Поиск комбинированных структур в ДНК-последовательностях

Гусев В. Д., Мирошниченко Л. А.
luba@math.nsc.ru
Новосибирск, Институт математики СО РАН

Элементарными структурами в ДНК-последовательностях будем считать повторы следующих трех типов: *прямые* (. . . *gaactc* . . . *gaactc* . . .), *симметричные* (. . . *gaactc* . . . *ctcaag* . . .) и *комплементарные симметричные* (. . . *gaactc* . . . *gagttc* . . .). В последнем случае речь идет о симметричных повторах с точностью до переименования элементов алфавита в соответствии с известным в молекулярной биологии отношением комплементарности: $a \leftrightarrow t$, $c \leftrightarrow g$. *Комбинированными* назовем структуры, состоящие из двух разнотипных повторов (прямой плюс симметричный, прямой плюс комплементарный симметричный, и т. п.) с ограничениями снизу на длины повторяющихся цепочек и сверху — на те же длины и расстояния между соседними элементами структуры. Ограничения снизу нужны для отсеивания случайных («шумовых») структур, а сверху — для обеспечения компактности структуры. Порядок чередования цепочек, образующих повторы разных типов — произвольный, возможны наложения и совпадения цепочек, относящихся к разным повторам.

Целью работы является реализация и исследование алгоритма выявления комбинированных структур в ДНК-последовательностях. Алгоритм может быть использован для обнаружения возможных регуляторных областей и «горячих точек» генома, выявления потенциально многофункциональных фрагментов с наложением структур, поиска нестандартных (из-за учета симметрии и переименования элементов алфавита) образцов с двумя переменными [1].

Описание алгоритма

Для поиска комбинированных структур мы используем разработанный нами ранее аппарат построения сложностного профиля символьной последовательности, опирающийся на факторизацию Лемпеля и Зива [2], но с расширенным спектром операций копирования. Применительно к нашему случаю мы добавляем операции симметричного копирования, как с переименованием элементов алфавита, так и без [3]. Поясним схему алгоритма на примере выявления комбинированной структуры, составленной из прямого и симметричного повторов, не совпадающих тождественно друг с другом.

Шаг 1. Задаем нижнюю и верхнюю границу длин повторяющихся фрагментов (соответственно, r и R), а также максимально допустимое

расстояние между соседними компонентами структуры. Эти параметры определяют максимально возможный размер структуры $W = 4R + 3d$.

Шаг 2. Используя операцию симметричного копирования, вычисляем сложностной профиль текста с помощью скользящего окна размера W [3]. Выделяем окна, содержащие хотя бы один *нерасширяемый* компонент (фактор) с длиной $r \leqslant l \leqslant R$ и указателем копирования, равным 1 (первый символ окна). Тем самым фиксируется новая (не рассматривавшаяся ранее) симметрия, которая при наличии прямых повторов в ближайшей ее окрестности может образовать комбинированную структуру. Нетрудно показать, что прямые повторы следует искать в выделенном окне, расширенном влево на $(2R + 2d)$ символов.

Шаг 3. Построив L-граммное дерево (trie-структура) для расширенного окна [3] , фиксируем листья, соответствующие двум (или большему числу) одинаковых, но позиционно разнесенных цепочек длины L. Пары, которые могут быть расширены по тексту до максимально возможной длины l $(r \leqslant l \leqslant R)$ при сохранении свойства идентичности, являются искомыми прямыми повторами.

Шаг 4. Совмещаем (позиционно) симметрию с каждым из найденных прямых повторов. Если в образовавшейся 4-компонентной структуре расстояния между соседними компонентами не превышают d, фиксируем наличие комбинированной структуры.

Если параметр r логарифмическим образом зависит от величины W, т. е. соответствует средней длине максимального случайного повтора для окна анализа, трудоемкость алгоритма в среднем имеет порядок NL, где N — длина текста, L — длина цепочек, представленных в поисковом (L-граммном) дереве (в наших экспериментах L выбиралось близким к r).

Апробация алгоритма

Алгоритм проверялся на хорошо изученном геноме фага λ (длина $N = 48502$) и подборке кодирующих участков различных генов человека (всего 1469 последовательностей). При значениях параметров $r = 7$, $R = 20$, $d = 13$ в λ выделено 16 комбинированных структур типа «прямой повтор + симметричный комплементарный», в подборке генов — 1785 структур. Приведем примеры наиболее интересных структур.

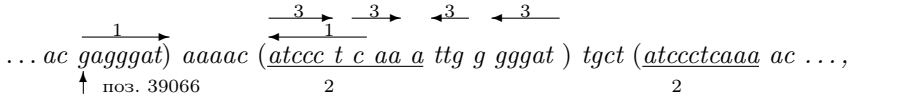
Пример 1 (фаг λ). В некодирующей области выделена структура

$$\ldots at\ \underline{gacAaaaa}\ \overrightarrow{attagcgcaag}\ aa\ \underline{gacaaaaa}\ tcac\ \overleftarrow{cttgcgctaat}\ gc \ldots,$$

↑ точка окончания транскрипции (поз. 27538)

содержащая прямой повтор (подчеркнут) и симметричный комплементарный повтор (указан стрелками сверху). Эта структура содержит точку окончания транскрипции по комплементарной цепи (A) и терминатор транскрипции (справа от A).

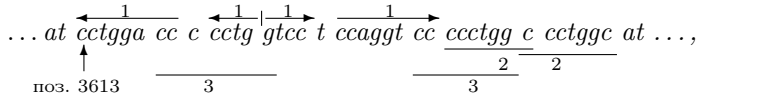
Пример 2 (фаг λ). В кодирующей области выделена структура



построенная на несовершенных тандемных повторах длины 19 (выделены скобками) с точно совпадающими ядрами длины 10 (подчеркнуты) и симметричными комплементарными повторами на стыке (см. 1) и внутри периода (см. 3). Выявленная структура лежит в области начала репликации фага λ. Представляет интерес наложение разных элементарных структур $(1, 2, 3)$ в центральной части фрагмента.

Структур типа «прямой повтор + симметричный» для обоих типов данных выявлено больше.

Пример 3 (ген «Collagen II α 1»). Выделена структура



содержащая несовершенную симметрию (1) с одной заменой и два прямых повтора (см. 2, 3). Структура возникает на периодичностях вида $(ccXggXccX)^n$, где $n = 4$, а X — произвольный нуклеотид. На аминокислотном уровне ей также соответствует симметричная структура $(PGP)^4$.

Заключение

Предложен эффективный алгоритм выявления в ДНК-последовательностях компактных комбинированных структур, состоящих из неслучайных повторов разного типа. Эксперименты на текстах с известной разметкой демонстрируют наличие таких структур в регуляторных областях и «горячих точках» генома. Близкими в идейном плане являются работы по отысканию фрагментов ДНК с аномально низкой сложностью [3] и образцов с двумя переменными [1]. Наше продвижение относительно [3] состоит в выявлении альтернативных накладывающихся друг на друга структур в зоне аномальной сложности. Продвижение по отношению к [1] состоит в расширении трактовки повтора и варьировании понятия образца.

Работа выполнена при поддержке РФФИ, проект № 06-06-80467.

Литература

[1] *Neraud J.* Algorithms for detecting morphic images of a word // Information and Computation. — 1995. — V. 120. — Pp. 126–148.

[2] *Lempel A., Ziv J.* On the complexity of finite sequences // IEEE Trans. Inform. Theory. — 1976. — V. IT–22, № 1. — Pp. 75–81.

[3] *Гусев В. Д., Немытикова Л. А.* Учет проявлений повторности, симметрии и изоморфизма в символьных последовательностях // Вычислительные системы. — Вып. 167. — Новосибирск, 2001. С. 11–33.