Спектрально-аналитический метод распознавания неточных повторов в символьных последовательностях

[1]*А.Н. Панкратов <pan@impb.ru>*
[1]*Р.К. Тетуев <ruslan.tetuev@gmail.com>*
[1]*М.И. Пятков <mpyatkov@gmail.com>*
[2]*В.П. Тойгильдин <vladislav.toigildin@cs.msu.su>*
[2]*Н.Н. Попова <popova@cs.msu.su>*
[1] *Институт математических проблем биологии РАН,*
142290, Россия, г. Пущино Московской обл., ул. Институтская, дом 4
[2] *Московский государственный университет имени М.В. Ломоносова,*
119991, Россия, Москва, Ленинские горы, д. 1., стр. 52, факультет ВМК

Аннотация. Предложены теоретическое обоснование и алгоритмическая реализация спектрально-аналитического метода распознавания повторов в символьных последовательностях. Теоретическое обоснование основывается на теореме об эквивалентном представлении символьной последовательности вектором непрерывных характеристических функций. Сравнение фрагментов характеристических функций производится в стандартной метрике в евклидовом пространстве коэффициентов разложения рядов Фурье по ортогональным многочленам. Существенным свойством данного подхода является способность оценивать повторы на разных масштабах. Другим важным свойством является возможность эффективного распараллеливания по данным. При разработке алгоритмов предпочиталась схема вычислений с минимальным количеством обращений к оперативной памяти, подразумевающая повторяющиеся и отложенные вычисления. В данной парадигме разработан алгоритм вычисления коэффициентов разложения по ортогональным многочленам за счет использования рекуррентных соотношений. Показано, что алгоритм вычисления коэффициентов разложения по ортогональным многочленам может быть эффективно векторизован за счет вычислений с фиксированной длиной вектора. Распараллеливание и векторизация реализованы с использованием стандарта OpenMP и расширения Cilk Plus языка C/C++. Разработанный метод эффективно масштабируется в зависимости от параметров задачи и числа ядер процессора на системах с общей памятью.

Ключевые слова: спектрально-аналитический метод; ряды Фурье; ортогональные многочлены; рекуррентные соотношения; OpenMP; Cilk Plus

DOI: 10.15514/ISPRAS-2015-27(6)-21

Для цитирования: Панкратов А.Н., Тетуев Р.К., Пятков М.И., Тойгильдин В.П., Попова Н.Н. Спектрально-аналитический метод распознавания неточных повторов в символьных последовательностях. Труды ИСП РАН, том 27, вып. 6, 2015 г., стр. 335-344. DOI: 10.15514/ISPRAS-2015-27(6)-21.

1. Введение

Спектрально-аналитический подход является комбинированным численно-аналитическим методом решения информационных задач, основанным на представлении функций отрезками ортогональных рядов с последующей обработкой в пространстве коэффициентов разложения. Его применение к задаче поиска повторов в биоинформационных последовательностях было показано в работах [1-6].

При разработке программы большое внимание было уделено эффективной реализации предложенного метода, поскольку сравнение с существующими методами поиска повторов, основанными на методах дискретной математики, возможно было на основе законченных программ. Перевод задачи в область математического анализа стимулировал также разработку математического обоснования такой редукции.

2. Описание и обоснование метода

Для адаптации спектрально-аналитического подхода к задачам биоинформатики потребовалось обобщить понятие точечной матрицы. При этом данный подход не теряет общности и может быть применен к поиску повторов в любых символьных последовательностях. В данной работе подход описан в максимально общем виде с сохранением части терминологии и примеров из области биоинформатики.

2.1 Теорема о разложении символьной последовательности

Метод основан на спектральном разложении функций, составляющих характеристическое описание текстовой последовательности, при котором важны следующие свойства функций: 1) полнота и 2) непрерывность. Полнота описания означает, что исходная последовательность может быть восстановлена по характеристическим кривым. Второе свойство необходимо для оценки повторов по форме изменения характеристик. Его выполнение в случае символьных последовательностей обеспечивается тем, что вычисляются кривые содержания подмножеств нуклеотидов в окне заданной длины вдоль последовательности макромолекулы. К этому типу кривых относится хорошо известная и изученная в биоинформатике кривая GC-содержания. При этом размер окна, который является параметром такого описания, вводит фактически понятие масштаба для рассматриваемой символьной последовательности.

Для общего случая сформулируем и докажем следующую **теорему**:

для произвольной символьной последовательности в алфавите из M символов существует $Log_2 M$ характеристических функций, из которых исходная последовательность может быть восстановлена, при этом функции являются - значными, где K – параметр масштаба.

Для доказательства закодируем символы последовательности числовым вектором в двоичной системе счисления. Потребуется не меньше $Log_2 M$ бит. Теперь рассмотрим скользящее окно ширины K, и просуммируем количество единиц определенного бита всех символов последовательности в этом окне. Определенную таким образом функцию, зависящую от начала положения окна в последовательности, назовем характеристической функцией последовательности, соответствующей заданному биту двоичной кодировки символов. Каждый бит каждого символа последовательности можно восстановить из соответствующей ему характеристической функции. При этом значение характеристической функции равно количеству символов, в кодировке которых в соответствующем бите стоит единица, т.е. является функцией содержания некоторого подмножества символов (не менее половины из всего алфавита) в окне, скользящем вдоль последовательности.

Например, в случае геномных последовательностей, заданных в 4-х буквенном алфавите $\{A, T, G, C\}$, можно использовать двухбитную кодировку, достигая таким образом 4-х кратного сжатия геномных файлов, заданных исходно в 8-битной кодировке. При этом в качестве характеристических функций могут выступать кривые содержания нуклеотидов G,C и G,A в скользящем окне длины K [5].

2.2 Структурная схема метода

Характеристические кривые, которые составляют описание объекта, разбиваются на перекрывающиеся фрагменты длины W с шагом d. После этого производится попарно сравнение всех фрагментов f_i, g_i, рассматриваемых как дискретные функции с нумерацией отсчетов в пределах окна длины W, на основе стандартной метрики в евклидовом пространстве:

$$\rho(f,g) = (f-g, f-g) = \frac{1}{W}\sum_{i=1}^{W}(f_i - g_i)^2$$

Для сокращения вычислений расстояний между фрагментами используется аппроксимация фрагментов характеристических функций отрезками ортонормированного ряда. Поэтому оценка расстояния осуществляется по формуле:

$$\rho(f,g) = \sum_{i=1}^{N}(c_i - d_i)^2$$

где c_i, d_i – коэффициенты разложения ряда Фурье, а N – их количество (причем, $N \ll W$). Использование спектрального разложения позволяет не только экономно производить оценку расстояния, но также производить преобразования для оценки инвертированных и комплементарных последовательностей в пространстве коэффициентов разложения, что означает одновременное распознавание всех типов повторов без преобразования самой последовательности [5].

Для распознавания повторов используется пороговое решающее правило: если $\rho < \varepsilon$, где ε - пороговое значение, то фрагменты считаются похожими, а если $\rho \geq \varepsilon$, фрагменты не похожи. При наличии нескольких характеристических кривых, составляющих полное описание объекта, распознавание по ним ведется одновременно, а итоговый результат является логическим умножением решающих правил по каждой из характеристических функций. Такой подход улучшает устойчивость распознавания к ошибкам. Это следует из того, что решающее правило срабатывает в районе минимумов метрики ρ, рассматриваемой как функции от номера фрагмента. Таким образом, множество минимумов определяет множество кандидатов на повтор. В случае двух признаков, например, GC- и GA- кривых, множество повторов берется как пересечение множеств кандидатов на повтор, полученных по каждому из признаков отдельно.

После проведения этих операций результаты сравнения отображаются на точечной матрице, одна точка на которой, однако, соответствует сравнению двух целых фрагментов, а не просто сайтов последовательности. Точечная матрица является одним из наглядных стандартных представлений результатов сравнения двух последовательностей, позволяющим отобразить выравнивание неточных повторов, а также их взаимное расположение. Обобщенная точечная матрица позволяет получить новые возможности для выравнивания неточных повторов. Например, было показано, что неточный протяженный тандемный повтор может быть отображен совершенным квадратом на точечной матрице. Это достигается за счет правильного подбора соотношения между размерами окна и шага окна аппроксимации. На основе этого важного результата построен полностью автоматизированный метод распознавания тандемных повторов и найдены неизвестные ранее повторы.

Структурная схема метода выглядит следующим образом:

1) Предварительная обработка символьной последовательности. На этом этапе происходит формирование исходного алфавита: удаление ненужных символов, перекодировка символов последовательности.
2) Преобразование символьной последовательности в пучок непрерывных характеристических функций на основе доказанной теоремы.
3) Преобразование характеристических функций в спектральное представление. В отличие от предыдущих шагов этот этап подразумевает необратимое сжатие информации.

4) Спектральное сравнение фрагментов последовательностей.
5) Отображение и анализ точечной матрицы с целью выявления протяженных повторов, тандемных повторов и исследования взаимного расположения повторов.
6) Верификация повторов путем выравнивания методами динамического программирования.

3. Эффективная реализация метода

Критически важными с самого начала развития метода были вопросы его вычислительной сложности и эффективной реализации. Все этапы алгоритма являются независимыми друг от друга и хорошо распараллеливаются по данным. Вычислительная сложность всех этапов линейно зависит от длин последовательностей, кроме построения точечной матрицы, которое в общем случае является квадратичным по сложности вычислительным процессом, т.к. зависит от произведения длин анализируемых последовательностей. Однако, введенное функциями содержания естественное понятие масштаба при анализе нуклеотидных последовательностей позволяет утверждать, что построение точечной матрицы фиксированного размера на разных масштабах производится за линейное время в зависимости от длин анализируемых последовательностей [6].

3.1 Основные принципы

При реализации разработанного метода в виде алгоритмов и программ учитывался современный уровень развития вычислительной техники с широким применением параллельных и векторных вычислений, графических ускорителей и спецвычислителей, распределенных и облачных технологий. Для экономии памяти и пересылок между узлами предпочтение отдается вычислительным схемам с минимальным использованием промежуточных результатов вычислений.

Например, при вычислении коэффициентов разложения значения ортогональных многочленов вычисляются по рекуррентным соотношениям и не запоминаются в промежуточных массивах. Это позволяет эффективно использовать кэш процессора, а также приводит к хорошей масштабируемости на многоядерных процессорах и графических ускорителях, так как снижает нагрузку на оперативную память при многопоточных вычислениях.

Другой пример экономии памяти связан с использованием точечной матрицы. Для лучшей масштабируемости приложения в зависимости от числа процессоров и длины последовательности предлагается не сохранять матрицу, а вычислять значения ее элементов по требованию. Это позволяет даже длинные последовательности обрабатывать на одном узле.

3.2 Рекуррентный алгоритм вычисления коэффициентов разложения

Рассмотрим более подробно процесс вычисления коэффициентов разложения по ортогональным многочленам на примере многочленов Чебышева. Коэффициенты разложения функции $f(t)$ вычисляются по формулам:

$$c_j = \frac{2}{W}\sum_{i=1}^{W} f(t_i)\,T_j(t_i)$$

где t_i – узлы квадратурной формулы Гаусса, $T_j(t)$ - многочлены Чебышева непрерывного аргумента t, удовлетворяющие рекуррентному соотношению с начальными условиями:

$$T_0(t) = 1, T_1(t) = t, T_{j+1}(t) = 2tT_j(t) - T_{j-1}(t)$$

В [7] представлен алгоритм вычисления коэффициентов разложения на языке C++. Ниже представлена оптимизированная версия этого алгоритма на языке C++11 с расширением Cilk Plus:

```
void chebft(int n, double t[n], double f[n], int m, double c[m], int l){
        double p1[l], p2[l], p3[l];
        c[:] = 0.0;
        for (int i = 0; i < n; i += l){
                if (i > n - l) l = n - i;
                p2[0:l] = t[i:l] * f[i:l] * 2.0 / n;
                p1[0:l] = f[i:l] * 2.0 / n;
                c[0] += __sec_reduce_add(p1[0:l]);
                for (int j = 1; j < m; j++){
                        p3[0:l] = p2[0:l];
                        p2[0:l] = p1[0:l];
                        p1[0:l] = 2.0 * t[i:l] * p2[0:l] - p3[0:l];
                        c[j] += __sec_reduce_add(p1[0:l]);
                }
        }
}
```

В отличие от исходного алгоритма [7] здесь сделаны следующие усовершенствования:

1) применено рекуррентное соотношение для вычисления значений многочленов;

2) умножение на значение функции и нормировка включены в рекуррентное соотношение умножением на начальные условия;

3) произведена векторизация по соседним узлам сетки, l – длина вектора.

В соответствии с внесенными изменениями представленный алгоритм назовем векторно-рекуррентным с фиксированной глубиной векторизации.

Единственные временные переменные в этом алгоритме – это массивы $p1, p2, p3$, размер которых зависит от длины l, значения которой следует выбирать равным длине векторных регистров процессора.

Разработанная вычислительная схема обладает не только высокой эффективностью, но масштабируемостью на многоядерных процессорах. При вычислении коэффициентов разложения от массива функций в задаче поиска повторов достигается практически идеальное масштабирование по числу используемых ядер при распараллеливании с помощью директив OpenMP.

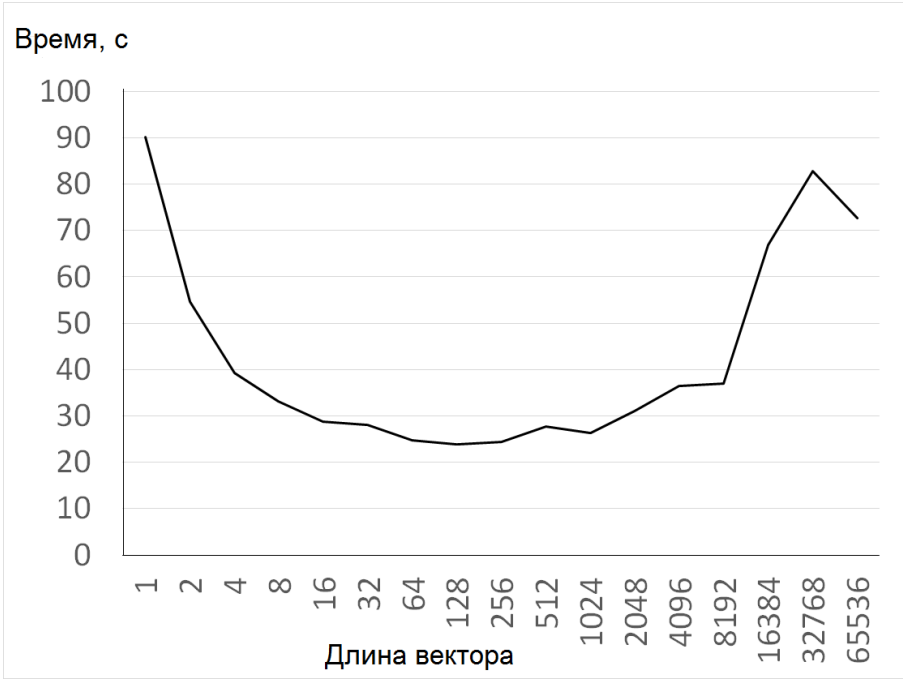

Рис. 1 Время выполнения программы в зависимости от длины вектора.

На рис. 1 представлены результаты расчетов $N = 10^5$ коэффициентов разложения на сетке $W = 10^5$ в зависимости от длины вектора. Вычисления производились на процессоре AMD Phenom. Следует отметить, что при максимальной длине вектора, приближающейся к длине всей сетки, эффект от использования векторных операций процессора практически исчезает. Это происходит из-за того, что в этом случае кэш-память процессора фактически не используется. Кроме того, оптимальным значением глубины векторизации кода для данного процессора является 128, при этом достигается 4-х кратное ускорение за счет утилизации векторных инструкций процессора (SSE, AVX). Таким образом, построен алгоритм, который эффективно масштабируется в

зависимости от длины сетки, т.к. использует фиксированную глубину векторизации.

Заключение

Можно выделить существенные свойства спектрально-аналитического метода, определяющие его эффективность при распознавании неточных повторов:

1) интегральное оценивание повторов, которое позволяет нивелировать локальные неточности в повторах сигнала;
2) выбор масштаба за счет изменения размеров окна и его шага, что позволяет производить гибкое выравнивание неточных повторов;
3) использование спектрального разложения сигналов, которое обуславливает значительное сокращение вычислений;
4) высокая степень распараллеливания и векторизации вычислений.

В данной работе показана принципиальная возможность поиска повторов и высокая вычислительная эффективность предложенных алгоритмов в случае построения точечных матриц. Достигнутые результаты позволяют сделать вывод о возможности дальнейшего совершенствования качества работы метода и его применимости к конкретным задачам. Например, перспективным является вопрос о применимости разработанного метода в задаче поиска неточных повторов по заданному образцу среди множества геномов. Решение этой задачи возможно с привлечением алгоритмического аппарата, связанного с распределенными и облачными вычислениями.

Работа выполняется при поддержке грантов РФФИ №14-07-00654, 14-07-00924, 14-07-31306, 15-29-07063.

Список литературы

[1]. Дедус Ф.Ф., Куликова Л.И., Махортых С.А., Назипова Н.Н., Панкратов А.Н., Тетуев Р.К. Аналитические методы распознавания повторяющихся структур в геномах. Доклады Академии Наук, 2006, т.411, №5, с.599-602, doi: 10.1134/S1064562406060354.

[2]. Дедус Ф.Ф., Куликова Л.И., Махортых С.А., Назипова Н.Н., Панкратов А.Н., Тетуев Р.К. Распознавание структурно-функциональной организации генетических последовательностей. Вестник московского университета. Серия 15: Вычислительная математика и кибернетика, 2007, т.31, №2, с.12-16, doi: 10.3103/S0278641907020021.

[3]. Pankratov A.N., Gorchakov M.A., Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova D.A., Olshevets M.M., Pyatkov M.I., Rudnev V.R., Tetuev R.K., and Filippov V.V. Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences. Pattern Recognition and Image Analysis, 2009, Vol. 19, №4, pp. 687–692, doi: 10.1134/S105466180904018X.

[4]. Тетуев Р.К., Назипова Н.Н., Панкратов А.Н., Дедус Ф.Ф. Поиск мегасателлитных тандемных повторов в геномах эукариот по оценке осцилляций кривых GC-содержания. Математическая биология и биоинформатика, 2010, Т.5, №1, с.30-42, doi: 10.17537/2010.5.30.

[5]. Панкратов А.Н., Пятков М.И., Тетуев Р.К., Назипова Н.Н., Дедус Ф.Ф. Поиск протяженных повторов в геномах на основе спектрально-аналитического метода. Математическая биология и биоинформатика, 2012, Т.7, №2, с.476–492, doi: 10.17537/2012.7.476.

[6]. Pyatkov M.I., Pankratov A.N. SBARS: fast creation of dotplots for DNA sequences on different scales using GA-, GC-content. Bioinformatics, Vol. 30, №12, 2014, pp. 1765–1766, doi: 10.1093/bioinformatics/btu095.

[7]. W.H.Press, S.A.Teukolsky, W.T.Vetterling, B.P.Flannery Numerical Recipes. The Art of Scientific Computing. Third Edition. Cambridge University Press, 2007, 1256 pp.

Spectral Analytical Method of Recognition of Inexact Repeats in Character Sequences

[1]A.N. Pankratov <pan@impb.ru>
[1]R.K. Tetuev <ruslan.tetuev@gmail.com>
[1]M.I. Pyatkov <mpyatkov@gmail.com>
[2]V.P. Toigildin <vladislav.toigildin@cs.msu.su>
[2]N.N. Popova <popova@cs.msu.su>
[1] Institute of Mathematical Problems of Biology, 4 Institutskaya Str., Pushchino, Moscow Region, 142290, Russian Federation
[2]Lomonosov Moscow State University, 2nd Education Building, Faculty CMC, GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation

Abstract. Proposed are theoretical basis and algorithmic implementation of spectral-analytical method of recognition of repeats in character sequences. The theoretical justification is based on the theorem on equivalent representation of the character sequence by the vector of continuous characteristic functions. Comparison of fragments of characteristic functions is performed in the standard metric in Euclidean space of expansion coefficients of the Fourier series of orthogonal polynomials. An essential feature of this approach is the ability to evaluate repeats at different scales. Another important feature is the possibility of efficient parallelization of data. In the development of algorithms we preferred scheme of computing with a minimal amount of references to memory, implying repetitive calculations and evaluations on demand. In this paradigm, proposed is an algorithm for calculating the coefficients of expansions in the orthogonal polynomials through the use of recurrence relations. It is shown that the algorithm for calculating the coefficients of expansions in the orthogonal polynomials can be effectively vectorized by computing with a fixed vector length. Parallelization and vectorization implemented using the OpenMP standard and extension Cilk Plus of language C/C++. The developed method effectively scales, depending on the parameters of the problem and the number of processor cores on systems with shared memory.

Keywords: spectral-analytical method; Fourier series; orthogonal polynomials; recurrence relations; OpenMP; Cilk Plus

DOI: 10.15514/ISPRAS-2015-27(6)-21

For citation: Pankratov A.N., Tetuev R.K., Pyatkov M.I., Toigildin V.P., Popova N.N. Spectral Analytical Method of Recognition of Inexact Repeats in Character Sequences. Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 6, 2015, pp. 335-344 (in Russian). DOI: 10.15514/ISPRAS-2015-27(6)-21

References

[1]. Dedus F.F., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Pankratov A.N. and Tetuev R.K. Analytical Recognition Methods for Repeated Structures in Genomes. Doklady Mathematics, 2006, Vol. 74, №3, pp. 926-929, doi: 10.1134/S1064562406060354.

[2]. Dedus F.F., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Pankratov A.N., and Tetuev R.K. Recognition of the Structural–Functional Organization of Genetic Sequences. Moscow University Computational Mathematics and Cybernetics, 2007, Vol. 31, No. 2, pp.49–53, doi: 10.3103/S0278641907020021.

[3]. Pankratov A.N., Gorchakov M.A., Dedus F.F., Dolotova N.S., Kulikova L.I., Makhortykh S.A., Nazipova N.N., Novikova D.A., Olshevets M.M., Pyatkov M.I., Rudnev V.R., Tetuev R.K., and Filippov V.V. Spectral Analysis for Identification and Visualization of Repeats in Genetic Sequences. Pattern Recognition and Image Analysis, 2009, Vol. 19, №4, pp. 687–692.

[4]. Tetuev R.K., Nazipova N.N., Pankratov A.N., Dedus F.F. Search for Megasatellite Tandem Repeats in Eukaryotic Genomes by Estimation of GC-content Curve Oscillations. Math. Biol. Bioinf. 2010, 5(1):30-42, doi: 10.17537/2010.5.30.

[5]. Pankratov A.N., Pyatkov M.I., Tetuev R.K., Nazipova N.N., Dedus F.F. Search for Extended Repeats in Genomes Based on the Spectral-Analytical Method. Math. Biol. Bioinf. 2012;7(2):476-492, doi: 10.17537/2012.7.476.

[6]. Pyatkov M.I., Pankratov A.N. SBARS: fast creation of dotplots for DNA sequences on different scales using GA-, GC-content. Bioinformatics, Vol. 30, №12, 2014, pp. 1765–1766, doi: 10.1093/bioinformatics/btu095.

[7]. W.H.Press, S.A.Teukolsky, W.T.Vetterling, B.P.Flannery Numerical Recipes. The Art of Scientific Computing. Third Edition. Cambridge University Press, 2007, 1256 pp.