

УДК: 577.212.2

Поиск регулярных последовательностей в промоторах из геномов различных групп организмов с использованием критерия серий

©2008 А.А. Шеленков, Е.В. Коротков

Центр «Биоинженерия» Российской академии наук, Москва, 117312, Россия

Аннотация. В статье вводится понятие «регулярность» для описания структурных свойств последовательностей ДНК, расширяющее понятие скрытой периодичности. Предложен метод обнаружения регулярности с использованием критерия серий. Проведенный поиск регулярных последовательностей в эукариотических промоторах показал, что более 60% из них обладают регулярностью на статистически значимом уровне. Обсуждаются возможные биологические функции регулярности и возможность использования данной характеристики для аннотации промоторов.

Ключевые слова: регулярность, промоторы, последовательности ДНК.

ВВЕДЕНИЕ

В настоящее время проводится широкомасштабный анализ последовательностей различных геномов, в частности, генома человека. Одной из важнейших задач этого анализа является характеристика и определение функций различных генов. В последнее десятилетие был предложен ряд достаточно надежных методов предсказания участков, кодирующих белок (например, [1]). Однако предсказание регуляторных участков, в частности, промоторов, все еще остается сложной задачей, хотя также был предложен ряд методов для их обнаружения [2–4]. Промотор – это участок генома, расположенный вблизи сайта инициации транскрипции и играющий ключевую роль в генетической регуляции [5]. Промоторы получают сигналы от различных источников (например, от клеточных рецепторов) и контролируют уровень инициации транскрипции, которая в значительной степени определяет экспрессию гена [6]. Таким образом, обнаружение промоторов является важным шагом для проведения аннотации генов.

Для того, чтобы разделить участки геномов, содержащие и не содержащие промоторы (последних, очевидно, большинство) был использован целый ряд признаков, например CpG островки [3, 7], ТАТА-боксы [4, 8], СААТ-боксы [4, 8], некоторые характерные сайты связывания факторов транскрипции [4, 8, 9], матрицы пентамеров [2], олигонуклеотиды [10], а также комбинированные подходы [11].

Кроме того, были использованы различные процедуры распознавания образов, такие как нейронные сети [4, 7, 8], линейный и квадратичный дискриминантный анализ [3, 9], интерполяционная Марковская модель [4], анализ независимых составляющих [12].

Однако анализ экспериментальных данных показал [13], что вопрос выбора правильных биологических сигналов, используемых в программах предсказания промоторов, все еще остается открытым. Ни один из этих сигналов не описывает все

разнообразии промоторов, и каждый признак, полученный на основе изучения промоторных последовательностей, имеет свои ограничения в использовании [11].

Таким образом, существует необходимость выделения некоторой новой характеристики последовательностей промоторов, которая являлась бы специфичной по отношению к этим элементам, но при этом обладала бы достаточной гибкостью для того, чтобы соответствовать многообразию видов таких последовательностей.

Ранее нами была предложена скрытая периодичность в качестве характеристики, позволяющей проводить аннотацию последовательностей с неизвестной функцией [14–16]. В частности, нами было показано, что последовательности, обладающие скрытой периодичностью с длиной периода 2–100, являются потенциальными минисателлитами. Скрытая периодичность является достаточно общим явлением, свойственным различным группам организмов. Тем не менее, данная характеристика не позволяет выделить промоторные участки из всего множества регуляторных элементов генома [15].

В данной работе мы предлагаем новую характеристику символьных последовательностей – регулярность, приводим метод ее обнаружения, основанный на использовании критерия серий, а также применяем этот метод для поиска регулярных последовательностей в промоторах из геномов различных групп организмов. Под регулярностью мы понимаем статистически значимое подобие распределений символов по участкам последовательности между исследуемой последовательностью и искусственной периодической последовательностью с некоторой длиной периода. Строгое определение регулярной последовательности дано в разделе 1.5.

Метод, приведенный в данной работе, можно использовать в том числе и для поиска скрытой периодичности, но он позволяет обнаруживать и последовательности, которые формально нельзя причислить к периодическим, но при этом имеющие сходную структурную организацию. Основными достоинствами метода являются его относительная нечувствительность к наличию вставок и делеций символов, отсутствие необходимости предварительного задания типа последовательности для поиска (например, в виде матрицы частот), и самое главное, возможность обнаруживать регулярность, сильно размытую в результате эволюционного процесса.

В разделе «Результаты и обсуждение» будет показано, что большинство известных промоторов обладают регулярной структурой.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1. Общее описание критерия серий

Критерий серий (Вальда–Вольфовица) [17, 18] является непараметрическим критерием, который применяется для проверки гипотезы H_0 , утверждающей, что две (или более, в общем случае) группы данных представляют случайные независимые выборки с объемами n_1 и n_2 ($n_1 + n_2 = N$) из одной генеральной совокупности, то есть, функции распределения для этих двух выборок не отличаются друг от друга. При использовании критерия серий для одной выборки оценивается число *серий* в ряде, в котором каждый элемент может принимать в общем случае k различных значений. Рассмотрим частный случай $k = 2$. Результаты наблюдений записываются в виде вариационного ряда объединенной выборки T , а принадлежность данных к той или иной группе обозначается с помощью кодирующей переменной, принимающей два значения (0 и 1, где 0 означает принадлежность к первой выборке, а 1 – ко второй). Значения кодирующей переменной образуют ряд R , называемый *последовательностью категорий*, или *последовательностью кодов*. Для применения критерия серий мы сортируем элементы ряда T в возрастающем порядке, при этом проводя аналогичные перестановки в ряду R .

Рассмотрим пример. Пусть получены две выборки: $\{7, 12, 18, 24, 32\}$ и $\{14, 25, 28, 30, 36\}$. Объединяя их в один вариационный ряд, получаем $\{7, 12, 18, 24, 32, 14, 25, 28, 30, 36\}$, при этом ряд категорий (значений кодирующей переменной) будет иметь вид $\{0, 0, 0, 0, 0, 1, 1, 1, 1, 1\}$. После проведения сортировки по возрастанию получаем: $T = \{7, 12, 14, 18, 24, 25, 28, 30, 32, 36\}$, $R = \{0, 0, 1, 0, 0, 1, 1, 1, 0, 1\}$.

Серия – это последовательность элементов ряда, содержащая элементы только одного типа, ограниченная элементами другого типа, либо началом или концом ряда.

Для приведенного выше примера получаем: 00 1 00 111 0 1 (серии подчеркнуты). Таким образом, данный ряд содержит 6 серий.

Статистикой критерия является число серий в последовательности кодов. Если гипотеза H_0 верна, то обе выборки должны быть хорошо перемешаны в общем вариационном ряду и число серий должно быть велико. Если же выборки получены из генеральных совокупностей с разными распределениями, различающимися средними значениями или разбросом, то число серий, по-видимому, будет мало.

Расчет статистик для критерия серий основан на предположении, что для случайного ряда число серий будет принимать значения из некоторого диапазона, то есть, неслучайность выборки означает, что число серий в ней меньше некоторого минимального числа или больше некоторого максимального. При определении этих экстремальных значений учитывается число серий в выборке, а также частота появления каждого из возможных значений в ней. Значения априорных вероятностей появления каждого из элементов для расчета не нужны.

1.2. Применение критерия серий для поиска регулярности

В данной работе мы проводили сравнение промоторных последовательностей ДНК из различных геномов с искусственно полученными периодическими последовательностями. Поскольку в расчетах используется искусственная последовательность, то применение стандартных формул критерия серий [18, 19] не позволяет получить статистику, имеющую нормальное распределение, поэтому для оценки статистической значимости сходства последовательностей проводилось имитационное моделирование (метод Монте-Карло).

В качестве результатов наблюдений выступают номера позиций, на которых в последовательности расположен некоторый символ (a, t, c или g).

Сравнение проводилось следующим образом. Исходная последовательность сканировалась с помощью окна длиной 500 нуклеотидов, перемещающегося по ней с шагом в 100 нуклеотидов. Для последовательности, попавшей в окно, создавалась искусственная периодическая последовательность путем задания позиций, на которых располагались определенные заранее нуклеотиды. Например, для длины периода = 5, позиции искусственной периодической последовательности были 1, 6, 11, 16, 21 и т.д. Эта последовательность сравнивалась с участком исходной последовательности, попавшим в окно, путем построения вариационного ряда для номеров позиций исследуемого символа так, как это было описано в разделе 1.1.

Например, если символ 'a' находился в исходной последовательности на позициях окна 7, 10, 14, 15 и 18, то вариационный ряд для искусственного периода = 5 имел вид: $\{1, 6, 7, 10, 11, 14, 15, 16, 18, 21\}$, что соответствовало ряду значений кодирующей переменной $\{0, 0, 1, 1, 0, 1, 1, 0, 1, 0\}$. Здесь значение кодирующей переменной равно 0 для периодической последовательности и 1 для исследуемой последовательности.

Обозначим число серий в полученном ряду как r_0 (в приведенном примере $r_0 = 7$). Следует отметить, что критерий серий позволяет проводить достоверное сравнение только выборок, сопоставимых по объему [19]. Если же, например, число символов в исследуемой последовательности будет намного больше (более чем в 2 раза) или намного меньше количества символов в искусственной периодической последовательности, то применение критерия серий не позволит обнаружить

периодичность при ее наличии. Модифицируем приведенный выше пример – пусть символ 'а' находился в исходной последовательности на позициях окна 6, 7, 8, 9, 12, 13, 14, 16, 17, 18, 19. В этом случае ряд значений кодирующей переменной будет иметь вид $\{0, 0, \underline{1}, \underline{1}, \underline{1}, \underline{1}, 0, \underline{1}, \underline{1}, \underline{1}, 0, \underline{1}, \underline{1}, \underline{1}, \underline{1}, 0\}$ (серии подчеркнуты). Данный пример показывает, что при увеличении числа символов в каждой из ячеек, образуемых символами искусственной периодической последовательности (т.е., находящихся между этими символами), число серий не увеличивается. Для больших длин периода (> 4) это может привести к тому, что некоторые последовательности, обладающие периодичностью, не будут выявлены.

В целях разрешения данной проблемы мы вносили дополнительные искусственные символы, таким образом, разбивая ячейки на несколько частей. Например, для периода = 5 положение искусственных символов могло быть следующим: 1, **3**, 6, **8**, 11, **13**, 16, **18**, 21 (жирным выделены новые символы). Мы рассматривали все возможные позиции размещения новых символов, в том числе при внесении более чем одного символа в каждую ячейку (внесение символов на соответствующие позиции проводилось одновременно для всех ячеек). При этом случаи, когда число искусственных символов превышало число символов исходной последовательности, исключались из рассмотрения в целях сохранения адекватности получаемых результатов. Подобное разбиение увеличивает число серий и позволяет выявить большее число периодических последовательностей. Обозначим максимальное число серий, полученное для исследуемого положения окна при всех рассматриваемых множествах искусственных символов, как r .

Для определения статистической значимости проводилось имитационное моделирование методом Монте-Карло. В данной работе для каждого положения окна проводилось 200 испытаний. В каждом испытании выполнялось случайное перемешивание символов исходной последовательности, после чего вновь полученная последовательность сравнивалась с искусственной периодической так, как это было описано выше. При этом вновь рассматривались все возможные множества искусственных символов, и выбиралось максимальное число серий. Таким образом, в результате применения метода Монте-Карло было получено 200 значений для максимального числа серий. После этого для полученного множества значений рассчитывалось среднее значение (μ_r) и стандартное отклонение (σ_r). Тогда формула для расчета статистической значимости будет иметь вид:

$$Z = \frac{r - \mu_r}{\sigma_r}. \quad (1)$$

Большое положительное значение статистики критерия (мы использовали значения ≥ 4.0 , выбор порогового значения рассмотрен в разделе 1.4) означает, что наблюдаемое число серий намного превосходит ожидаемое, то есть, вариационный ряд хорошо перемешан. Данный факт говорит о том, что распределения исследуемого символа в двух последовательностях являются схожими на статистически значимом уровне. В случае большого отрицательного значения мы говорим о слабой перемешанности ряда, или же о том, что символ по-разному распределен в исследуемых последовательностях. Если же значения Z небольшие по модулю, то на статистически значимом уровне мы не можем ничего утверждать о сходстве либо различии распределений символа.

Наличие числа серий, большего ожидаемого значения, позволяет сделать предположение о потенциальном наличии периодичности в исходной последовательности (возможно, размытой) или, в общем случае, о наличии регулярности (определение регулярности будет дано в разделе 1.5).

1.3. Расчет объединенной статистики

Выше было рассмотрено в качестве примера применение критерия серий для сравнения двух последовательностей на основании изучения перемешанности позиций одного символа в каждой из этих последовательностей. Однако для проведения более полного сравнительного анализа последовательностей необходимо рассматривать перемешанность по всем четырем символам. При этом проводится вычисление статистики Z (формула (1)) для каждой из букв в отдельности. Обозначим результаты для a , t , c и g как Z_a , Z_t , Z_c и Z_g , соответственно. Все указанные величины имеют нормальное распределение, что будет показано ниже. Воспользуемся свойством стандартного нормального распределения, чтобы получить новую суммарную величину, также распределенную нормально. Мы получаем:

$$Z_{sum} = \frac{Z_a + Z_t + Z_c + Z_g}{\sqrt{4}} = \frac{Z_a + Z_t + Z_c + Z_g}{2} \sim N(0;1) \quad (2)$$

Проводился расчет величины Z для каждого из нуклеотидов по формуле (1), а затем рассчитывалось объединенное значение Z_{sum} по формуле (2).

Все последовательности, обнаруженные путем сканирования с помощью окна, подвергались фильтрации таким образом, как это описано в следующем разделе.

Итак, в данной работе мы сравниваем две последовательности путем расчета для них значения Z_{sum} по формуле (2).

Также проводилось варьирование границ изучаемой последовательности, а именно, правая и левая ее границы независимо изменялись с шагом 2, и каждый раз рассчитывалась статистика Z_{sum} . Таким образом, изучались все подпоследовательности из данного окна с длиной от 50 до 500 нуклеотидов. Изменение границ было необходимо для определения в окне подпоследовательности, обладающей наибольшим значением статистики критерия из всех перекрывающихся друг с другом подпоследовательностей. В качестве выходных данных выдавались одна или несколько неперекрывающихся подпоследовательностей окна, значение статистики критерия для которых было больше порогового. Подпоследовательности с длиной, меньшей 50, не рассматривались, поскольку размер выборки в этом случае является недостаточно большим для получения достоверных результатов на выбранном уровне статистической значимости. Данный алгоритм применялся последовательно для значений длин периодов от 2 до 12 нуклеотидов.

Параметры сканирования последовательности (окно = 500, шаг внутри окна = 2) были эмпирически подобраны таким образом, чтобы обеспечить приемлемую скорость проведения вычислений, но при этом не пропустить последовательности, имеющие статистически значимое подобие с искусственной периодической последовательностью. При исследовании последовательностей промоторов было использовано статичное окно длины 500, соответствующее длине промотора (выбор исходных данных описан в разделе 1.6.).

Поскольку интерес представляют участки, имеющие значимое отклонение от среднего значения не только для всех четырех букв, но и для любого их подмножества (в том числе, и для одной буквы), то для каждой подпоследовательности из окна рассчитывались значения статистики критерия для всех возможных комбинаций букв (по формулам, аналогичным (2)), а затем из них выбиралось максимальное. Помимо прочего, проведение подобных расчетов позволяло избежать сильного влияния отрицательных значений статистики по одной или нескольким буквам на суммарный результат Z_{sum} .

Для проверки корректности предположения о нормальном распределении статистики критерия для одного символа, мы провели исследования спектра величины Z при сравнении последовательности, сгенерированной с помощью датчика случайных чисел (период датчика $\sim 2 \cdot 10^{18}$) с искусственными периодическими

последовательностями, имеющими различную длину периода в диапазоне 2–12. Была сгенерирована последовательность длиной 10 миллионов символов (то есть, приблизительно в 10 раз превышающая суммарную длину исследуемых последовательностей промоторов), при этом исходные частоты появления каждого из нуклеотидов были равны ($= 0.25$). Эти последовательности сравнивались с искусственной периодической с помощью метода, описанного в разделе 1.2. Полученный спектр приведен на рис. 1 для длины периода $= 4$. Среднее значение $= 0.0124$, стандартное отклонение $= 1.1267$, значений ≥ 5 не встречается. Число значений, больших 4, равно 3, максимальное из них составило 4.1982. Таким образом, мы видим, что предположение о нормальном распределении статистики критерия адекватно отражает существующие закономерности.

Для остальных длин периода были получены аналогичные результаты.

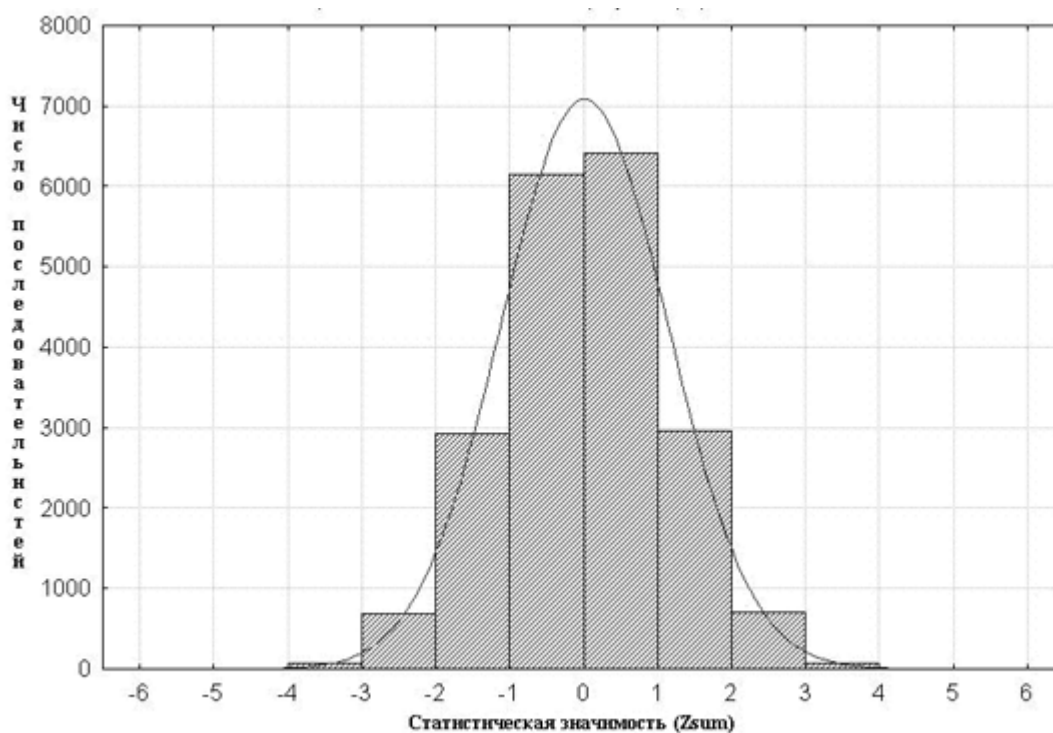


Рис. 1. Спектр величины статистики критерия для случайных последовательностей, полученный при проведении имитационного моделирования

1.4. Обработка найденных периодических последовательностей

На первом шаге обработки найденных последовательностей необходимо исключить их перекрытие между собой. Такое перекрытие (полное или частичное) может образоваться, если одна и та же последовательность была обнаружена при разных положениях сканирующего окна. Хранение всех таких последовательностей не дает никакой дополнительной информации, но при этом значительно замедляет дальнейшую их обработку, поэтому если две найденные последовательности перекрывались между собой более чем на 90%, мы оставляли только одну из них, имеющую большую значимость Z_{sum} , а другая исключалась из дальнейшего рассмотрения.

Мы используем критерий серий для сравнения исследуемой последовательности с искусственной периодической, имеющей некоторую длину периода p . Однако такое применение не позволяет сделать однозначный вывод о том, что величина Z_{sum} будет максимальной именно для этой длины периода, а не для какой-то другой. Довольно часто встречается ситуация, когда последовательность имеет значение Z_{sum} , большее порогового, как для длины периода $= 2$, так и для длины $= 4, 8$ и т.д. Таким образом, кратные периоды также могут давать большую статистическую значимость. Однако

для получения значимых и достоверных результатов необходимо определить, является ли полученное значение статистики критерия максимальным именно для исследуемой длины периода. Поэтому на данном шаге обработки результатов мы вычисляли значение Z_{sum} для длин периода 2–12 для каждой последовательности из множества оставшихся после проведения предыдущего шага обработки. В случае, если максимальное значение статистики критерия не соответствовало ранее обнаруженной длине периода, то последовательность исключалась из рассмотрения.

Кроме того, для обеспечения максимальной достоверности результатов был проведен еще один шаг фильтрации. Для каждой из изучаемых последовательностей мы построили спектр информационного разложения [14]. Метод информационного разложения позволяет выявить скрытую периодичность символьной последовательности при отсутствии вставок и делеций символов. В случае, если максимальное значение построенного спектра соответствовало длине периода, отличной от найденной для рассматриваемой последовательности, но при этом значение в спектре было ≥ 4.0 , мы исключали исследуемую последовательность из рассмотрения, поскольку в этом случае полученный результат нельзя было считать достоверным.

Множество последовательностей, прошедших все шаги фильтрации, рассматривалось нами в качестве множества результатов. Все данные, приведенные в разделе «Результаты и обсуждение», прошли вышеуказанные шаги фильтрации.

Пороговое значение статистики критерия было выбрано нами исходя из анализа случайных последовательностей, имевших общую длину, приблизительно в 10 раз превышающую суммарный размер анализируемых участков ДНК. В результате было выбрано значение $Z_{sum} = 4.0$. Как было отмечено выше, мы выяснили, что критерием серий на этой длине выявляется три случая регулярности со значением статистики критерия, равным или превышающим выбранное пороговое (для длины периода = 4; для остальных длин периода результаты аналогичны). В целях получения более точной оценки, мы провели такое же исследование для случайной последовательности длиной 100 миллионов символов. В этом случае было получено 39 регулярных участков со статистической значимостью, большей пороговой. Максимальное полученное значение составило 4.6442. Таким образом, мы можем сделать вывод, что на изучаемом множестве промоторов (общая длина ~ 1 млн. символов) следует ожидать наличие не более чем одной случайной последовательности, в которой критерием серий выявляется регулярность со статистической значимостью ≥ 4.0 . Отметим также, что случаев с $Z_{sum} \geq 5.0$ выявлено не было.

Если рассматривать пороговое значение в качестве аргумента нормального распределения, то оно соответствует вероятности $3.2 \cdot 10^{-5}$ обнаружения подпоследовательности с $Z_{sum} \geq 4.0$ в случайной нуклеотидной последовательности при проведении одного испытания.

При использовании меньших пороговых значений число регулярных подпоследовательностей случайной последовательности, выявленных критерием серий, не позволяло обеспечить достоверность получаемых результатов.

1.5. Определение регулярной последовательности. Схема регулярности

Назовем «регулярными последовательностями» последовательности ДНК, имеющие одинаковое распределение позиций их символов по интервалам, соответствующим заданной длине периода, с распределением символов в искусственной периодической последовательности с этой же длиной периода.

Для выявления сходства или различия распределения позиций символов исследуемой последовательности по заданным интервалам мы используем критерий серий. Кроме того, используются как стандартные искусственные периодические

последовательности, так и последовательности с добавлением дополнительных символов внутри отдельных периодов (см. раздел 1.2).

«Схемой регулярности» мы называем схематичное изображение расположения символов искусственной периодической последовательности, распределение символов которой оказалось наиболее близким к распределению в исследуемой последовательности. При этом в схему попадают только символы, значение статистики для которых было больше 2.0. Например, приведенная ниже схема означает, что максимальное значение статистики критерия было получено для искусственных периодических последовательностей, имеющих символ 'a' в 1 и 6 позиция периода, символ 'c' в 1 и 5 позициях, символ 'g' – во 2, 3, и 7 позициях. Регулярность в данном случае наблюдается по трем символам – “acg”.

```
a-----a--
c----c----
-gg---g-
```

Все последовательности, описанные в разделе «Результаты и обсуждение», являются регулярными. Мы выбрали данный термин потому, что хотя данные последовательности формально нельзя причислить к периодическим, они, тем не менее, обладают некоторой регулярной структурой, имеющей статистически значимое отличие от случайных последовательностей.

1.6. Входные данные и особенности их обработки

Исходные последовательности промоторов были получены из базы данных EPD [20], версия 93 (общее количество последовательностей, содержащихся в базе, исключая предварительные данные – 4809). Было выбрано 2236 последовательностей, представляющих все группы организмов. Для этого был использован запрос к базе данных «Получение представительного множества последовательностей, не имеющих большого уровня сходства друг с другом» (Representative set of not closely related sequences). Использование данной опции гарантирует, что никакие 2 последовательности из полученного множества не будут иметь уровень подобию, больший 50%. Диапазон извлекаемых промоторных участков последовательностей составлял (–500, +100), где +1 – сайт начала транскрипции. Однако поиск регулярности проводился только в последовательностях промоторов без учета прилегающих генов. Таким образом, исследуемое множество промоторов составили 2236 последовательностей длины 500 каждая.

Для каждой из последовательностей проводились расчеты, описанные в разделах 1.3. и 1.4. Поскольку интересным для исследования представляется именно регулярность (в частном случае, размытая периодичность), а не слабая перемешанность, то в качестве результатов мы приводим только последовательности, для которых значение статистики критерия было положительным и большим порогового.

2. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

2.1. Пример использования разработанного метода

В данном разделе приведен пример последовательности, в которой разработанный нами метод обнаруживает регулярность на статистически значимом уровне, тогда как другие методы (например, [14, 21, 22]), в том числе основанные на преобразовании Фурье [22], не могут обнаружить периодичность.

В качестве исходной была выбрана последовательность, обладающая совершенной периодичностью с длиной периода 5, имеющая длину 500 н.п. Эта последовательность состояла из 100 копий {attcg}. Затем в нее были случайным образом внесены вставки и делеции символов, а также замены, имитирующие мутации. На каждом шаге

рассчитывалась статистика критерия по формуле (2) для длины периода 5. На одном из шагов была получена следующая последовательность:

```
atcggcaatctgctatgtatcttctgattattcggatccggttcgcggtttcgatctccgattt
cgagtcctactccgaattggattccggttctaatacggcgtttcgatccttctcattcgattcc
gatttcgattttcgattcgattcgattagattaaaggattcgatcgaagtctctcttgcgatt
cggaggcccttcgatggatgattcgaaccagggttcggttaaatacgcattcattccccttgca
cctgtcgccttcgcatcattcctgattattcagatgtcgcattcgagggtctcgatgatgaatt
cgattgccggatcctttcgattgttcgatcgattcgggtcatcgatgtcggccattggtcac
ccatcgaattgggggggctaagatttcgatcgatactgcgcccctcattccgattcgtcgat
tcgattgatattttcgattcgatgagattcgttttcgatagatcattatcgatcgattttcaa
aaag
```

Длина последовательности = 500. Значение статистики критерия для нее составило $Z_{sum} = 5.50$ ($Z_a = 3.4757$, $Z_t = 1.7913$, $Z_c = 2.4999$, $Z_g = 3.2317$). В то же время, другими методами не было выявлено периодичности с периодом 5 на статистически значимом уровне в этой последовательности. Таким образом, разработанный нами метод является намного менее чувствительным к вставкам и делециям символов.

2.2. Обнаружение уже известных последовательностей, обладающих скрытой периодичностью

Исходя из определений скрытой периодичности и регулярности, можно сделать вывод о том, что большинство последовательностей, обладающих скрытой периодичностью, должны обладать также и регулярностью. Чтобы проверить данное предположение, мы провели поиск скрытой периодичности на исследуемом нами множестве промоторов с помощью метода информационного разложения [14]. Общее число периодических последовательностей с длиной периода 2–12 составило 109, из них 62 имели длину более 50 нуклеотидов (минимальная длина для обнаружения регулярности). Сравнение с результатами поиска регулярности показало, что все эти последовательности перекрываются на 80% и более с регулярными участками, полученными для соответствующей длины периода.

Таким образом, в данном случае поиск регулярности позволяет выявить в том числе и последовательности, обладающие скрытой периодичностью. В частности, это еще раз подтверждает, что регулярность является свойством последовательности, а не артефактом используемого метода ее поиска.

2.3. Пример регулярной последовательности промотора

Приведем пример регулярной последовательности, обнаруженной в промоторе.

>EP77714

КООРДИНАТЫ В ИСХОДНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ = 137, 216

ДЛИНА ПЕРИОДА = 10

УЧАСТОК ИСХОДНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ:

```
tttttgcagctttattagcgcgatgtcaccgtttatcagtgatggactgcacaaacgatatt
tctagactttctagtcga
```

СТАТ. ЗНАЧИМОСТЬ СИМВОЛОВ: $Z_a = 1.1427$ $Z_t = 2.9953$ $Z_c = 2.0747$
 $Z_g = 2.0747$

СУММАРНАЯ СТАТ. ЗНАЧИМОСТЬ: $Z = 4.1437$

ПОЗИЦИЯ ОКНА: 1 ЛЕВАЯ ГРАНИЦА = 136 ПРАВАЯ ГРАНИЦА = 216

СХЕМА:

```
tt-t-----t
c-----c---
g-----g---
```

Итак, регулярная последовательность была обнаружена в промоторе с кодом EP77714, ее длина составила 80 нуклеотидов. Для ее получения была использована искусственная периодическая последовательность с периодом 10. Следует отметить, что ни одна из полученных величин статистической значимости Z для отдельных букв не превышает порогового значения, то есть, регулярность является коллективным свойством всех четырех символов. Символ 'a' не отмечен в схеме регулярности, поскольку значение статистики критерия для него было меньше 2.0.

2.4. Результаты поиска регулярных последовательностей

Мы провели поиск регулярных последовательностей в промоторных последовательностях из базы данных EPD [20], используя следующие параметры сканирования: размер окна = 500, шаг внутри окна = 2. Таким образом, длина обнаруживаемых регулярных последовательностей находилась в диапазоне 50–500. Данные по количеству обнаруженных последовательностей приведены в Таблице 1.

Таблица 1. Число регулярных последовательностей, найденных в последовательностях промоторов для длин периода 2–12.

Длина периода	Число регулярных последовательностей
2	111
3	227
4	56
5	118
6	131
7	142
8	91
9	185
10	363
11	374
12	296

Общее число промоторов – 2236, число промоторов, содержащих регулярные последовательности – 1365

Таким образом, более 60% промоторов имеют регулярную структуру на статистически значимом уровне. В большинстве остальных промоторов также были обнаружены регулярные последовательности, однако значение статистики критерия для них было меньше порогового.

Несомненно, самым важным этапом изучения регулярности является ее биологическая интерпретация. Рассмотрим распределение длин регулярных последовательностей (табл. 2), а также их местоположение в промоторах.

Таблица 2. Распределение длин найденных регулярных последовательностей.

Диапазон длин последовательности, н.п.	Количество выявленных последовательностей
50 - 100	546
100 - 150	264
150 - 200	94

200 - 250	56
250 - 300	27
300 - 350	15
350 - 400	9
400 - 450	9
450 - 500	1

Наибольшая длина регулярной последовательности составила 476 н.п. Тем не менее, как видно из таблицы 2, большинство последовательностей имело длину в диапазоне 50–150. Поскольку при фильтрации перекрывающихся последовательностей мы отбирали те из них, которые обладали большей статистической значимостью, а не длиной, то подобное распределение длин является вполне ожидаемым. Использование именно такого критерия отбора позволяет установить более четкие границы регулярности и повысить достоверность получаемых результатов.

Более важным для понимания биологической значимости регулярности является изучение распределения местоположения регулярных последовательностей в промоторах. Мы разбили длину промотора на интервалы длиной 10 нуклеотидов и определили число последовательностей, попавших в каждый из этих интервалов. Поскольку минимальная длина регулярной последовательности составляла 50 н.п., то очевидно, что последовательность всегда попадала в несколько интервалов. Прежде всего мы хотели выяснить, является ли случайным распределение регулярных последовательностей по длине промотора. В случае, если данное распределение будет иметь экстремумы на определенных участках промотора, то существует возможность привязки регулярности к некоторой биологической функции, которой обладают эти участки.

Мы использовали имитационное моделирование (метод Монте-Карло) для определения, является ли распределение регулярных последовательностей случайным. Для этого мы случайным образом размещали регулярные последовательности по всей длине промотора. При этом длины регулярных последовательностей соответствовали их длинам в исходном множестве. Например, если было обнаружено 22 последовательности длиной 96, то такое же количество последовательностей этой длины случайным образом размещалось по длине промотора. Фактически, с помощью датчика случайных чисел определялась только точка начала последовательности. Ее координата могла изменяться в диапазоне $(1; 500-k)$, где k – длина регулярной последовательности, то есть последовательность не могла выйти за границы промотора.

Подобное размещение последовательностей повторялось 200 раз, при этом каждый раз определялось число последовательностей, попавших в интервалы длиной 10 (то есть последовательностей, которые содержали указанные интервалы целиком или частично). На основе полученных данных для каждого интервала были рассчитаны среднее значение (m) и дисперсия. После этого мы сравнили данные величины с реально полученным распределением. Интервалы, в которых реальное число регулярных последовательностей было больше ожидаемого на величину трех или более стандартных отклонений (σ), приведены в таблице 3.

Таблица 3. Статистика по интервалам, число регулярных последовательностей в которых значимо превысило ожидаемое значение.

Интервал (включая правую границу)	Наблюдаемое количество последовательностей	Диапазон ожидаемых значений количества последовательностей ($m-3\sigma; m+3\sigma$)	Отклонение реального числа последовательностей от среднего, σ
1 – 10	28	(0.00, 13.58)	9.0

10 – 20	60	(18.42, 56.69)	3.5
20 – 30	95	(44.48, 93.76)	3.2
400 - 410	319	(208.12, 291.51)	5.0
410 – 420	314	(193.67, 272.20)	6.2
420 – 430	308	(174.23, 252.37)	7.3
430 – 440	315	(154.10, 227.40)	10.2
440 – 450	294	(126.71, 201.35)	10.4
450 – 460	276	(98.61, 170.47)	11.8
460 – 470	220	(72.11, 134.18)	11.3
470 – 480	182	(46.21, 98.77)	12.5
480 – 490	112	(21.45, 61.11)	10.7
490 – 500	74	(0.00, 19.61)	9.2

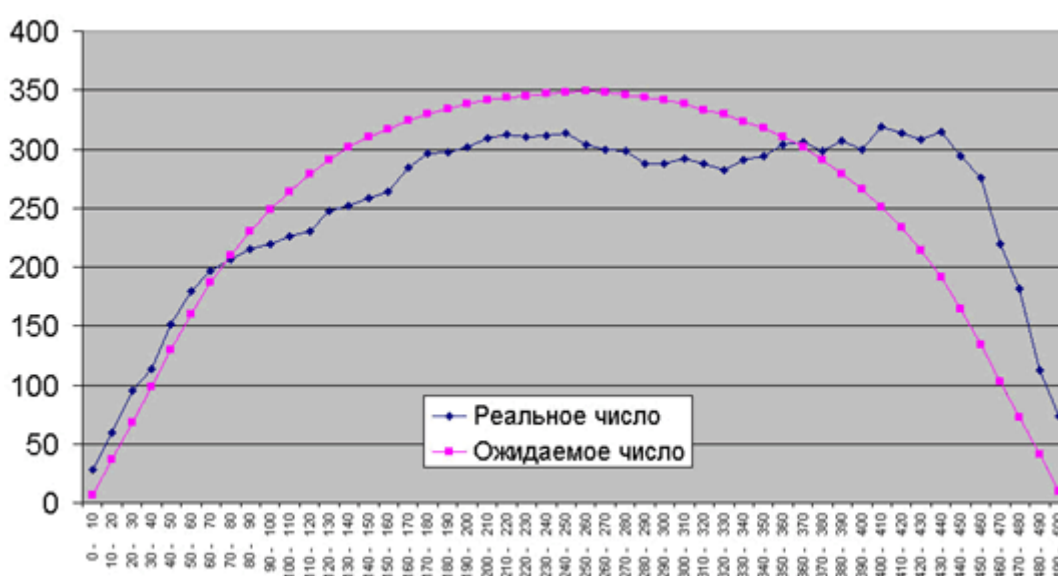


Рис. 2. Расположение регулярных последовательностей. Синим цветом показано реальное число последовательностей в каждом из интервалов, а красным – ожидаемое значение для случайного размещения последовательностей такой же длины.

Кроме того, в интервалах 110–160 и 250–330 наблюдалось отклонение около 3σ в отрицательную сторону. Поскольку на данных участках количество регулярных последовательностей отличается от ожидаемого намного меньше, чем в случае положительного отклонения, то мы не рассматриваем более подробно эти участки, считая, что они «наводятся» участками с положительным отклонением, то есть не имеют биологической значимости сами по себе. График расположения регулярных последовательностей по интервалам представлен на рис. 2.

Возможная биологическая роль участков, приведенных в таблице 3, обсуждается в следующем разделе.

2.5. Обсуждение

Тестирование разработанного математического метода показало, что он способен выявлять как явную, так и скрытую периодичность со сравнительно небольшим числом делеций или вставок. При наличии длинных вставок нуклеотидов в район ДНК с имеющейся регулярностью метод может не выделить такую подпоследовательность, регулярность в которой была бы на статистически значимом уровне.

Изучению промоторных последовательностей генов в настоящее время уделяется большое внимание, так как эти последовательности определяют активность генов в эукариотических и прокариотических клетках. Если будет возможно классифицировать промоторные последовательности по тем или иным количественным признакам последовательности и найти связь этих признаков с экспрессией генов в определенных клетках или в определенные моменты развития организма, то это может открыть путь к реконструкции генетических сетей в клетке ([23, 24]). С другой стороны, разработка все более точных алгоритмов поиска промоторных последовательностей может сделать более точным и поиск генов в геномах клеток эукариот ([23–26]), так как обнаружение промоторного участка точно указывает на начало гена. Для решения этих задач необходима разработка новых математических подходов, способных выявить в промоторных последовательностях закономерности, которые можно будет использовать как для точного поиска промоторных последовательностей, так и для их классификации.

В целях выяснения возможной биологической значимости, интервалы, богатые регулярными последовательностями, которые были выявлены в разделе 2.4, представляется интересным наложить на схему связывания факторов транскрипции. РНК-полимераза обычно связывается с промотором в районе $(-45, +5)$. Из работы [27] мы получаем, что участок промотора от -38 до $+5$ является местом связывания факторов транскрипции TFIIB $(-38, -32)$, TBP $(-31, -24)$, TFIIB $(-23, -17)$ и TAF1 $(-2, +5)$. В нашей системе координат это соответствует участку $(462, 505)$.

Сравнивая эти данные с таблицей 3, мы можем видеть, что районы связывания факторов транскрипции и РНК-полимеразы полностью приходятся на интервалы, в которых число регулярных последовательностей значительно превысило ожидаемое. Исходя из полученной картины, мы можем предположить, что найденная нами регулярность последовательностей важна для связывания факторов транскрипции и РНК-полимеразы с ДНК. Вероятно, для связывания белков с ДНК необходимо определенное регулярное чередование нуклеотидов в последовательности ([28, 29]). Такое чередование может быть связано с эволюционным родством различных транскрипционных факторов, в силу чего определенную схожесть имеют и последовательности ДНК, с которыми эти факторы могут связываться, что на уровне последовательности оснований ДНК может выявляться как регулярность. На статистически значимом уровне регулярность в районе $(-100, +1)$ была обнаружена примерно для 60% последовательностей. Для остальных последовательностей регулярность этих районов была замечена, но только она обладала уровнем $Z < 4.0$. Это может быть связано с тем, что набор транскрипционных факторов может меняться от одного промотора к другому, что может приводить как к изменению статистической значимости регулярности, так и к тому, что в данном интервале мы обнаруживаем регулярность последовательности ДНК в интервале $400-500$, а не $450-500$.

Кроме того, выявленная нами регулярность может быть связана с изгибом молекулы ДНК в районе связывания РНК-полимеразы [30, 31]. Определенная регулярность в чередовании нуклеотидов, выявленная нами, может создавать изгиб молекулы ДНК, что, возможно, облегчает сборку транскрипционного комплекса [32, 33].

ЗАКЛЮЧЕНИЕ

Нами предложена новая количественная характеристика свойств промоторных последовательностей – регулярность. Предложенный нами метод обнаружения регулярности, основанный на использовании критерия серий, позволяет обнаружить регулярность в условиях наличия относительно небольшого числа вставок и делеций. Кроме того, используя данный метод, можно также выявить последовательности,

обладающие скрытой периодичностью, которая является частным случаем регулярности.

Наличие регулярных последовательностей является общим явлением для большинства (более 60% на статистически значимом уровне) эукариотических промоторов, что позволяет провести их классификацию и, таким образом, сделать первый шаг в разработке новых универсальных методов поиска промоторов в неаннотированных последовательностях различных геномов.

СПИСОК ЛИТЕРАТУРЫ

1. Claverie J.M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* 1997. **6**. 1735-1744.
2. Bajic V.B., Chong A., Seah S.H., et al. An intelligent system for vertebrate promoter recognition. *IEEE Intell. Syst. Mag.* 1997. **17**. 64–70.
3. Davuluri R.V., Grosse I., Zhang M.Q. Computational identification of promoters and first exons in the human genome. *Nature Genet.* 2001. **29**. 412–417.
4. Ohler U., Liao G.C., Niemann H., et al. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* 2002. **3**. 0087.1–0087.12.
5. Pedersen A.G., Baldi P., Chauvin Y., et al. The biology of Eukaryotic promoter prediction: a review. *Comp. Chem.* 1999. **23**. 191–207.
6. Dieterich C., Grossmann S., Tanzer A., et al. Comparative promoter region analysis powered by CORG. *BMC Genomics.* 2005. **6(1):24**.
7. Bajic V.B., Seah S.H. Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res.* 2003. **13**. 1923–1929.
8. Knudsen S. Promoter2.0: for the recognition of PolIII promoter sequences. *Bioinformatics.* 1999. **15**. 356–361.
9. Solovyev V.V., Shahmuradov I.A. PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Res.* 2003. **31**. 3540–3545.
10. Scherf M., Klingenhoff A., Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* 2000. **297**. 599–606.
11. Xie X., Wu S., Lam K.-M., Yan H. PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. *Bioinformatics.* 2006. **22**. 2722–2728.
12. Matsuyama Y., Kawamura R. Promoter recognition for E. coli DNA segments by independent component analysis. *Proc. Comput. Syst. Bioinformatics Conf.* 2004. 686–691.
13. Bajic V.B., Tan S.L., Suzuki Y., et al. Promoter prediction analysis on the whole human genome. *Nat. Biotechnol.* 2004. **22**. 1467–1473.
14. Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method to analyze symbolical sequences. *Phys. Let. A.* 2000. **312**. 198–210.
15. Shelenkov A.A., Skryabin K.G., Korotkov E.V. Search and Classification of Potential Minisatellite Sequences from Bacterial Genomes *DNA Res.* 2006. **13(3)**. 89–102.
16. Shelenkov A.A., Korotkov A.E., Korotkov E.V. MMsat – a database of potential micro- and minisatellites. *Gene.* 2008. **409**. 53–60.
17. Hoel P.G. Introduction to Mathematical Statistics, 3rd ed. New-York: Wiley. 1966.
18. Браунли К.А. *Статистическая теория и методология в науке и технике*. Москва: Наука. 1977.
19. Sheskin D.J. *Handbook of Parametric and Nonparametric Statistical Procedures*, 2nd ed. New York: Chapman & Hall/CRC. 2000.

20. Schmid C.D., Perier R., Praz V., Bucher P. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res.* 2006. **34**. D82–5.
21. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 1999. **27**(2). 573–580.
22. Sharma D., Issac B., Raghava G.P., Ramaswamy R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation, *Bioinformatics.* 2004 **20**. 1405–1412.
23. Werner T. The state of the art of mammalian promoter recognition. *Brief Bioinform.* 2003. **4** (1). 22–30.
24. Fickett J.W., Hatzigeorgiou A.G. Eukaryotic promoter recognition. *Genome Res.* 1997. **7** (9). 861–78.
25. Novichkov P.S., Gelfand M.S., Mironov A.A. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics.* 2001. **17** (11). 1011–8.
26. Hertel K.J. Combinatorial control of exon recognition. *J. Biol Chem.* 2008. **283** (3). 1211–5.
27. Zhang M.Q. Computational analyses of eukaryotic promoters. *BMC Bioinformatics.* 2007. **8** (Suppl. 6). S3.
28. Ioshikhes I., Trifonov E.N., Zhang M.Q. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc Natl Acad Sci USA.* 1999. **96**(6). 2891–2895.
29. Kutuzova G.I., Frank G.K., Makeev V.I., et al. Fourier analysis of nucleotide sequences. Periodicity in *E. coli* promoter sequences. *Biofizika.* 1997. **42** (2). 354–62.
30. Tchermaenko V., Radlinska M., Lubkowska L., et al. DNA bending in transcription initiation. *Biochemistry.* 2008. **47** (7). 1885–1895.
31. Mizuno T. Static bend of DNA helix at the activator recognition site of the *ompF* promoter in *Escherichia coli*. *Gene.* 1987. **54**. 57–64.
32. Bolshoy A., Nevo E. Ecologic genomics of DNA: upstream bending in prokaryotic promoters. *Genome Res.* 2000. **10**. 1185–1193.
33. Ozoline O.N., Deev A.A., Trifonov E.N. DNA bendability – a novel feature in *E. coli* promoter recognition. *J. Biomol Struct Dyn.* 1999. **16** (4). 825–31.

Материал поступил в редакцию 17.03.2008, опубликован 17.04.2008.