

Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра суперкомпьютеров и квантовой информатики

ТОЙГИЛЬДИН Владислав Петрович

**Разработка и исследование параллельного
алгоритма поиска неточных повторов в геноме.**

ДИПЛОМНАЯ РАБОТА

Научный руководитель
к.ф.-м.н., доцент
Н.Н. Попова

Москва, 2015

Аннотация

Русская аннотация

Аннотация

Английская аннотация

Содержание

Введение	2
1 Исследование и построение решения	4
1.1 Математическая модель поиска повторов в биологических последовательностях	4
1.2 Спектрально-аналитический метод поиска повторов	5
1.2.1 Профили последовательностей ДНК	5
1.2.2 Спектральное индексирование профилей	5
1.2.3 Спектральное сравнение профилей	6
1.2.4 Построение матрицы гомологии	6
1.3 Параллельный метод поиска повторов	6
2 Структура программной реализации	8
2.1 Графические интерфейсы	8
3 Результаты вычислительных экспериментов	9
3.1 Сравнение с GPU	9
3.2 Сравнение блочного метода с матричным	9
3.3 Масштабируемость	9
3.4 Правильность полученных результатов	9
Заключение	10
Список литературы	11

Введение

В последние десятилетия мы можем наблюдать тенденцию замены привычных экспериментов на вычислительные, что стало возможным благодаря развитию ЭВМ. Тому есть две причины: экономический фактор и возможность ускорения экспериментов. Все это привело к тому, что методы математического анализа находят применение в областях изначально очень далеких от точных наук.

После открытия структуры ДНК и ее роли в передаче наследственной информации в середине XX века получает свое бурное развитие молекулярная генетика. Почти сразу же стало ясно, что для анализа структуры биологических последовательностей нужны принципиально другие методы, отличные от естественно-научных, применяющихся в биологии. Так зародилась биоинформатика - наука, образовавшаяся на стыке молекулярной биологии, генетики, математики и компьютерных технологий, основной целью которой является разработка вычислительных алгоритмов для анализа и систематизации данных о структуре и функциях биологических последовательностей.

С появлением в начале XXI века новых методов секвенирования (расшифровки биологических последовательностей) резко увеличился объем генетической информации. Рост банков данных носит экспоненциальный характер и есть все основания предполагать, что в ближайшее время этот процесс только усилится. Увеличивающийся объем данных открывает возможности для проведения полномасштабных исследований на уровне целых геномов, однако это требует новых подходов. Дело в том, что почти все существующие алгоритмы обработки биологических последовательностей - это адаптированные алгоритмы обработки текстовых строк, т.к. биологические данные в расшифрованном виде представляют собой последовательности символов. Однако данные алгоритмы изначально не учитывают мутационных процессов, таких как вставка, замена, делеция. Учет же таких точечных мутаций является дорогой с вычислительной точки зрения операцией, вносящей существенную нелинейность в подобные алгоритмы, что значительно увеличивает время их выполнения и делает непригодными для практических целей.

Все это привело к тому, что существующие вычислительные мощности не удовлетворяют потребностям биологов. Становится актуальной разработка программных средств, позволяющих быстро обрабатывать большие биологические данные. Данная дипломная работа посвящена параллельным методам анализа и обработки биологических последовательностей на суперкомпьютерах с целью сокращения времени обработки и увеличения объема обрабатываемых данных.

Одной из частных задач молекулярной генетики является поиск повторяющихся элементов, изучение их структуры и распределения в биологических последовательностях. Повторы играют важную роль в функционировании организма, т.к. составляют значительную часть генома, возможно участвуют в его реорганизации и при попадании в кодирующие области повторы могут быть причиной нарушения функций этих геномов, что ведет к развитию заболеваний. Таким образом повторы могут использоваться для диагностики генетических заболеваний и определения родства организмов.

Существует масса алгоритмов, ориентированных на поиск относительно точных или коротких (до 500 н.п.) повторов. Однако именно протяженные (от 1000 н.п.) повторы могут

служить источником для фундаментальных исследований при решении эволюционных и филогенетических задач - определении родства групп организмов на генном уровне. Для полногеномного сравнения на практике часто применяется ДНК-ДНК гибридизация - биохимическая реакция, протекающая с участием ДНК двух организмов, проводимая для получения количественной оценки их схожести. Однако данный метод очень дорогостоящий, долгий и недостаточно точный.

В данной работе рассматривается метод поиска протяженных неточных повторов, разработанный коллективом кафедры математических методов прогнозирования и института математических проблем в биологии Российской академии наук (Пушино). Данный метод позволяет эффективно решать задачу поиска неточных протяженных повторяющихся структур в генетических текстах. Основное отличие данного метода состоит в переходе от дискретного анализа к непрерывному анализу. Таким образом для задачи дискретной по своей природе, применяется континуальный подход, основанный на приближении непрерывных функций с помощью ортогональных многочленов. Благодаря этому метод решает недостатки дискретного подхода и обладает следующими свойствами:

- Линейная сложность алгоритма
- Высокая степень устойчивости к мутациям
- Возможность распараллеливания алгоритма на кластерных системах

Для выполнения дипломной работы необходимо было решить следующие задачи:

- 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.

Глава 1

Исследование и построение решения

1.1 Математическая модель поиска повторов в биологических последовательностях

Формально **биологическая последовательность**:

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

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

Введем обозначение. **Подпоследовательностью** $X|_i^k$ последовательности X называется часть последовательности X с элемента с номером i длиной k при условии $i + k \leq |X|$:

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

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

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

где:

- $\rho(X_1, X_2)$ - 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 расстояния редактирования**
- K - длина последовательностей

Пусть есть две 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 $\{i_x, i_y, k\}$, $i_x, i_y, k \in \mathbb{N}$, таких что:

$$\begin{aligned} i_x + k &\leq N_x \\ i_y + k &\leq N_y \\ (X|_{i_x}^k, Y|_{i_y}^k) &\text{ - повтор длины } k \end{aligned}$$

То есть задача найти все такие подпоследовательности последовательностей X и Y , что эти подпоследовательности будут повторами.

1.2 Спектрально-аналитический метод поиска повторов

В работе исследуется спектральный метод поиска повторов в последовательностях ДНК предложенный и разработанный коллективом кафедры математических методов прогнозирования и института математических проблем в биологии Российской академии наук (Пушино).

Метод разбивается на четыре основных этапа:

1. Получение профилей последовательностей
2. Спектральное индексирование полученных профилей
3. Сравнение коэффициентов спектрального разложения
4. Формирование матрицы гомологии

1.2.1 Профили последовательностей ДНК

Одной из главных особенностей рассматриваемого метода поиска повторов является то, что на первом этапе последовательность преобразуется из дискретной в непрерывную область. Это достигается построением т.н. профилей.

Под GC-профилем последовательности $X = (x_n)_{n=1}^{N_x}$ с окном w в дальнейшем будем понимать такую последовательность $P_{GC}(X, w) = (p_n^{GC})_{n=1}^{N_p}$, $N_p = N_x - w + 1$, что

$$p_i^{GC} = \sum_{k=i}^{i+w} I^{GC}(x_k), i = \overline{1, N_p}$$

где

$$I^{GC} = \begin{cases} 1, & x_n \in \{G, C\} \\ 0, & \text{иначе} \end{cases}, n = \overline{1, N_x}$$

Понятие GA-профиля определяется аналогично:

$$P_{GA}(X, w) = (p_n^{GA})_{n=1}^{N_p} : p_i^{GA} = \sum_{k=i}^{i+w} I^{GA}(x_k), i = \overline{1, N_p},$$

где

$$I^{GA} = \begin{cases} 1, & x_n \in \{G, C\} \\ 0, & \text{иначе} \end{cases}, n = \overline{1, N_x}$$

1.2.2 Спектральное индексирование профилей

На этом этапе профили переводятся в спектральное представление с использованием в качестве базиса полиномов Чебышева дискретного аргумента или функций Фурье.

Под спектральным представлением сигнала $P = (p_n)_{n=1}^{N_p}$ будем понимать вектор $\overline{C} = C_m(P) = (c_0, \dots, c_{m-1})$, где c_0, \dots, c_{m-1} - первые m коэффициентов разложения сигнала P по некоторой системе ортогональных функций $u_0(x), \dots, u_{m-1}(x), \dots$. В случае использования в качестве базиса функций Фурье имеем:

$$\begin{aligned} u_0(x) &= 1, \\ u_1(x) &= \cos x, u_2(x) = \sin x, \\ u_3(x) &= \cos 2x, u_4(x) = \sin 2x, \\ &\dots \end{aligned}$$

Полиномы Чебышева определяются при помощи рекуррентного соотношения:

$$\begin{aligned} u_0(x) &= 1, \\ u_1(x) &= x, \\ u_{n+1}(x) &= 2xu_n(x) - u_{n-1}(x) \end{aligned}$$

Вычисление коэффициентов разложения в обоих случаях выполняется по рекуррентным соотношениям

Рекуррентная схема вычисления коэффициентов разложения профилей позволяет хорошо использовать кэш-память процессора, что оказывает существенное влияние на производительность алгоритма.

1.2.3 Спектральное сравнение профилей

На этой стадии спектральное представление профилей используется для производства сравнения на основе некоторого специально разработанного критерия.

Теперь можно ввести понятие расстояния редактирования для рассматриваемого метода поиска повторов.

Пусть X_1, X_2 - две последовательности ДНК, такие что $|X_1| = |X_2| = K \geq w$. Будем понимать под GC- и GA-расстоянием редактирования для X_1 и X_2 :

$$\begin{aligned} \rho^{GC}(X_1, X_2) &= \left\| \overline{C_1^{GC}} - \overline{C_2^{GC}} \right\|, \\ \rho^{GA}(X_1, X_2) &= \left\| \overline{C_1^{GA}} - \overline{C_2^{GA}} \right\|, \\ \|\overline{C}\| &= \sum_{i=0}^{m-1} c_i^2 \end{aligned}$$

под повтором будем понимать пару последовательностей X_1, X_2 , удовлетворяющих системе

$$\begin{cases} \rho^{GC}(X_1, X_2) < \epsilon \\ \rho^{GA}(X_1, X_2) < \epsilon \end{cases}$$

1.2.4 Построение матрицы гомологии

Под матрицей гомологии окном профилей w , окном аппроксимации a и шагом аппроксимации s для последовательностей X_1, X_2 будем понимать матрицу

$$\begin{aligned} M(X, Y, w, s, a) &= (m_{ij})^{L_x * L_y}, \\ L_x &= \left\lceil \frac{N_x - a - w + 1}{s} \right\rceil, L_y = \left\lceil \frac{N_y - a - w + 1}{s} \right\rceil \end{aligned}$$

$$m_{ij} = \begin{cases} 1, & x_n \in \{G, C\} \\ 0, & \text{иначе} \end{cases}$$

На рисунке 1.1 приведена гомологическая матрица, образованная в результате сравнения последовательности с самой собой. Матрица имеет симметричный вид.

1.3 Параллельный метод поиска повторов

И здесь тоже кое-чего да написано.

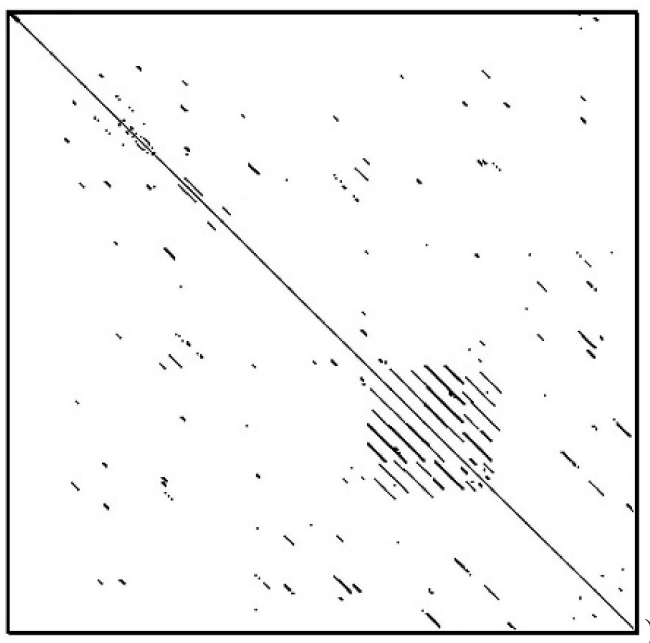


Рисунок 1.1: Пример гомологической матрицы

Глава 2

Структура программной реализации

2.1 Графические интерфейсы

Глава 3

Результаты вычислительных экспериментов

3.1 Сравнение с GPU

3.2 Сравнение блочного метода с матричным

3.3 Масштабируемость

3.4 Правильность полученных результатов

Заключение

Список литературы

- [1] *Автор*. Название книги / Ed. by Редактор. — Издательство, 2012.
- [2] *Автор*. название тезисов конференции // Название сборника. — 2012.
- [3] Название буклета.
- [4] Название статьи / Автор1, Автор2, Автор3, Автор4 // *Журнал*. — 2012. — Vol. 1. — P. 100.
- [5] “this is english article” / Author1, Author2, Author3, Author4 // *Journal*. — 2012. — Vol. 2. — P. 200.