РАСПАРАЛЛЕЛИВАНИЕ СПЕКТРАЛЬНОГО АЛГОРИТМА ПОИСКА ПОВТОРОВ В ГЕНОМНЫХ ПОСЛЕДОВАТЕЛЬНОСТЯХ

М.И. Пятков

Введение

Разработка новых методов клонирования и определения последовательности оснований (секвенирования) нуклеиновых кислот положила начало новому этапу развития молекулярной биологии. Процесс секвенирования стал более дешевым и быстрым, предоставляя более широкие возможности к обработке и анализу генетических последовательностей.

Выявление и анализ закодированных в последовательностях функциональных сигналов требует применения современных методов информатики – разработки и оптимизации алгоритмов с использованием всех возможностей процессоров таких как многопоточность и векторные команды.

Одной из актуальных задач бионформатики является исследование повторяющихся элементов(повторов) в геноме человека и изучение их структуры. Количество повторяющихся последовательностей в геноме человека составляет около 50% всего генома, тогда как количество генов, кодирующих белки примерно на порядок меньше.

В данной работе предлагается алгоритм поиска длинных разнесенных повторов. Лежащий в основе алгоритма обобщенный спектрально-аналитический метод, позволяет значительно ускорить процесс анализа последовательности за счет применения средств распараллеливания и векторизации. Также предлагается матрица спектральной схожести генетических последовательностей. Близкая к точечной матрице гомологии, она предоставляет более быстрый, чем дот-матрица, инструмент для сравнительного анализа и визуализации внутренней структуры больших отрезков геномов (порядка 10e6 нуклеотидов), их тандемных и разнесенных повторов.

Общая схема алгоритма

На изображении ниже представлена общая схема работы алгоритма распознавания повторов в генетических последовательностях.

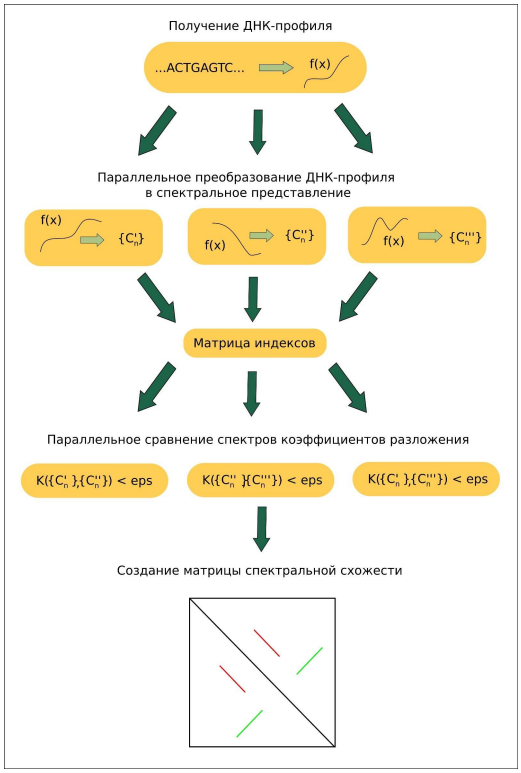

Рис. 1

Далее каждый шаг рассмотрен более подробно.

Получение ДНК-профиля

В первую очередь необходимо перевести ДНК-последовательность, представленную в виде набора символов-нуклеотидов (A, T, G, C), в числовую функцию, которую в дальнейшем мы и будем анализировать. Рассмотрим генетическую последовательность, как последовательность символов $s_i=\{A,T,G,C\}$, имеет смысл разделить символы на две группы $\{A,T\}$ и $\{G,C\}$ по биологическому свойству (комплементарности). В соответсвии с эти разделением возникает функция кодирования:

$$f(s_i)=\begin{cases}0, & s_i\in\{A,T\}\\ 1, & s_i\in\{G,C\}\end{cases}$$

Мы рассматриваем представление дискретной последовательности s_i на протяженном участке сигнала p_j,

$$p(j)=\sum_{i=j}^{j+n} f(s_i),$$

где n — длина скользящего окна, которое является параметром алгоритма. Сигнал p_j можно назвать профилем $\{G,C\}$ содержания в последовательности s_i. Для того, чтобы полностью описать последовательность мы используем два профиля содержания по $\{G,C\}$ и $\{G,A\}$ нуклеотидам, это в дальнейшем позволяет отфильтровать шум на матрице спектральной схожести.

Преобразование ДНК-профиля в спектральное представление

Следующим шагом является перевод ДНК-профиля спектральное представление. На профиле p_j выбирается некоторое окно w с длиной l, после это окно смещается на шаг k и таким образом покрывается весь профиль p_j. На каждый шаг происходит преобразование интервала лежащего в окне w в спектральное представление путем разложения по базису Чебышева дискретного аргумента, по реккурентной формуле:

$$(n+1)\theta_{n+1}(t)=(2\mathrm{n}+1)(2\mathrm{t}-T+1)\theta_n(t)-n(T^2-n^2)\theta_{n-1}(t),$$

формула для получения коэффициентов разложения выглядит следующим образом:

$A_n=\sum_{t=0}^{T-1} f(t)\theta_n(t)$, где $f(t)$ - функция дискретного аргумента в окне w.

Для каждого интервала мы получаем спектры коэффициентов разложения, которые индексируются в виде матрицы для дальнейшего сравнения.

Сравнение спектров коэффициентов разложения

Для сравнения спектров коэффициентов в матрице индексов, применяется метрика, основанная на неравенстве треугольника. После преобразований формулы, критерий приобретает вид, оптимальный для наших целей:

$$\theta=\frac{\|f-g\|}{\|f\|+\|g\|}$$

где f и g, некоторые функции. Для работы с коэффициентами критерий приводится к следующему виду:

$$\theta_N=\frac{\|\sum_{n=0}^{N} A_n\phi_n-\sum_{n=0}^{N} B_n\phi_n\|}{\|\sum_{n=0}^{N_{max}} A_n\phi_n\|-\|\sum_{n=0}^{N_{max}} B_n\phi_n\|}$$

$$\theta_N\le\theta_{N+1}<\varepsilon$$

Монотонность метрики позволяет не обрабатывать полностью заведомо разные спектры, то есть, если уже при сравнении первых коэффициентов, значение θ превышает заданный ε, алгоритм переходит к следующей паре векторов.

Особенности базиса позволяют очень быстро искать инвертированные спектры коэффициентов, путем замены знака нечетных коэффицентов на противоположный у одного из сравниваемых спектров.

$$f(x) = -g(x) \quad , \quad \begin{cases} C^f_{2k} = C^g_{2k} \\ C^f_{2k+1} = -C^g_{2k+1} \end{cases}$$

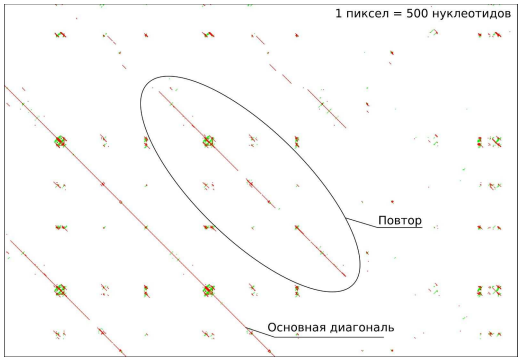

Рис. 2

Получение и анализ матрицы спектральной схожести

Результаты сравнения спектров записываются в матрицу спектральной схожести. Графически это выглядит так: прямые повторы изображены красным цветом и параллельны основной диагонали, инвертированые повторы изображаются зеленым и перпендикулярны основной диагонали.

После того как матрица построена, она автоматически анализируется для получения координат повторов.

Оптимизация

При реализации алгоритма в виде программы учитывалась необходимость использования современных возможностей процессоров – в частности SIMD(Single Instruction, Multiple Data) и многопоточности. Особенности математического аппарата лежащего в основе алгоритма, позволяют значительно ускорить вычислительную часть программы. Блок преобразования ДНК-профиля в спектральное представление является вычислением соответсвующих интегралов коэффициентов разложения, при этом отсутсвует зависимость по данным, что позволяет успешно распараллелить данный блок, почти идеально, на количество потоков по количеству ядер с эквивалентным ускорением. В данном блоке были протестированы два подхода вычисления коэффициентов разложения. Первый, требует больше оперативной памяти, т.к. предварительно индексирует в матрице коэффициенты одной последовательности, достоинство такого подхода – отсутствие необходимости производить большой объем вычислений во вложенном цикле, как недостаток – увеличение требований к пропускной способности памяти, что особенно важно при массивно-параллельных вычислениях, так как отдельные узлы кластера или ядра могут вообще не иметь общего доступа ко всей оперативаной памяти системы. Второй способ предполагает вычисление и сравнение векторов коэффициентов на ходу, как плюс – скромные требования на размер оперативной памяти, но минус те же требования на пропускную способность памяти. Так как наши исследования проходили на SMP машине, вполне предсказуемо оказалось, что подход с индексацией оказался эффективнее.

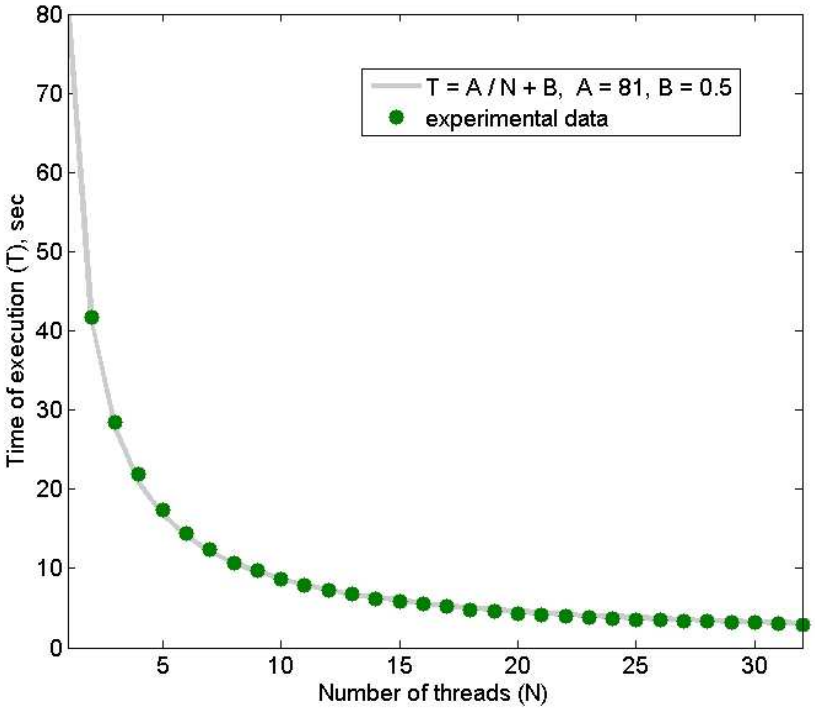

Рис. 3

Для реализации многопоточности использовалась библиотека OpenMP.

Аппаратно-ускоренные и векторные инструкции (SIMD) были реализованы с помощью библиотеки Intel IPP (Intel Performance Primitives), данные оптимизированные функции мы использовали в блоке сравнения спектров коэффициентов разложения.

Стенд на котором мы исследовали эффективность распараллеливания был предоставлен компанией Intel в рамках конкурса Intel Manycore Testing Lab, и представлял собой четырехпроцессорную машину с Intel Xeon X7560, где каждый процессор обладал 8 ядрами без учета Hyperthreding, т.е в сумме реальных 32 ядра, на которых и происходили замеры производительности.

Как можно увидеть из графика представленного выше на 32 ядрах параллельная часть алгоритма с

индексированным подходом, отвечающая за вычисление коэффициентов разложения, показала ускорение приблизительно в 27 раз.

Работа частично поддержана проектами РФФИ № 08-07-00353, №10-01-00609 и компанией Intel.

ЛИТЕРАТУРА:

1. Ф.Ф. Дедус, Л.И. Куликова, С.А. Махортых, Н.Н. Назипова, А.Н. Панкратов, Р.К. Тетуев "Аналитические методы распознавания повторяющихся структур в геномах" //ДАН. 2006. Т. 411. № 5. С. 599-602.
2. Р.К. Тетуев , Н.Н. Назипова , А.Н. Панкратов, Ф.Ф. Дедус "Поиск мегасателлитных тандемных повторов в геномах эукариот по оценке осцилляций кривых GC-содержания" //Математическая биология и биоинформатика. 2010. Т. 5. № 1. С. 30-42.
3. A.N. Pankratov, M.A. Gorchakov, and F.F. Dedus, N.S. Dolotova, L.I. Kulikova, S.A. Makhortykh, N.N. Nazipova, D.A. Novikova, M.M. Olyshevets, M.I. Pyatkov, V.R. Rudnev, R.K. Tetuev, and V.V. Filippov. "Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences" // Pattern Recognition and Image Analysis, 2009, Vol. 19, No. 4, pp. 687–692.
4. М.И. Пятков, А.Н. Панкратов, Р.К. Тетуев, Ф.Ф. Дедус. "Оптимизация спектрального алгоритма распознавания повторяющихся последовательностей в геномах" 14-ая Международная пущинская школа-конференция молодых ученых "Биология – наука XXI века" , Пущино, 19-23 апреля 2010 г., Сборник тезисов, Том №2, с.288-289.