная версия алгоритма



Программная реализация параллельного алгоритма



Основные результаты

- Разработан и реализован на суперкопьютерах Ломоносов, BlueGene/P и многопроцессорной рабочей станции параллельный алгоритм поиска повторов в биологических последовательностях. Алгоритм реализует парные сравнения последовательностей большого размера. Параллельная реализация алгоритма выполнена с использованием технологии MPI с возможностью подключения cuda-модулей для использования графических ускорителей.
- Проведено тестирование разработанного алгоритма на искусственных данных, а так же на решении реальной задачи сравнения биологических последовательностей крысы и мыши. Полученные результаты параллельного метода совпадают с последовательной реализацией метода.

Основные результаты

- Проведено исследование эффективности и масштабируемости разработанной параллельной программы.
- Показано, что использование графических процессоров позволяет ускорить алгоритм.
- Разработан графический интерефейс для параллельной программы, позволяющий пользователю запускать программу как на локальном, так и на удаленной системе. Интерфейс реализован с использованием кроссплатформенной библиотекой Qt, что позволяет использовать программу в различных операционных средах.

Основные результаты

- Часть работы с использованием графических процессоров была поддержана компанией Nvidia в конкурсе студенческих работ.