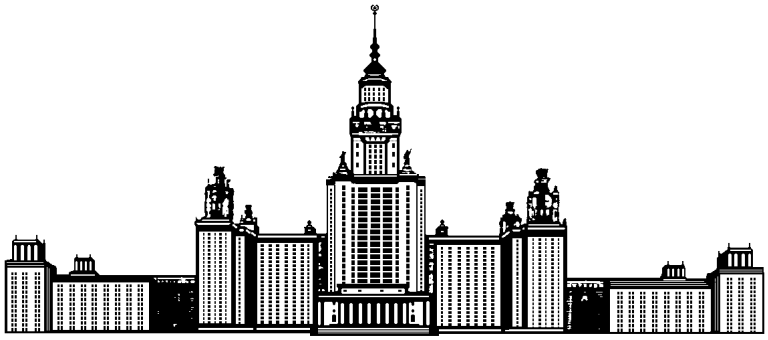

Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра суперкомпьютеров и квантовой информатики

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

Классификация живых организмов всегда была одной из основных задач биологии. Наблюдения за животными и анатомический подход неизбежно приводили к ошибкам в классификации, которые сейчас можно исправить благодаря развитию современных технологий. Одной из таких технологий стала ДНК-ДНК гибридизация - биохимическая реакция, протекающая с участием ДНК двух организмов для оценки их схожести. Но, к сожалению, такая технология очень дорогостоящая и недостаточно точная.

В последние десятилетия, мы можем наблюдать тенденцию замены экспериментов их моделированием, что стало возможным используя вычисления на ЭВМ. Тому есть две причины: экономический фактор и возможность ускорения экспериментов. 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 молекул, прежде всего нуклеиновых кислот и белков. Одной из важных задач биоинформатики стала задача полногеномного сравнения ДНК последовательностей и оценки их схожести.

Отдельного внимания заслуживает подзадача поиска повторов, т.к. повторы составляют до 50% нашего генома. Существует два вида повторов - разнесенные и тандемные. Нас интересуют тандемные повторы - повторы, образовавшиеся в результате дублирования фрагментов ДНК, когда копии фрагментов следуют точно друг за другом. В зависимости от размера образца выделяют 4 вида тандемных повторов: мегасателлиты (длина свыше 1000 нуклеотидных пар), сателлиты (от 100 до 1000 н.п.), минисателлиты (от 7 до 100 н.п.) и микросателлиты (до 6 н.п.). Также повторы различаются по уровню сохранности образца. Дело в том, что со временем копированные образцы подвергаются мутациям, каждая копия дивергирует и тандемный повтор становится размытым. Поэтому выделяют три типа повторов по степени отличия от образца: точные, несовершенные и размытые.

Практическая значимость тандемных повторов выходит из работ, доказывающих, что тансоноспецифичность определенных повторяющихся участков генома позволяет уточнить существующую классификацию организмов, используя геномные данные.

В настоящее время существуют алгоритмы, позволяющие успешно находить тандемные повторы небольших размеров, используя спектрально-статистические подходы. Но до сих пор не существует специализированного метода для поиска мегасателлитных размытых повторов. А ведь именно мегасателлитные повторы играют фундаментальную роль в передаче по наследству генетических болезней и эволюции генома, что делает задачу нахождения тандемных мегасателлитных размытых повторов особенно актуальной.

Такое положение дел в последнее время наметилось из-за отставания процесса вычислительного анализа генетических данных от стремительно развивающейся экспериментальной базы в современной биологии. Т.к. геном можно представить текстовой строкой из 4-х буквенного алфавита, то почти все алгоритмы обработки генетических последовательностей были получены путем простой адаптации алгоритмов обработки текстовой информации к условиям генетических текстов. В связи с этим, временная сложность алгоритмов существенно нелинейна и при увеличении масштаба сравнения и объёмов данных наблюдается резкое снижение эффективности таких алгоритмов. Главным замедляющим фактором при сравнении схожих

генетических данных являются мутации, "исправление" которых существенно увеличивает время анализа.

В данной работе рассматривается метод поиска тандемных протяженных размытых повторов, разработанный коллективом разработчиков кафедры математических методов прогнозирования и института математических проблем в биологии Российской академии наук (Пущино). 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 подохода и обладает следующими свойствами:

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

Целью курсовой работы является разработка параллельного алгоритма поиска тандемных протяженных размытых повторов и анализ его эффективности. Исходя из цели, можно выделить следующие задачи, поставленные в курсовой работе:

1. Изучение предложенного метода поиска повторов.
2. Разработка и реализация параллельного алгоритма.
3. Исследование масштабируемости алгоритма.

Глава 1

Исследование и построение решения

1.1 Математическая модель поиска повторов в биологических последовательностях

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.

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

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

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

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

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 растояния редактирования**
- K - длина последовательностей

Пусть есть две последовательности $X = (x_n)_{n=1}^{N_x}$ и $Y = (y_n)_{n=1}^{Nk_y}$.
Под **задачей поиска повторов** будем понимать нахождение всех троек $\{i_x, i_y, k\},\ i_x, i_y, k \in \mathbb{N}$, таких что:

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

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 (Пущино).

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 $\overline{C} = C_m(P) = (c_0, ..., c_{m-1})$, где $c_0, ..., c_{m-1}$ - первые m коэффициентов разложения сигнала P по некоторой системе ортогональных функций $u_0(x), ..., u_{m-1}(x), ...$ В случае использования в качестве базиса функций Фурье имеем:

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

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:

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

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

$$M(X, Y, w, s, a) = (m_{ij})^{L_x * L_y},$$

$$L_x = [\tfrac{N_x - a - w + 1}{s}], L_y = [\tfrac{N_y - a - w + 1}{s}]$$

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

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.

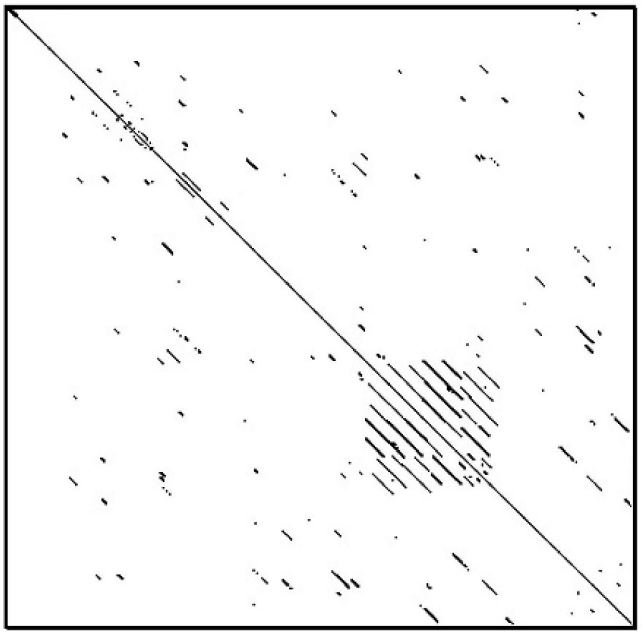

Рисунок 1.1: Пример гомологической матрицы

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 полученнных 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.