Геномика и биоинформатика

УДК: 577.21

Множественное выравнивание промоторных последовательностей из генома человека

© **2020** Е.В. КОРОТКОВ[1, 2*], А.М. КАМИОНСКАЯ[1], М.А. КОРОТКОВА[2]

[1] *Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, 119071*

[2] *Национальный исследовательский Ядерный Университет, Москва, 115409*

**e-mail:* bioinf@yandex.ru

Поступила в редакцию 27.03.2020 г.
После доработки 20.04.2020 г.
Принята к публикации 28.07.2020 г.

Разработан новый алгоритм множественного выравнивания нуклеотидных последовательностей MAHDS. С его помощью впервые создано статистически значимое множественное выравнивание промоторных последовательностей из генома человека. На основании построенных выравниваний создано 25 классов промоторных последовательностей с объемом каждого класса больше 100 последовательностей. Классы промоторов могут быть использованы для поиска промоторных последовательностей в эукариотических геномах.

Ключевые слова: промотор, класс, динамическое программирование, геном человека

doi: 10.21519/0234-2758-2020-36-4-7-14

Промоторная последовательность как у прокариот, так и у эукариот расположена до первого основания первого кодона (обозначается как +1) кодирующей последовательности гена [1]. Выделяется так называемый коровый промотор длиной от 60 до 120 оснований с которым связывается РНК-полимераза [2, 3]. Более протяженный участок длиной 600 оснований от -499 до +100 содержит в своем составе коровый промотор, а также участки связывания различных транскрипционных факторов [4]. В данной работе область исследования ограничена только эукариотическими промоторными последовательностями. В состав промотора включаются некоторые мотивы, которые представляют собой короткие консервативные последовательности. Среди них известны так называемая TATA последовательность, которая занимает позиции от -31 до -26 нуклеотида [5], и B recognition element (BRE), который находится от -37 до -32 нуклеотида в последовательности промотора. Обнаружены короткие последовательности, которые обеспечивают связывание различных белковых факторов с промоторной последовательностью [6]. Многие из этих последовательностей приходятся на участок промотора от +1 до +40 нуклеотида. Последовательность промотора не обладает симметрией и это позволяет ДНК полимеразе начать транскрипцию в правильном направлении. Уровень транскрипции зависит от того, под контролем какого промотора находится ген. По активности, промоторы можно разделить на два класса. К первому можно отнести конститутивные промоторы, которые обеспечивают постоянный уровень транскрипции. Ко второму — регулируемые промоторы, при использовании которых скорость транскрипции зависит от присутствия определенных транскрипционных факторов или других внешних условий.

Промоторные последовательности сильно отличаются друг от друга [7]. Это связано с необходимостью управления уровнем активности многих генов. При инициации транскрипции осуществляется сборка комплекса, в состав которого входит РНК-полимераза, промоторная последовательность и еще ряд транскрипционных факторов [8]. Набор этих факторов может меняться от одного гена к другому, что в свою очередь связано с сильным разнообразием промоторных последовательностей. На сегодняшний день для различных геномов эукариот известны сотни тысяч промоторов, последовательности которых собраны в различных базах данных [4]. В данной

работе мы использовали базу данных EPD, https://epd.epfl.ch//index.php.

Однако, несмотря на такое большое количество промоторных последовательностей, найти статистически значимое множественное выравнивание между ними до сих пор не удавалось [9]. Это привело к тому, что идентифицировать (аннотировать) промоторы по нуклеотидной последовательности достаточно сложно. Типичная схема для аннотации состоит в построении статистически значимого множественного выравнивания. Затем это выравнивание используется для профильного анализа или для построения скрытой марковской модели. Такая схема приводит к низкому числу ложных позитивов (ошибка первого рода) при анализе геномов. Алгоритмы для предсказания промоторных последовательностей в настоящее время используют другие математические методы ввиду отсутствия статистически значимого множественного выравнивания, такие как TSSW[7], PePPER[10], G4PromFinder [11] и многие другие. В среднем, различные алгоритмы предсказывают один ложный промотор (ложный позитив) на 10^3–10^4 оснований ДНК, в то время как человеческий геном содержит приблизительно один ген на 10^5 оснований. В результате среди ложных предсказаний невозможно выделить истинный промотор. Таким образом, если мы хотим использовать математические методы для поиска промоторов, то мы должны уменьшить число ложных предсказаний промоторов (число ложных позитивов) примерно в 10^2 или в 10^3 раз. Если правильно предсказывать промоторные последовательности, то можно находить не выявленные ранее гены или неизвестные ранее точки инициации транскрипции. Возможность находить новые гены открывается ввиду того, координаты найденного промотора с некоторой точностью указывают на +1 позицию первого основания первого кодона гена. Поиск новых точек инициации транскрипции обусловлен тем, что промотор содержит в своем составе точку инициации транскрипции [1] и обнаружение нового промотора равносильно обнаружению новой точки инициации транскрипции с определенной точностью.

Ранее мы разработали математический метод для создания множественного выравнивания для сильно различающихся нуклеотидных последовательностей (MAHDS), который может быть использован на сайте http://victoria.biengi.ac.ru/mahds/auth. Под сильно различающимися последовательностями будем понимать последовательности, накопившие более 2,5 случайных замен (x) на один нуклеотид относительно друг друга. MAHDS позволяет строить статистически значимые выравнивания для x в интервале от 2,5 до 4,4. Это превосходит возможности разработанных ранее алгоритмов, так как они могут строить статистические значимые множественные выравнивания до x<2,5 [12]. MAHDS был применен для построения множественного выравнивания промоторных последовательностей из генома A.thaliana. По построенному множественному выравниванию мы оценили x=3,6 [12] для промоторных последовательностей этого генома. Такое количество замен оснований между промоторами показывает, что статистически значимое множественное выравнивание невозможно рассчитать существующими методами [12]. В данной работе мы построили множественное выравнивание для промоторных последовательностей, накопленных в банке данных EPD [4] из генома человека. Мы также разработали метод классификации промоторов на основе проведенного множественного выравнивания. Всего удалось обнаружить 25 классов промоторных последовательностей с объемом классов более 100 промоторов. Полученные классы могут использоваться для поиска промоторных последовательностей в геноме человека.

МАТЕМАТИЧЕСКИЕ МЕТОДЫ И АЛГОРИТМЫ

В работе мы использовали 22 694 промоторных последовательностей из генома человека, которые мы получили с сайта: https://epd.epfl.ch//index.php [13]. Обозначим множество этих последовательностей как S. Каждый промотор имеет длину равную 600-ам нуклеотидов и включает в свой состав последовательность от -499 до +100 относительно первого основания первого кодона (+1 позиция гена). Для классификации промоторных последовательностей мы вначале случайно выбирали 100 промоторных последовательностей (множество S_1) из 22 694 последовательностей. Использование только 100 последовательностей связано с необходимостью максимально ускорить выполнение генетического алгоритма (см. ниже в этом разделе). Затем для последовательностей из множества S_1 методом MAHDS мы строили множественное выравнивание. В основу метода MAHDS положено двумерное динамическое программирование. Все промоторы длиной N=600 оснований из множества S_1, для которых строится множественное выравнивание, объединяются в одну последовательность L. Основная идея метода MAHDS состоит в том, чтобы отказаться от прямого расчета множественного выравнивания путем любого сравнения последовательностей из множества S_1. Вместо прямого расчета мы проводим оптимизацию образов

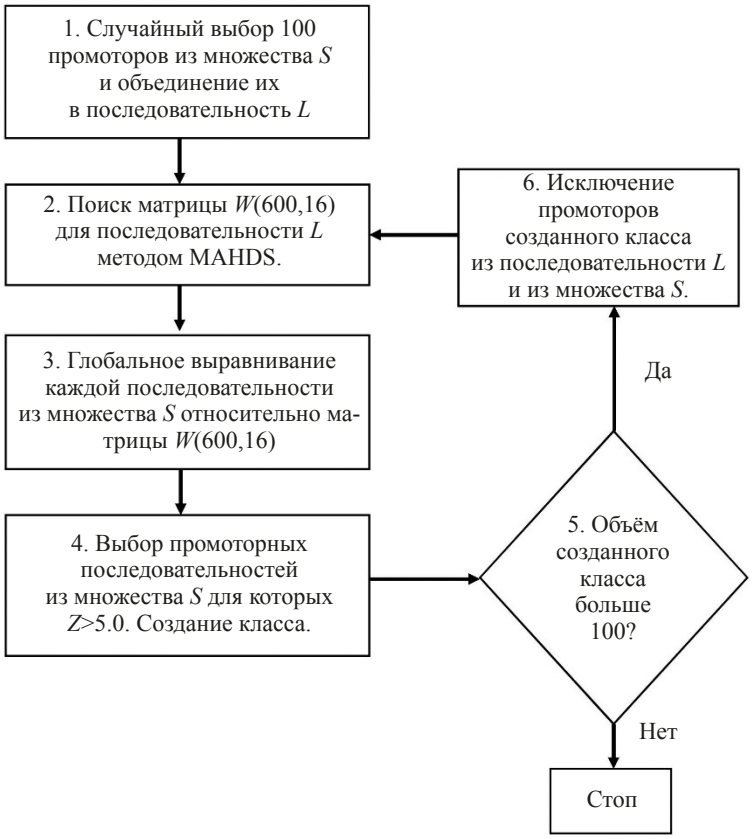

Рис. 1. Блок-схема алгоритма для создания классов промоторных последовательносей.

Fig. 1. The block diagram of the algorithm for creating classes of promoter sequences.

множественного выравнивания относительно последовательностей из множества S_1. В качестве таких образов выступают позиционно-весовые матрицы (ПВМ). Число столбцов в такой матрице равно длине промоторных последовательностей (т. е. 600 оснований), а число строк равно 16. В матрице 16 строк используются для того, чтобы рассматривать не отдельные основания, а пары оснований в позициях i и i+1, где i меняется от 1 до N-1. Это означает, что ПВМ учитывают, как подобие последовательностей во множестве S_1, так и корреляцию соседних оснований. Каждая такая матрица $M(k)$ есть определенный образ какого-то множественного выравнивания нуклеотидных последовательностей $S_1(k)$. Это множественное выравнивание можно построить, если построить глобальное выравнивание последовательности L относительно $M(k)$, например как это было описано ранее [14]. Однако, проблема заключается в том, что для созданной нами последовательности L мы не знаем соответствующую матрицу M и поэтому не можем построить оптимальное глобальное выравнивание для промоторов, собранных во множестве S_1. Однако эту проблему можно решить, если мы сможем оптимизировать какую-либо матрицу M, созданную случайным образом. В качестве параметра для оптимизации можно рассматривать значение функции сходства F_{max} при проведении выравнивания последовательности L относительно матрицы M [14]. Оптимизацию мы проводили генетическим алгоритмом [15]. В результате для случайной матрицы M(16,600) мы рассчитывали некую оптимальную матрицу M(16,600). Под оптимальной матрицей понимается такая матрица, которая имеет наибольшее значение F_{max}. Поэтому мы проделали оптимизационную процедуру для множества случайных матриц $M(k)$, k=1,10^3. В итоге среди полученных оптимальных матриц $M'(k)$ мы выбирали такую, которая имеет наибольшее значение F_{max}. Назовем эту матрицу как W(16,600). Теперь, уже зная последовательность L и матрицу W(16,600), мы при помощи обычного двумерного глобального выравнивания строим множественное выравнивание промоторов из множества S_1 [14]. В результате мы получали множественное выравнивание для которого мы рассчитали позиционно весовую матрицу W(16,600).

Очевидно, что не все промоторные последовательности из множества S могут иметь статистически значимое выравнивание с матрицей W(16,600). Поэтому мы выравнивали каждую промоторную последовательность из множества S относительно матрицы W(16,600) [16] (рис. 1, пункт 3).

Затем в первый класс мы выбирали такие промоторные последовательности из множества *S*, которые имеют неслучайное выравнивание относительно матрицы *W*(16,600) (рис. 1, пункт 4). Такие промоторные последовательности можно отнести к одному классу. Этот класс характеризуется матрицей *W*(16,600). Если объем созданного класса превышал 100 промоторных последовательностей, то класс считался созданным (рис. 1, пункт 5). Если он был менее 100 последовательностей, то создание классов заканчивалось.

Для создания дальнейших классов мы удаляли из множества *S* все последовательности, которые имеют неслучайное выравнивание относительно матрицы W(16,600) и получали новое множество *S'* (рис. 1, пункт 6). Далее для нового множества *S'* итеративно повторялась процедура создания класса, как это было сделано выше для множества *S*.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Всего удалось создать 25 статистически значимых классов промоторных последовательностей с объемами классов в интервале от приблизительно 2700 до 100. На рис. 2 показан размер созданных 25 классов.

Множественное выравнивание для множества последовательностей S_1 для каждого созданного класса можно найти на сайте https://yadi.sk/d/9tws31Cc6bJ6OA. Фрагмент матрицы *W*(16,600) для первого класса промоторных последовательностей показан на Таблице 1.

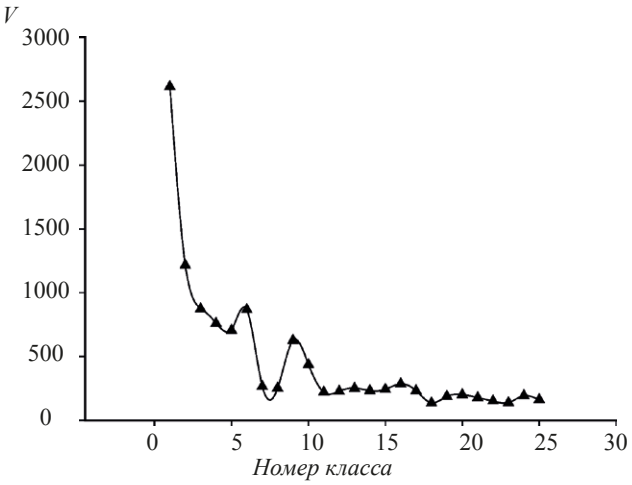

Рис. 2. Количество промоторных последовательностей в каждом из созданных классов. *V* обозначает число промоторных последовательностей в каждом созданном классе.

Fig. 2. The number of promoter sequences in each of the created classes. *V* denotes the number of promoter sequences in each created class.

Мы случайным образом перемешали все промоторные последовательности, собранные во множестве *S*. Затем создали классы по алгоритму, показанному на рис. 1. Размер первых 25 классов для случайных последовательностей составляет 12±9 последовательностей. Таким образом, число ложных позитивов не превышает 12% для последнего класса и менее 0,5% для первого класса. Это показывает, что мы выделили статистически значимые классы промоторных последовательностей.

Мы изучили консервативность оснований в первых двух классах промоторных последовательностей. Для этого мы определяли

Таблица 1

Фрагмент матрицы *W*(16,600) для первого класса промоторных последовательностей из генома человека. Здесь *W*(a), *W*(t), *W*(c), *W*(g), показывают значение матрицы *W* в позиции *i* при условии, что в позиции *i*+1 будут основание a, t, c и g соответственно.

Fragment of the matrix *W* (16.600) for the first class of promoter sequences from the human genome is shown. Here *W*(a), *W*(t), *W*(c), *W*(g) are the value of the matrix *W* in position *i* provided that at the position *i*+1 there are a base a, t, c and g, respectively.

i	Нуклеотид в позиции *i*	*W*(a)	*W*(t)	*W*(c)	*W*(g)
500	A	–1,51	–1,54	–1,68	0,47
500	T	–0,02	–0,07	1,61	0,40
500	C	–0,41	0,78	–0,37	–1,03
500	G	2,17	0,40	0,37	0,06
501	A	–0,48	–0,53	0,01	1,32
501	T	–1,54	–0,07	0,36	0,82
501	C	–0,41	1,61	–0,02	–1,03
501	G	–0,80	–1,70	1,07	0,77
502	A	–0,48	–0,53	–2,10	0,05
502	T	1,00	–1,08	0,36	–0,86
502	C	–0,41	1,61	–0,37	0,72
502	G	0,05	–1,28	0,72	1,83
503	A	–1,00	1,00	–1,25	1,32
503	T	–0,02	–0,57	–0,05	–0,44
503	C	–0,83	–0,47	0,33	–0,33
503	G	–0,80	–0,86	1,07	2,18
504	A	–1,00	–0,53	–1,68	0,47
504	T	–1,54	1,43	–0,89	–0,02
504	C	–1,25	0,36	–0,02	1,07
504	G	0,05	–0,02	1,07	1,47
505	A	–1,51	–2,05	–0,83	0,47
505	T	–1,04	0,43	–0,89	2,50
505	C	–0,83	0,78	–0,02	–1,03
505	G	–1,23	–1,28	2,47	2,53

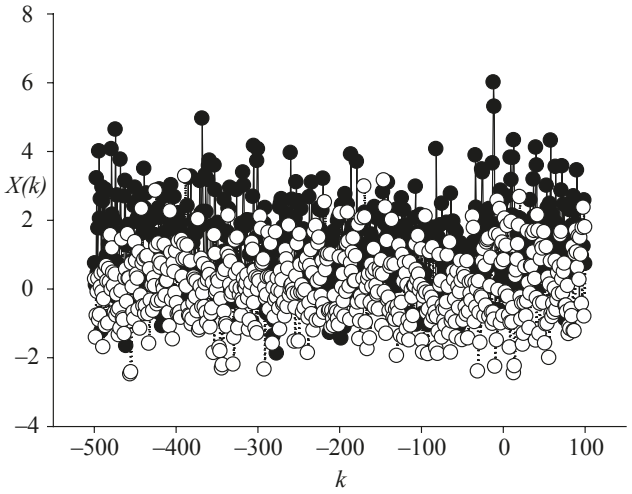

Рис. 3. Зависимость $X(k)$ от номера основания в промоторной последовательности k для первого класса созданных промоторных последовательностей. $X(k)$ есть аргумент нормального распределения, k есть номер нуклеотида в промоторной последовательности относительно первого основания кодона. Черные круги – последовательности промоторов, белые круги – случайные последовательности.

Fig. 3. The dependence of $X(k)$ on the base number in the promoter sequence k for the first class of created promoter sequences. $X(k)$ is argument of a normal distribution, k is the base number in the promoter sequence relative to the first base of the first codonof a gene. Black circles are promoter sequences, white circles are random sequences.

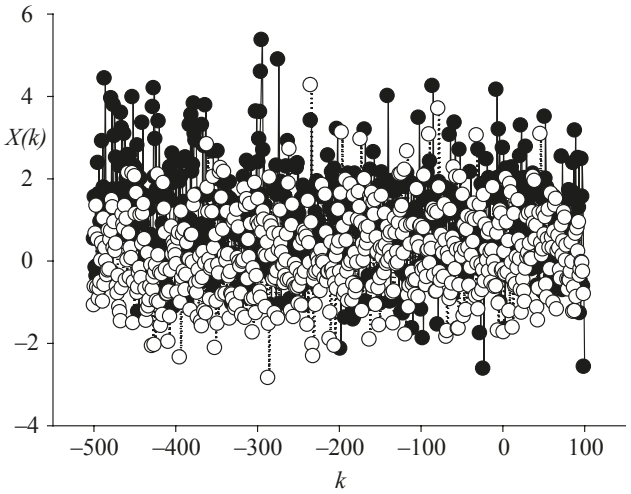

Рис. 4. Зависимость $X(k)$ от номера основания в промоторной последовательности k для второго класса созданных промоторных последовательностей. $X(k)$ есть аргумент нормального распределения, k есть номер нуклеотида в промоторной последовательности относительно первого основания кодона. Черные круги – последовательности промоторов, белые круги – случайные последовательности.

Fig. 4. The dependence of $X(k)$ on the base number in the promoter sequence k for the second class of created promoter sequences. $X(k)$ is argument of a normal distribution, k is the base number in the promoter sequence relative to the first base of the first codonof a gene. Black circles are promoter sequences, white circles are random sequences.

вероятности появления оснований как $p(i)=n(i)/K$ в соотвествующих выравниваниях с сайта https://yadi.sk/d/9tws31Cc6bJ6OA. Здесь $n(i)$ есть число оснований типа i во всех последовательностях из созданного множественного выравнивания. $i \in \{a,t,c,g\}$. K — суммарная длина всех последовательностей во множественном выравнивании. Затем мы рассчитывали вероятности встретить пару оснований $f(i,j)=p(i)p(j)$. Мы также рассчитывали $t(i,j,k)$. Это число пар (i, j) в позициях k и k+1. k меняется от 1 до N-1. Пусть у нас имеется Q последовательностей в выравнивании. Затем мы рассчитали:

$$x(i,j,k) = \frac{t(i,j,k) - Qf(i,j)}{\sqrt{Qf(i,j)(1-f(i,j))}} \qquad (1)$$

Здесь $Qf(i,j)$ — ожидаемое число соседней пары оснований (i,j). Всего для каждого k мы имеем 16 пар оснований. Затем мы рассчитывали:

$$Y(k) = \sum_{i=1}^{4}\sum_{j=1}^{4} x(i,j,k)^2 \qquad (2)$$

Величина $Y(k)$ имеет x^2 распределение с 15 степенями свободы. Затем мы рассчитывали $X(k)=\sqrt{2Y(k)} - \sqrt{2n-1}$, где n=15. Значение $X(k)$ тем больше, чем меньше эволюционная дивергенция нуклеотидов в позициях k и k+1. Максимум значений $X(k)$ достигается в том случае, когда каждой из позиций k и k+1 присутствует только одно основание ДНК для всех промоторных последовательностей.

Мы построили зависимость $X(k)$ для k от –499 до +99. для первых двух классов созданных промоторных последовательностей. Здесь позиции +1 соответствует первое основание первого кодона. Эти зависимости показаны на рисунках 3 и 4.

Одновременно мы построили зависимость $X(k)$ для случайно-перемешанных последовательностей. Это означает, что мы брали множественное выравнивание, созданное для класса промоторных последовательностей и случайно перемешивали нуклеотиды. При этом мы не меняли позиции вставок или делеций. Полученные графики для первых двух классов представлены на рис.3 и 4.

Промоторные последовательности самого большого класса из генома человека имеют много консервативных позиций для различных k (рис.3). Это можно заметить по тому, что $X(k)$

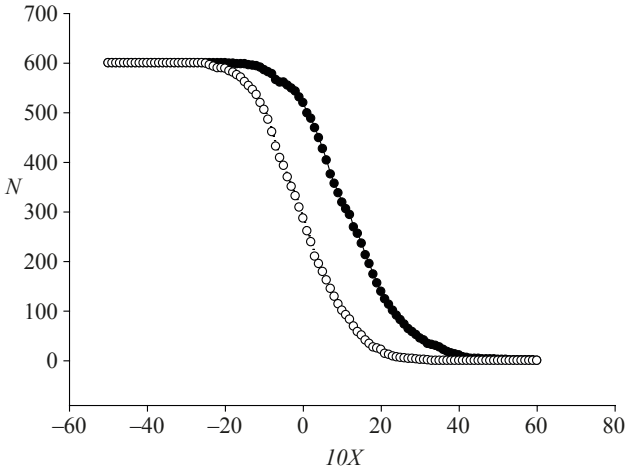

Рис. 5. Показана зависимость $N(X)$ для первого класса промоторных последовательностей из генома человека, X есть аргумент нормального распределения. Черные круги — промоторные последовательности первого класса, белые круги — случайные последовательности.

Fig. 5. The dependence $N(X)$ for the first class of promoter sequences from the human genome is shown. X is argument of a normal distribution. Black circles are promoter sequences of the first class, white circles are random sequences.

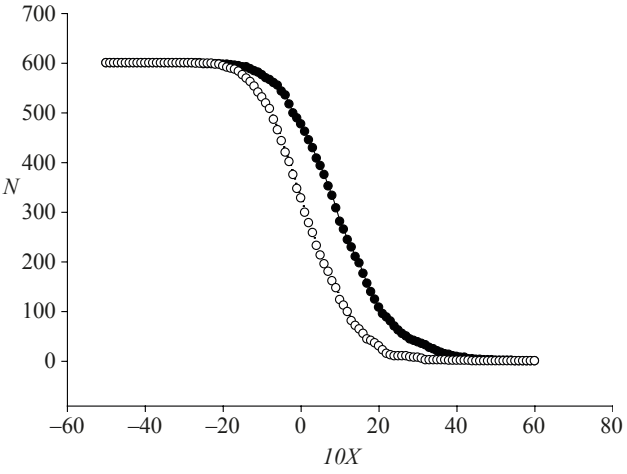

Рис. 6. Показана зависимость $N(X)$ для второго класса промоторных последовательностей из генома человека, X есть аргумент нормального распределения. Черные круги — промоторные последовательности первого класса, белые круги — случайные последовательности.

Fig. 6. The dependence $N(X)$ for the second class of promoter sequences from the human genome is shown. X is argument of a normal distribution. Black circles are promoter sequences of the first class, white circles are random sequences.

для случайно-перемешанных последовательностей значительно меньше, чем для промоторных последовательностей. Можно заметить консервативный участок промотора в районе k=–20, что соответствует TATA району промотора [17]. Также значимые пары оснований встречаются в позициях от +1 до +80. Это может быть связано как с инициацией транскрипции [1], так и с инициацией трансляции [18]. Вполне вероятно, что определенным факторам транскрипции или трансляции нужны определенные основания ДНК или только их некоторые комбинации (или мотивы). Однако, это явление не наблюдается для второго класса промоторных последовательностей. В них значения $X(k)$ для k>+1 незначительно превышает $X(k)$ для всех других k. Это может свидетельствовать о различных механизмах инициации транскрипции генов в геноме человека или о использовании различных факторов инициации как транскрипции, так и трансляции.

Важно также отметить, что последовательности промоторов содержат часто встречающиеся пары оснований ДНК по всей длине от –499 до +1. Это может показывать, что многие ДНК мотивы в этом районе важны для связывания факторов транскрипции [19].

Мы исследовали отличие значений $X(k)$, полученных для промоторных последовательностей от значений $X(k)$, рассчитанных для случайных последовательностей. Для этого мы посчитали величину $N(X)$, которая показывает число значений $X(k)$ от X до $+\infty$. Всего мы имеем 599 значений $X(k)$ для каждого класса промоторных последовательностей, поэтому $N(X)$ меняется от 0 до 599. Для первого и второго класса промоторных последовательностей зависимости $N(X)$ от X показаны на рисунках 5 и 6. $N(X)$ более чем в 2–5 раза больше для промоторных последовательностей, чем для случайных последовательностей для X от 0,0 до 3,0. При сравнении рисунков 3 и 5, а также 4 и 6 можно заметить, что отличие выравнивания промоторных последовательностей от случайных последовательностей происходит как за счет нескольких десятков $X(k)$>3,0, так и за счет нескольких сотен значений 0,0<$X(k)$<3,0. Это в свою очередь означает, что множественное выравнивание промоторных последовательностей обогащено неслучайными парами оснований.

Полученные результаты показывают, что возможно рассчитать статистически значимое множественное выравнивание промоторных последовательностей из генома человека. Полученные классы промоторов могут быть использованы для поиска промоторных последовательностей в как в геноме человека, так и в других

эукариотических геномах. Это может позволить более полно идентифицировать как сами гены, так и точки начала транскрипции.

ФИНАНСИРОВАНИЕ

Работа выполнена при частичной поддержке РФФИ (грант № 20-016-00057).

ЛИТЕРАТУРА

1. Nogales E., Louder R.K., He Y. Structural Insights into the Eukaryotic Transcription Initiation Machinery *Ann. Rev. Biophys.*, 2017, 46, (1), 59–83. doi:10.1146/annurev-biophys-070816-033751
2. Juven-Gershon T. et al. The RNA polymerase II core promoter — the gateway to transcription. *Curr. Opin. Cell. Biol*, 2008, 20, (3), 253–259. doi: 10.1016/j.ceb.2008.03.003
3. Smale S.T., Kadonaga J.T. The RNA polymerase II core promoter. *Ann. Rev. Biochem*., 2003, 72, 449–479. doi: 10.1146/annurev.biochem.72.121801.161520
4. Dreos R., Ambrosini G., Périer R.C., Bucher P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools *Nucleic Acids Res.*, 2015, 43, D92-D96. doi:10.1093/nar/gku1111
5. Lagrange T., Kapanidis A.N., Tang H., Reinberg D., Ebright R.H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Gen. Dev.*, 1998, 12, (1), 34–44. doi:10.1101/gad.12.1.34316406
6. Roeder R. The role of general initiation factors in transcription by RNA polymerase II *Trends Biochem*. Sci., 1996, 21, (9), 327–335. https://doi.org/10.1016/S0968-0004(96)10050-5
7. Solovyev V. V., Shahmuradov I.A., Salamov A.A. Identification of promoter regions and regulatory sites. *Methods Mol. Biol.* 2010, 674, 57–83. doi:10.1007/978-1-60761-854-6_5
8. Lee T.I., Young R.A. Transcription of Eukaryotic Protein-Coding Genes *Ann. Rev. Genet*, 2000, 34, 77–137. doi:10.1146/annurev.genet.34.1.77 Vol. 34, № 1. P. 77–137.
9. Zeng J., Zhu S., Yan H. Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. *Brief. Bioinform.*, 2009, 10, (5), 498–508. doi: 10.1093/bib/bbp027
10. De Jong A. et al. PePPER: a webserver for prediction of prokaryote promoter elements and regulons *BMC Genomics*, 2012, 13, (1), 299. doi:10.1186/1471-2164-13-299
11. Di Salvo M. et al. G4PromFinder: An algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs *BMC Bioinformatics*, 2018, 19, (1), 36. doi:10.1186/s12859-018-2049-x
12. Korotkov E.V. et al. Multiple alignment of promoter sequences from the A.thaliana genome *Neural Computing and Applications*. 2020, under consideration.
13. Dreos R. et al. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms *Nucleic Acids Res*, 2017, 45(D1):D51-D55. doi:10.1093/nar/gkw1069
14. Laskin A.A. et al. The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts. The NAD-binding protein sites *Mol Biol (Mosk)*, 2003, 37, (4), 663–673.
15. Pugacheva V., Korotkov A., Korotkov E. Search of latent periodicity in amino acid sequences by means of genetic algorithm and dynamic programming // *Stat. Appl. Genet. Mol. Biol.* 2016, 15, (5), 381–400. doi:10.1515/sagmb-2015–0079
16. Frenkel F.E., Chaley M.B., Korotkov E.V., Skryabin K.G. Evolution of tRNA-like sequences and genome variability // *Gene*, 2004, 335, 57–71. doi:10.1016/j.gene.2004.03.005
17. Patikoglou G.A. et al. TATA element recognition by the TATA box-binding protein has been conserved throughout evolution // *Gen. Dev*., 1999, 13, (24), 3217–3230. doi:10.1101/gad.13.24.3217
18. Hellen C.U.T., Sarnow P. Internal ribosome entry sites in eukaryotic mRNA molecules // *Gen. Dev*., 2001, 15, (13), 1593–1612. doi: 10.1101/gad.891101.
19. Smith N.C., Matthews J.M. Mechanisms of DNA-binding specificity and functional gene reg*ulation by transcription factors //* Curr. Opin.Struct. Biol. 2016, 38, 68–74. doi: 10.1016/j.sbi.2016.05.006

Multiple Alignment of Promoter Sequences from the Human Genome

E.V. KOROTKOV[1, 2*], A.M. KAMIONSKAYA[1], and M.A. KOROTKOVA[2]

[1] *Fundamentals of Biotechnology, Federal Research Centre, Russian Academy of Sciences, Moscow, 119071, Russia*

[2] *National Research Nuclear University, Moscow Engineering Physics Institute, Moscow, 115409, Russia*

**e-mail:* bioinf@yandex.ru

Received March 27, 2020
Revised April 20, 2020
Accepted July 28, 2020

Abstract–A new algorithm for multiple alignment of nucleotide sequences of MAHDS has been developed. A statistically significant multiple alignment of promoter sequences from the human genome was first created using this algorithm. Based on the constructed alignments, 25 classes of promoter sequences were created with the volume of each class exceeding 100 sequences. The classes of promoters can be used to search for promoter sequences in eukaryotic genomes.

Key words: promoter, class, dynamic programming, human genome.

Funding–The work was partially supported by the Russian Foundation for Basic Research (Grant no. 20-016-00057).

doi: 10.21519/0234-2758-2020-36-4-7-14