АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И ТАБЛИЦ ДАННЫХ
(Вычислительные системы)
1994 год Выпуск 150

УДК 519.767

ИСПОЛЬЗОВАНИЕ СЕРИЙНЫХ ХАРАКТЕРИСТИК ДЛЯ ИССЛЕДОВАНИЯ ЭФФЕКТА КЛАСТЕРИЗАЦИИ ЭЛЕМЕНТОВ В ДНК-МОЛЕКУЛАХ

Л.А. Немытикова

В в е д е н и е

Традиционные марковские модели представления первичных структур ДНК-молекул плохо учитывают эффект кластеризации элементов определенного вида вдоль длины последовательности [1]. О наличии этого эффекта свидетельствуют работы Блайсдела [2], Сприжицкого [3] и автора [4].

В указанных работах под кластеризацией понималось либо а) наличие аномально длинных (по отношению к марковской модели нулевого порядка) серий из элементов R,Y (R = {A,G} - пурины, Y={C,T} - пиримидины) или S,W (S={C,G} - сильные (по числу водородных связей) нуклеотиды, W={A,T} - слабые), либо б) аномально высокое число серий определенной длины из элементов того же (что в п."а") типа. (Под *серией* понимается цепочка из однотипных элементов, ограниченная по краям элементами другого типа.)

Блайсдел [2], в частности, обнаружил, что в эукариотических последовательностях семейства позвоночных *кодирующие* части характеризуются избытком коротких (l = 1,2) и дефицитом длинных (l = 3-7) серий слабых и сильных оснований; *некодирующие* части - дефицитом коротких и избытком длинных серий пуринов и пиримидинов. В [4] исследовались РНК-последователь -

ности (кодирующие сегменты генома вируса гриппа). Было показано, что на этих последовательностях выполняются обе отмеченные выше закономерности.

Сприжицкий [3] провел объемное исследование на первичных структурах ДНК-последовательностей 5 семейств (прокариоты, грибы, беспозвоночные, позвоночные, млекопитающие). Он выявил резкое преобладание R- и Y-серий у позвоночных и млекопитающих как в кодирующих, так и в некодирующих участках, начиная с $l = 4$ и выше. У прокариотов, грибов и беспозвоночных аномально высокое число R- и Y-серий наблюдается лишь в некодирующих участках, притом при длинах серий $l \geq 10$. У всех пяти семейств отмечается резкое преобладание длинных ($l \geq 10$) W-серий в некодирующих частях, что же касается S-серий, то это справедливо только для позвоночных и млекопитающих и в меньшей степени для прокариотов. В кодирующих частях всех 5 семейств не отмечается особых преобладаний по длинным S- и W-сериям.

Приведенный обзор результатов свидетельствует о многообразии форм проявления эффекта кластеризации символов определенного типа в ДНК-молекулах. Это побудило авторов [5] разработать специальную "блочную" модель ДНК-последовательности, где под блоком понимается серия в агрегированном алфавите.

Кластеризуемость (в той или иной форме) определенных элементов алфавита вдоль длины НК-последовательности носит, повидимому, столь же фундаментальный характер, как, например, правила динуклеотидного предпочтения, сформулированные Рут Нуссинов [6]. Исследование всех проявлений этого эффекта представляет большой интерес с биологической, классификационной и алгоритмической точек зрения.

Целью данной работы является: 1) введение нового критерия кластеризации и апробация его на подборках кодирующих и некодирующих последовательностей (экзоны, интроны, эукариотические промоторы); 2) сопоставление введенного критерия с уже известными.

1. Агрегирование алфавита

В текстах, составленных из элементов относительно большого алфавита (порядка 10-100 символов и выше), сколь-либо протяженные серии встречаются редко, да и сами по себе они малоинформативны. Зато при малых алфавитах, особенно двоичном, серии играют заметную роль и серийные характеристики используются в различных статистических критериях.

Для того чтобы эффективно использовать серийные характеристики для описания текстов с большим алфавитом, требуется провести предварительное *агрегирование* алфавита, т.е. разбиение его на небольшое число непересекающихся подмножеств (например, на два подмножества). Текст в агрегированном алфавите удобно описывать на языке серий.

Чтобы проиллюстрировать идею агрегирования, приведем следующий пример,касающийся естественных языков [7]. В русском и других языках слов, обладающих зеркальной симметрией (типа: "тут, как, топот, Анна" и т.п.), очень мало. Несколько больше ритмически построенных слов типа: "мама, няня, вот-вот" и т.д. Если заменить конкретные буквы классами букв, например, слова "тут" и "как" записать в виде "глухая согласная-гласная-глухая согласная", то симметричных слов, построенных по этой схеме, окажется намного больше. Если же все слова записать в терминах всего двух классов букв (согласная и гласная), то *симметричных* и *ритмичных* слов окажется более 70%.

Для *малых* алфавитов целесообразно проводить все возможные варианты агрегирования с вычислением соответствующих серийных характеристик [4]. В случае *больших* алфавитов вариант агрегирования часто подсказывается спецификой предметной области (например, все множество аминокислот ($|A| = 20$) можно разбить на два класса - гидрофобные и гидрофильные - или на три класса - положительно заряженные, отрицательно заряженные и нейтральные).

При работе с текстом на неизвестном языке, когда вариант агрегирования заранее неочевиден, возможен и дешифровочный подход: можно попытаться найти такое разбиение алфавита на подмножества, которое оптимизировало бы тот или иной критерий, в частности, серийный. Так, для автоматического разбиения алфавита неизвестного языка на гласные и согласные можно использовать критерий максимально возможной чередуемости элементов этих двух множеств [8] (текст тем благозвучнее, чем регулярнее перемежаются в нем гласные и согласные). На языке серий это можно было бы сформулировать следующим образом: найти такое разбиение алфавита, при котором общее число серий было бы максимальным (обычно этому соответствует аномально высокое число коротких (длины 1-2) серий).

В данной работе предпринята попытка ввести критерий кластеризуемости, учитывающий несколько агрегирований одновременно.

2. Критерий кластеризуемости

Алфавит ДНК-молекул состоит из 4 элементов $\Sigma = \{A,C,G,T\}$, которые допускают различные варианты осмысленного объединения (агрегирования). По типу азотистого основания, входящего в состав нуклеотидов, их можно разделить на пурины $R = \{A,G\}$ и пиримидины $Y = \{C,G\}$; по числу водородных связей, образуемых при комплементарном связывании, - на слабые $W = \{A,T\}$ и сильные $S = \{C,G\}$; по заряду - на амино (положительные) $M = \{A,C\}$ и кето (отрицательные) $K = \{G,T\}$ [9]. Все остальные разбиения алфавита - неравномощные (их 4).

В данной работе, исходя из соображений симметрии и удобства интерпретации, ограничимся лишь рассмотрением трех упомянутых выше разбиений. Имея в виду, что при каждом агрегировании алфавит делится на два подмножества, будем кодировать элементы этих подмножеств соответственно нулями и единицами.

Введем параметр s, принимающий всего два значения: s = 0 будет использоваться для характеризации серий из 0, а s = 1 - для характеризации серий из 1. Рассмотрим наиболее интересные, с нашей точки зрения, серийные характеристики: l_{smax} - длина максимальной серии типа s; r_{sj} - число s-серий длины j (j = 1, 2,...,l_{smax}); $r = \sum_{s} \sum_{j=1}^{l_{smax}} r_{sj}$ - общее число серий; $S_s = \sum_j S_{sj}$ - число разновидностей серий типа s [4] (в диапазоне от 1 до l_{smax} могут быть представлены не все серии, т.е. возможны ситуации, когда $r_{sj} = 0$ при некоторых значениях j; S_{sj} - число случаев, при которых $r_{sj} \neq 0$); $r_s(k) = \sum_{j=1}^{k} r_{sj}$ - число s-серий с длиной, не превышающей k [4]; $p_s(k) = \sum_{j=k}^{l_{smax}} r_{sj}$ - число s-серий, длина которых не меньше k; d(k) - длина максимального фрагмента, в котором расстояние между соседними элементами типа s не превышают k [4] (кластеры из s-элементов).

Характеристика r чаще всего используется в тестах на случайность или однородность [10]. Блайсдел и Сприжицкий в своих работах [2,3] использовали, в основном, характеристики r_{sj} при j = 1,2,...,k и $p_s(k)$ - для оценивания хвоста распределения. В работе [4] использовались все характеристики кроме $p_s(k)$.

В данной работе большое внимание уделено именно характеристикам $p_s(k)$, взятым за основу для введения еще двух (новых) серийных характеристик, связывающих сразу все три типа агрегирования:

$b(k) = \sum_{i=1}^{3} \sum_{s} p_s(k)$ - суммарное по всем агрегированиям число s-серий, длина которых не меньше k;

$q(k)/N$ - доля элементов текста, входящих хотя бы в одну серию (любого из 6 типов), длина которой не меньше k (коэффициент покрытия текста длины N "длинными" сериями).

ПРИМЕР 1. Пусть T = GCTTAAATACGAGGCCGGGCCTCCTCTA, N = 28, k = 7. Тогда $b(7) = 3$, $q(7) = 22$, $\frac{q(7)}{N} = \frac{11}{14}$. Действительно, в тесте T существуют три серии, длина которых не меньше 7: (T,A)-серия, (G,C)-серия и (C,T)-серия (выделены подчеркиванием), причем (G,C)- и (C,T)-серии пересекаются по элементу C. Заметим, что в общем случае перекрытию двух серий в агрегированном алфавите всегда соответствует моносерия. Длина покрытия текста T этими сериями - 22, коэффициент покрытия - 0.78.

Критерием кластеризуемости элементов алфавита вдоль последовательности будет являться аномально высокое наблюдаемое значение статистики $b(k)$ $(q(k)/N)$ по сравнению с ожидаемым для случайной последовательности с тем же частотным составом (в качестве грубого приближения ДНК-последовательности обычно используется модель независимых испытаний). Оценки, ожидаемые для случайной последовательности получаются имитационным моделированием путем случайного "перемешивания" исходного текста. При этом сохраняется частотный состав элементов. Эксперимент повторяется m раз (мы использовали значение m = 100). В каждом эксперименте (для каждой новой перемешанной последовательности) вычисляются все перечисленные выше характеристики. По результатам m экспериментов для каждого параметра α вычисляются его минимальное и максимальное значения $\alpha^{сл}_{min} = \min_i(\alpha_i)$, $\alpha^{сл}_{max} = \max_i(\alpha_i)$, среднее $\bar{\alpha} = \frac{1}{m}\sum_{i=1}^{m}\alpha_i$ и среднеквадратичное отклонение $\sigma = \left[\frac{1}{m}\sum_{i=1}^{m}(\alpha_i - \bar{\alpha})^2\right]^{1/2}$.

Вывод об аномальном поведении параметра α, наблюдаемого на реальной последовательности, делается с помощью одного из двух критериев: "жесткого":

$$\alpha \geq \min(\alpha^{сл}_{max}, (\bar{\alpha} + 3\sigma)) \text{ или } \alpha \leq \max(\alpha^{сл}_{min}, (\bar{\alpha} - 3\sigma)) \qquad (1)$$

и более "мягкого":

$$\alpha > (\bar{\alpha} + 2\sigma) \text{ или } \alpha < (\bar{\alpha} - 2\sigma). \qquad (2)$$

Трудоемкость вычисления всех серийных характеристик линейным образом зависит от длины последовательности N. Характеристики b(k) и q(k) вычисляются за один просмотр исходного текста, все остальные характеристики - за один просмотр агрегированного (двоичного) текста. Трудоемкость всего эксперимента с имитационным моделированием - $O(N \cdot m)$.

3. Апробация критериев на реальных данных

В качестве объектов для исследования были выбраны*):

а) подборка промоторных областей эукариотических генов (преимущественно человека). Подборка включает 110 последовательностей длины 100, 101 последовательность длины 400 (большая часть из них является расширением соответствующих 100-элементных текстов), 100 последовательностей длины 600 (эта подборка не пересекается с предыдущими);

б) подборка интронов генов человека (86 последовательностей с длинами от 21 до 3738, суммарная длины порядка 25000 символов);

в) подборка экзонов генов человека (100 последовательностей с длинами от 33 до 2190, суммарная длина порядка 25000 символов).

---

*) Автор благодарит сотрудников лаборатории Н.А.Колчанова (Институт цитологии и генетики СО РАН) за возможность воспользоваться соответствующими материалами.

Принципиальным отличием от экспериментов Сприжицкого [3] являлось то, что решение об аномальности по тому или иному критерию принималось для каждой последовательности отдельно, а не для всей совокупности последовательностей, объединенных в один текст. Как будет показано ниже, зависимость от длины является существенной.

По результатам экспериментов можно сделать следующие выводы:

1. Некодирующие последовательности (интроны и промоторы) в значительной степени характеризуются аномально высокими показателями характеристик $b(k)$ и $q(k)$ в диапазоне значений $k = 4-9$ (см. таблицу).

Таблица

Разбиение текстов по параметру "кластеризуемость"

Текст	$+2\sigma$	$+3\sigma$	Тексты типа случайных	Противоречивые тексты	-2σ
Промоторы (N = 100)	57.3	30	40.9	1.8	0
Промоторы (N = 400)	83	67	9	7	1
Интроны	60.5	23	29.1	4.6	5.8
Экзоны	34	6	52	3	11

В первом столбце этой таблицы указано процентное содержание текстов, удовлетворяющих критерию (2) ($\alpha > \bar{\alpha} + 2\sigma$, где α - наблюдаемое значение статистики $b(k)$ или $q(k)$, $\bar{\alpha}$ - ожидаемое); во втором столбце - процентное содержание текстов, аномальных по "жесткому" критерию (1) ($\alpha \geq \min(\alpha^{сл}_{max}, (\bar{\alpha} + 3\sigma))$). В третьем столбце - процент текстов, неотличимых от случайных ($\bar{\alpha} - 2\sigma < \alpha < \bar{\alpha} + 2\sigma$). В четвертом столбце - процент "противо-

речивых" текстов: при одних значениях k величина b(k) (или q(k)) аномально высокая при других k - аномально низкая. В пятом столбце указана доля текстов, в которых критерий аномальности по b(k) или q(k) срабатывает по нижнему ограничению ($\alpha < \bar{\alpha} - 2\sigma$), что соответствует хорошей чередуемости элементов при всех агрегированиях т.е. отсутствию кластеризуемости. Из этой же таблицы видно, что в кодирующих частях (экзонах) эффект кластеризации (в смысле критериев (1) и (2)) проявлен в незначительной степени.

2. Очень интересной представляется зависимость эффекта кластеризуемости от длины текстов (этот вопрос, как уже упоминалось выше,в [2-3] не исследовался). С *увеличением длины* последовательности эффект (там, где он имел место) проявляется в более *яркой* форме (показательными в этом плане являются две первые строки таблицы: 100- и 400-элементные промоторные области). Формально это объясняется тем, что оценка дисперсии в критериях (1) и (2) с увеличением N растет слабее, чем систематически накапливаемое смещение (в сторону преобладания длинных серий) в наблюдаемой реализации. К примеру, 100-элементный промотор (HSACTBPR) при k = 8 характеризуется значением b(8) = 4 и по результатам имитационного моделирования ($b^{сл}_{min} = 0$, $b^{сл}_{max} = 5$, $\bar{b}(8) = 2.17$, $\sigma(b) = 1.55$) не может быть отнесен к аномальным ($b(8) = 4 < \bar{b}(8) + 2\sigma(8) = 4.48$).Его 400-элементное расширение характеризуется значением b(8) = 16 и по результатам имитационного моделирования ($b^{сл}_{min} = 4$, $b^{сл}_{max} = 14$, $\bar{b}(8) = 8.98$, $\sigma(b) = 2.36$) уже относится к аномальным ($b(8) = 16 \simeq \bar{b}(8) + 3\sigma(8) > b^{сл}_{max} = 14$).

Тот же эффект наблюдается и у интронов (их длина сильно варьирует). Среди 5 интронов, попавших в четвертый столбец таблицы, четыре имеют длину меньше 85 символов. Аналогично,из 25 интронов, попавших в столбец 3, двадцать один имеют длину мень-

ше 126 символов. Зато все длинные интроны (с $N > 1000$ символов) обладают аномально высокими значениями $b(k)$ и $q(k)$ в диапазоне значений $k = 5\text{-}9$, значительно выходящими за пределы 3σ-интервала (например, в самом длинном интроне ($N = 3738$) $b(9) = 41$, тогда как $\bar{b}(9) = 22.2$, $\sigma = 3.7$, $b_{max}^{сл} = 30$).

В экзонах эффект "накопления" высоких показателей $b(k)$ и $q(k)$ с увеличением N не наблюдается, поскольку отсутствует систематическое смещение этих показателей в сторону превышения по отношению к ожидаемому значению: высокие значения $b(k)$ уравновешиваются низкими.

4. Сравнительный анализ серийных характеристик

1. Характеристики $b(k)$ и $q(k)$ коррелированы. Чаще всего (особенно при больших N) аномалии проявляются одновременно по $b(k)$ и по $q(k)$. Ситуации, когда это не так, объясняются следующим образом. Если наблюдается аномально высокое значение $b(k)$, а $q(k)$ не аномально, это соответствует высокой степени "перекрываемости" агрегированных серий, т.е. наличию длинных моносерий.

ПРИМЕР 2. Пусть T = GCTAGCCCCCCTTAGCTA.

При $k = 7$ $b(7) = 2$; $q(7) = 9$. Здесь высокая степень перекрываемости существенно ограничивает значение параметра q.

И, наоборот, если наблюдается аномально высокое значение $q(k)$, а $b(k)$ не является аномальным, это чаще всего соответствует наличию аномально длинных серий (в коротких текстах), либо слабой перекрываемости агрегированных серий (в длинных текстах).

2. В ходе экспериментов с указанными типами данных проверялись (кроме $b(k)$ и $q(k)$) и другие серийные характеристики, в частности l_{smax} ($s = 0,1$), r_{sj} ($1 \le j \le l_{smax}$), r и $p_s(k)$.

Аномально высокое значение параметра l_{smax} свидетельствует о наличии уникального по длине единичного кластера, как правило, функционально значимого. Примером могут служить некоторые TATA-боксы в эукариотических промоторах, имеющие длину 9 и более символов.

Параметр r_{sj} чаще всего бывает аномален при малых значениях j в агрегированиях типа S, W, что связано не с кластеризуемостью, а, наоборот, с очень хорошим перемешиванием S- и W-элементов (в кодирующих последовательностях). Аномально высокие значения r_{sj} при больших j встречаются не так часто и затруднительны для трактовки (избыток кластеров *фиксированной* длины).

Основной вклад в параметр r вносят величины r_{s1} и r_{s2}. Поэтому параметр r коррелирован с ними. О кластеризуемости сигнализируют очень низкие значения r. Параметр r достаточно информативен, но носит слишком интегральный характер: он не дает представления о том, в каком диапазоне длин серий наблюдается аномальность.

Параметры p_{sk} при не слишком малых значениях k ($k \geq 4$) в наилучшей степени приспособлены для выявления кластеризуемости элементов текста по отдельным агрегированиям. Они чаще выявляют аномалии, чем параметры r_{sj}, и в то же время на них не действует такой дестабилизирующий фактор как число коротких серий, способный замаскировать эффект кластеризации (такое случается при использовании параметра r).

4. Характеристики b(k) и q(k) - единственные (из рассмотренных выше), ориентированные на учет нескольких агрегирований сразу. Можно наблюдать в некоторых текстах (например, в промоторных областях) предрасположенность к кластеризации при разных агрегированиях. Объединение разрозненных эффектов может привести к их усилению, что формально проявляется в аномальности b(k) или q(k).

Другим существенным моментом, связанным с характеристиками $b(k)$ и $q(k)$, является возможность обнаружения в тексте с их помощью устойчивых комбинаций серий от разных агрегирований. Такие комбинации могут оказаться функционально значимыми (соответствующие примеры, связанные с "контрастным" оформлением TATA-бокса в некоторых эукариотических промоторах приведены в [11]). Методика выявления устойчивых комбинаций серий от разных агрегирований описана в следующем разделе.

Поскольку характеристики $b(k)$ и $q(k)$ определяются через $p_s(k)$, они коррелированы с ними, особенно в ситуациях, когда одно из агрегирований явно превалирует в смысле потенциальной кластеризуемости элементов (таковым часто является разбиение на пурины и пиримидины). Однако даже при достаточно сильной корреляции характеристики $b(k)$ и $q(k)$, с одной стороны, и $p_s(k)$, с другой, не заменяют друг друга. Приведем соответствующие примеры. В промоторе HSABL1B ($N = 100$) $b(9) = 5 = b^{сл}_{max}(9)$, т.е. он аномален по параметру b ($k=9$). В то же время ни при одном агрегировании характеристики $p_s(9)$ не аномальны (при других значениях k аномалии есть: $p_{(c,т)}(8) = 1 > \overline{p}_{(c,т)}(8) + 2\sigma$). Такая картина наблюдается в 5 (из 110) промоторах ($N = 100$) и в 4 (из 101) 400-элементных промоторах.

При пурин-пиримидиновом агрегировании наблюдается устойчивое преобладание длинных серий, дающих наибольший вклад в характеристику $b(k)$ ($q(k)$). Тем не менее в 13 из 110 промоторов ($N = 100$) параметр $p_s(k)$ при агрегировании R-Y не аномален, а $b(k)$ имеет аномально высокое значение. И ровно столько же случаев, когда тексты не имеют аномалий при агрегировании R-Y и не имеют аномалии по характеристикам $b(k)$ ($q(k)$).

Существуют и обратные примеры (их не меньше). Так в промоторе HSAFP1 ($N = 400$) выявлены следующие аномально высокие значения $p_s(k)$: $p_{(A,C)}(4)$; $p_{(A,G)}(k)$, при $k = 6,7,8$; $p_{(C,T)}(9)$.

Однако характеристики b(k), k = 4-9, осталась на уровне случайной.

В заключение данного раздела отметим, что рассмотренный нами набор характеристик достаточно полно отражает различные аспекты кластеризуемости элементов определенного типа вдоль последовательности. Интересно отметить в связи с этим, что из всех рассмотренных нами текстов очень мало оказалось таких, в которых не был бы обнаружен этот эффект хотя бы с помощью одной из характеристик.

5. Анализ взаимного расположения серий

Представляет интерес выяснить, существуют ли в анализируемых текстах какие-либо устойчиво повторяющиеся комбинации серий из разных агрегирований. Для ответа на этот вопрос предлагается следующая методика.

Осуществляем перекодирование каждого текста подборки, отражающее его "серийную структуру".Для этого задаем параметр k и выделяем в тексте все серии с длиной большей или равной k по всем трем агрегированиям. Каждую серию вне зависимости от ее длины обозначаем одной буквой в соответствии с составом ее элементов в агрегированном алфавите (всего выделяем 6 типов серий: Z = {R,Y,S,W,M,K}). Символы текста, не вошедшие ни в одну из серий (интервалы между сериями), кодируются с помощью символа N. При этом учитывается длина интервала L: если $L \leq k$, интервал кодируется одним символом N. Если $k < L \leq 2k$, интервал кодируется цепочкой NN и т.д. Учет длины интервала важен для оценки степени удаленности серий друг от друга (например, комбинация типа RNY, где R и Y - инвертированные комплементарные повторы, имеет больше шансов образовать шпилечную структуру, чем комбинация RNNNY, из-за меньшего размера петли).

ПРИМЕР 3. Пусть k = 7.

T=TCCGGGGCCGGGGGAGGGGGTGTAGGAGGAACCTAAGGAAGACATACGTCACAATTAAT.

После перекодирования текст будет иметь вид T=NSRKRNRNNW.

Нетрудно видеть, что при описанной схеме кодировки в перекодированных текстах отсутствует комбинация типа XX, где X∈Z. Они могли бы появиться при введении дифференциации серий по длинам.

Объединяем перекодированные тексты подборки в один текст $T = T_1*T_2*...*T_t$, где t - число текстов, а "*" - разделитель между текстами. Вычисляем для текста T все биграммные статистики вида F(XY) ($X \neq Y$, $X,Y \in Z$, F(XY) - частота встречаемости XY в T) и триграммные вида F(XQY) ($X,Y \in Z$, $Q \in Z \cup \{N\}$). Биграммные статистики учитывают рядом расположенные (в большинстве случаев перекрывающиеся) серии разного типа, триграммные - незначительно разнесенные серии (возможно, одного типа). Более сильно разнесенные серии (XNNY, XNNNY и т.д.), повидимому, не представляет особого интереса.

О значимости полученных статистик можно судить по итогам имитационного эксперимента с m-кратным перемешиванием текста T (по отношению к исходному тексту речь идет о перемешивании блоков, а не символов). Перемешивание проводится с соблюдением двух ограничений: 1) в перемешанном тексте (как и в исходном) не должны встречаться комбинации типа XX, где $X \in Z$; 2) перемешивание не затрагивает разделителей.

Приведем примеры аномалий, обнаруженных с помощью вышеописанной методики.

В 100-элементных промоторах биграмма RS имеет аномально высокую частоту встречаемости при значениях k = 5,6,7. При k = = 5 RS встречается 40 раз, $F_6(RS) = 25$, $F_7(RS) = 12$. В то же время комбинация SR не обнаруживает аномалий по частоте:

$F_5(SR) = 24 < 25.5 = \overline{F}_5(SR)$; $F_6(SR) = 9 < 12.4 = \overline{F}_6(SR)$; $F_7(SR) = 3 < 6.7 = \overline{F}_7(SR)$. (Заметим, что $\overline{F}_k(RS)$ очень близко к $\overline{F}_k(SR)$.)

Аналогичный эффект преобладания частоты RS над SR наблюдается и у интронов: $F_5(RS) = 19$; $F_5(SR) = 6$; $\overline{F}_5(RS) \simeq \simeq \overline{F}_5(SR) \simeq 9.6$.

При k = 5-8 наблюдается устойчивая аномалия по частоте встречаемости биграммы SY (F(SY) >> F(YS)).

Тот же эффект несимметрии частот наблюдается для биграммы YK ($F_6(YK) = 9$ - аномально высокая частота, $F_6(KY) = 0$ - аномально низкая частота).

Среди "разнесенных" серий можно отметить некоторые комбинации в интронах с аномально низкой частотой, например, YQS и RQW при k = 5,6 (Q ∈ {R,Y,S,W,K,M,N}). В 400-элементных промоторах низкая частота встречаемости наблюдается у комбинаций SQW и WQS. Обе комбинации при k = 5 встречаются всего по 2 раза.

З а к л ю ч е н и е

Распределение элементов по длине нуклеотидной последовательности не является случайным: очень часто нуклеотиды определенного типа образуют кластеры. Кластеризуемость элементов НК-последовательностей носит столь же универсальный характер как и правила динуклеотидного предпочтения, сформулированные Рут Нусинов.

Для выявления эффекта кластеризации используется идея агрегирования (укрупнения) алфавита. Если в результате агрегирования алфавит разбивается на два подмножества, то удобно описывать агрегированную (двоичную) последовательность с помощью серийных характеристик. В частности, в терминах этих характеристик можно формулировать различные критерии кластеризуемости элементов последовательности.

Известные критерии кластеризуемости формулировались применительно к фиксированному варианту агрегирования алфавита. В работе введен новый критерий кластеризуемости элементов НК-последовательностей, учитывающий возможность образования кластеров одновременно по всем равномощным двоичным агрегированиям нуклеотидного алфавита (таких агрегирований 3).

Смысл одновременного рассмотрения различных агрегирований заключается в том, что многие значимые конструкции, например, такие как шпилечные структуры, симметрии, ТАТА-боксы в промоторах и т.п., часто имеют ярко выраженную блочную структуру, где блоки являются сериями в агрегированном алфавите. Соседство серий разного типа, выявляемое с помощью введенного критерия, создает предпосылки для образования указанных выше структур.

Сопоставление введенного критерия с уже известными, имеющими дело с фиксированными агрегированиями, показало их частичную коррелированность (что вполне естественно), но не взаимозаменяемость. Существуют ситуации, когда последовательность не обнаруживает кластеризуемости ни по одному из агрегирований в отдельности, но по всем вместе является аномальной, т.е. содержит избыток серий по совокупности агрегирований. Справедливо и обратное: последовательность может обнаруживать кластеризуемость по одному из агрегирований, но не по всем сразу.

Проведена апробация нового критерия на реальном материале (эукариотические промоторы, интроны, экзоны). Показана высокая кластеризуемость (в соответствии с новым критерием) объектов типа "промоторные зоны" и "интроны". Отмечено, что с увеличением длины этих объектов срабатывает своего рода эффект "накопления": кластеризуемость проявляется ярче (иногда существенно ярче, чем в известных критериях).

Литература

1. GELEAND M.S. Computer functional analysis of nucleo - tide sequences: problems and approaches // Mathematical Methods of Analysis of Biopolymer Sequences (DIMACS series in descrete math.). - 1992.- Vol.8.- P.87-98 .

2. BLAISDELL B.E. A prevalent persistent global nonrandomness that distinguishes coding and non-coding eucaryotic nu - clear DNA sequences //J.Mol.Evol.- 1983.- Vol.19, N 2.- P.122-133.

3. СПРИЖИЦКИЙ Ю.А. Статистический анализ и распознавание функциональных участков генома: Дисс... канд.физ.-мат. наук: 03.00.02.- М., 1987.- 145 с.

4. ГУСЕВ В.Д., НЕМЫТИКОВА Л.А. Анализ серий в генетичес - ких текстах // Анализ временных рядов и символьных последова - тельностей.- Новосибирск, 1991.- Вып.141: Вычислительные сис - темы.- С.46-76.

5. SUBOCH G.M., SPRIZHITSKY Yu.A. Statistical signifi - cance of some complex nucleotide combinations: a comparison of DNA models // CABIOS.- 1990.- Vol.6, N 1.- P.43-48.

6. NUSSINOV R. Strong doblet preferences in nucleotide sequences and DNA geometry // J.Mol. Evol.- 1984.- Vol. 20. - P.111-119.

7. Лингвистические проблемы автоматизации редакционно-издательских процессов /Под редакцией Перебейнос и Феллер. - Киев: Наукова думка, 1986. - С.7-8.

8. СУХОТИН Б.В. Оптимизационные методы исследования язы - ка.- М.: Наука, 1976.

9. CORNISH-BOWDEN A. Nomenclature for incompletely spesi-fied bases in nucleic acid sequences: recomendation // NAR. - 1985.- Vol.13, N 9.- P.3021-3030.

10. ФЕЛЛЕР В. Введение в теорию вероятностей и ее приложения.- М.: Мир, 1964.- 498 с.

11. NUSSINOV R. The eucaryotic CCAAT and TATA boxes, DNA spacer flexibility and looping //J.Theor. Biol.-1992.-Vol.155.- P.243-270.

Поступила в ред.-изд.отд.

21 ноября 1994 года