ДОКЛАДЫ АКАДЕМИИ НАУК, 2006, том 411, № 5, с. 599-602

ИНФОРМАТИКА

УДК 519

АНАЛИТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ ПОВТОРЯЮЩИХСЯ СТРУКТУР В ГЕНОМАХ

© 2006 г. Ф. Ф. Дедус, Л. И. Куликова, С. А. Махортых, Н. Н. Назипова, А. Н. Панкратов, Р. К. Тетуев

Представлено академиком Ю.И. Журавлевым 26.04.2006 г.

Поступило 17.05.2006 г.

ВВЕДЕНИЕ

Данная работа посвящается преодолению серьезных проблем, которые встречаются при разработке новых подходов в задачах распознавания повторяющихся структур в геномах. Современная молекулярная биология стала источником беспрецедентно больших объемов экспериментальных данных, осмысливание которых невозможно без привлечения современных информационных технологий и эффективных математических методов распознавания (анализа) данных и моделирования (прогнозирования поведения) биологических систем и процессов. Таким образом, для современной биологии и генетики требуется участие математиков, в особенности тех, кто имеет опыт в распознавании сигналов при работе с большими массивами данных.

ПОИСК ПОВТОРЯЮЩИХСЯ СТРУКТУР

Геном биологического организма можно представить в виде линейной символьной последовательности, заданной над алфавитом, состоящем из четырех букв: *A*, *T*, *G*, *C*. Для поиска структурной и статистической периодичности в символьных последовательностях применялись методы, основанные на фурье-анализе [1], автокорреляционных функциях, теории информации [2], а также различные методы, основанные на статистических критериях. Эти методы не позволяют проводить поиск периодичности, искаженной вставками и выпадениями символов, что говорит об актуальности разработки новых простых и высокоэффективных методов поиска.

Одним из методов исследования статистических свойств полных геномов и их фрагментов является измерение количества информации (сложности) и избыточности. Для оценки количества информации в геномных последовательностях использовались различные методы: классическая мера информации, алгоритмическая сложность Колмогорова [3], энтропия Шеннона [4], сложность в смысле алгоритмов сжатия (например, алгоритма Лемпеля–Зива [5]). Однако с практической точки зрения ради вычислительного выигрыша, в дальнейшем мы использовали более простую, количественную оценку структуры генома – содержание (состав) определенных нуклеотидов в некотором скользящем окне. Например, для окна длины N, скользящего вдоль ДНК-последовательности $S = \{S_1, S_2, \ldots, S_n, \ldots, S_L\}$, $(G + C)$- и $(G + A)$-составы в каждой i-той позиции представлены значениями функций:

$$f_i^{(G,C)} = \frac{1}{N}\sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = C)\right|,$$

$$f_i^{(G,A)} = \frac{1}{N}\sum_{n=i+1}^{i+N} \left|(S_n = G) \cup (S_n = A)\right|,$$

где $i = 1, 2, \ldots, L - N + 1$.

В данной работе предлагается осуществлять поиск повторяющихся фрагментов как соответствий различных, но схожих участков функций вида

$$f_i = f_i^{(G,C)} \cdot f_i^{(G,A)}.$$

При этом преодоление трудностей, связанных с большими объемами данных, осуществляется в рамках обобщенного спектрально-аналитического метода (ОСАМ) [6] и является следствием адаптивности этого метода. С требуемой точностью полученные функции аппроксимативно представляются в виде:

$$f_i \approx \sum_{j=0}^{K} C_j \varphi_j(t_i)$$

по некоторому из базисов классических ортогональных функций $\{\varphi_j(t)\}$ при $K \ll L - N + 1$.

Используя последнее представление, после преобразования коэффициентов разложения C_j можно

Институт математических проблем биологии Российской Академии наук, Пущино Московской обл.

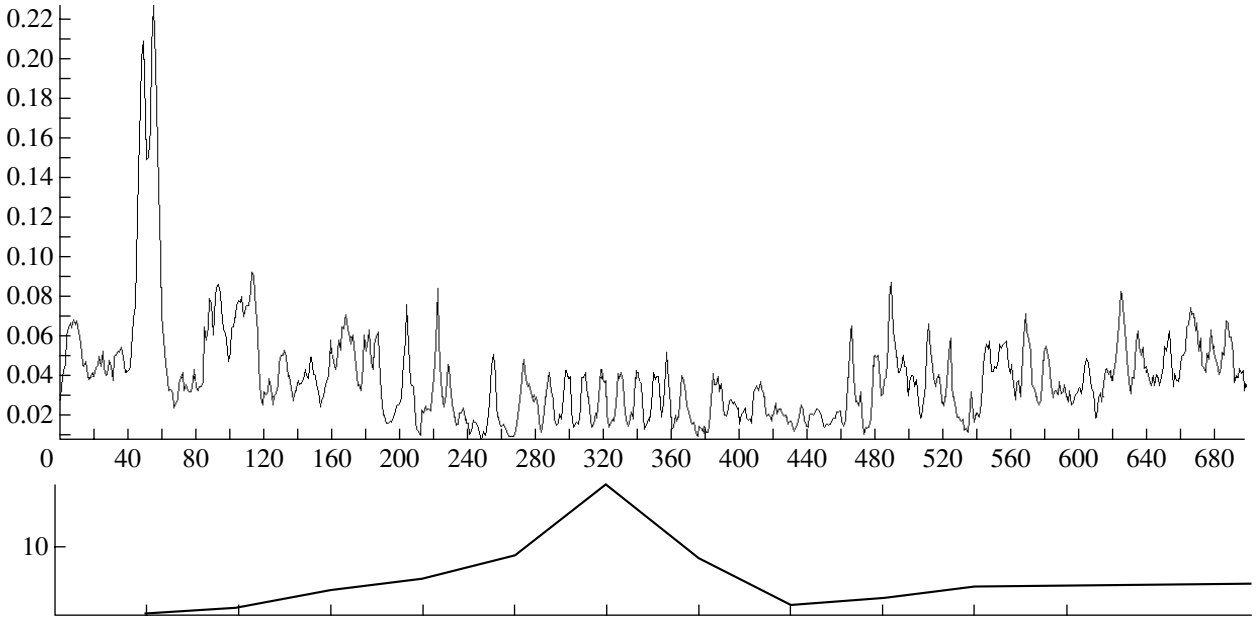

Рис. 1. График функции f_i (верхний) и соответствующего ей функционала оценки степени повторяемости (нижний).

построить некоторую функцию оценки подобия различных фрагментов функции f_i.

В качестве такой функции можно предложить функционал, коэффициенты разложения которого определены коэффициентами C_j, отражающий степень повторяемости функции f_i в окрестности i-ой позиции, как на рис. 1.

Таким образом, мы решаем задачу поиска так называемых тандемных повторов, когда в геноме существуют повторы символов, следующих друг за другом без разрывов.

Используемая нами стратегия поиска периодичностей позволяет искать не только тандемные, но и разнесенные неточные повторы, более просто в сравнении с другими подходами. Иллюстрацией к этому могут послужить обнаруженные в геноме Staphylococcus haemolyticus в ходе численного эксперимента два крупных разнесенных фрагмента: с 2495315-й до 2498579-й и с 2539025-й до 2542289-й позиции. Для поиска использовалась “близость” спектров разложения соответствующих участков кривой на рис. 2.

ВЫЧИСЛИТЕЛЬНАЯ СЛОЖНОСТЬ АЛГОРИТМОВ

Приведем в качестве типового примера сравнение оценок алгоритмической сложности в задаче поиска тандемных повторов в ДНК. Вычисления, основанные на простом переборе всех возможных вариантов точного совпадения повторов, дают квадратичную алгоритмическую сложность $O(n^2)$ относительно длины последовательности n. Часто при малом числе содержащихся в генетической последовательности повторов z относительно n, эта задача проще решаема методом Ландау и Шмидта [7] со сложностью $Q(n\ln n + z)$. Однако если предположить количество возможных несовпадений в повторах равным k, то сложности подобных алгоритмов возрастают соответственно, как $O(kn^2)$ и $O(kn\ln n + z)$.

Реализация предлагаемого выше решения данной задачи приводит к линейной алгоритмической сложности, которая не зависит от предполагаемого числа несовпадений (это сказывается на оценках полученных повторов), причем результаты работы устойчивы к возможным различным нелинейным искажениям входных данных, таким как вставки и выпадения фрагментов текста. Более того, предлагаемый подход подразумевает возможность распараллеливания вычислений, что приведет к дополнительному ускорению поиска тандемных повторов при реализации алгоритмов на вычислительных кластерах.

РЕЗУЛЬТАТЫ

До сих пор не было удовлетворительных методов, которые решали бы задачу обнаружения протяженных неточных тандемных повторов. Нами разработан достаточно эффективный и перспективный метод. Преимущества этого метода, использующего комбинацию численно-аналитических подходов, особенно хорошо выражены при решении подобных задач, учитывающих неточные совпадения и существенные искажения генетического текста.

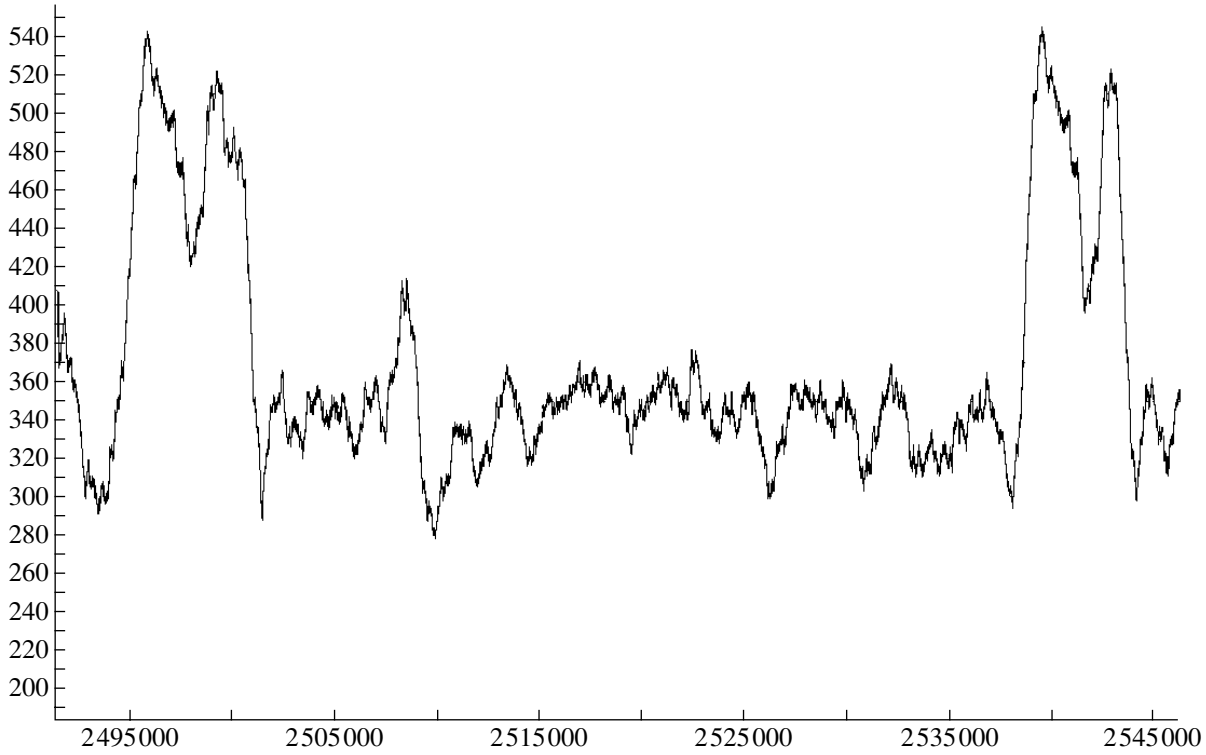

Рис. 2. Кривая изменения ($G + C$)-состава при длине окна в 1000 оснований.

Для апробации предлагаемого метода на основе разработанных алгоритмов был реализован комплекс программ. В качестве тестовой последовательности предложен геном бактерии Staphylococcus haemolyticus (идентификатор нуклеотидной последовательности базы данных GenBank AP006716), для которого уже было известно и ранее зарегистрировано несколько тандемных повторов (табл. 1). Список результатов работы программ оказался значительно более внушительным, чем ожидалось, включая помимо известных повторов целый ряд новых, ранее незарегистрированных в базе данных GenBank (табл. 2). Это позволяет предположить, что предлагаемый авторами подход анализа генетических последовательностей для некоторых задач обладает большей эффективностью в сравнении с ранее развиваемыми.

Первые опыты применения комбинированной вычислительной технологии вселяют уверен-

Таблица 1. Описанные ранее в базе данных GenBank тандемные повторы для Staphylococcus haemolyticus

Координата начала участка повторов	Длина повтора	Кратность повтора
328610	54	155
1182186	231	19
1502207	270	37

Таблица 2. Найденные предлагаемым программным способом тандемные повторы для Staphylococcus haemolyticus

Координата начала участка повторов	Длина повтора	Кратность повтора
44700	36	23
47386	108	15
96100	40	13
328610	54	155
1174866	231	16
1182186	231	19
1207631	141	3
1502207	270	37
1514160	270	8
2424573	72	7
2425484	258	8
2429172	90	7
2494500	274	3

ность в успешном решении поставленных задач распознавания.

Исследование проводится при поддержке РФФИ (гранты 04–01–00756, 04–01–00814, 04–02–17368, 04–07–90402, 06–07–89274, 06–07–89303) и Фонда содействия отечественной науке.

СПИСОК ЛИТЕРАТУРЫ

1. *Sharma D., Issac B., Raghava G.P.S., Ramaswamy R.* // Bioinformatics. 2004. V. 20. № 9. P. 1405–1412.
2. *Korotkov E.V., Korotkova M.A., Kudryashov N.A.* // Phys. Lett. 2003. № 312. P. 198–210.
3. *Колмогоров А.Н.* Теория информации и теория алгоритмов. М.: Наука, 1987. 213 с.
4. *Shannon C. E.* // Bell Syst. Techn. J. 1948. V. 27. P. 379–423.
5. *Lempel A., Ziv J.* // IEEE Trans. Inform. Theory. 1976. V. 24. № 5. P. 530–536.
6. *Дедус Ф.Ф., Махортых С.А., Устинин М.Н., Дедус А.Ф.* Обобщенный спектрально-аналитический метод обработки информационных массивов. М.: Машиностроение, 1999. 357 с.
7. *Gusfield D.* Algorithms on String, Trees, and Sequences. N.Y.: Cambridge Univ. Press, 1997. P. 255–257.