Математическая биология и биоинформатика
2016. Т. 11. № 1. С. 14–23. doi: 10.17537/2016.11.14

========== **БИОИНФОРМАТИКА** ==========

УДК 510.52:519.21

О количестве перекрытий слов в паттернах

Фурлетова Е.И.[*1], Ройтберг М.А.[1,2,3]**

[1]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, Россия*
[2]*НИУ Высшая школа экономики, Москва, Россия*
[3]*Московский физико-технический институт, Долгопрудный, Московская область, Россия*

Аннотация. Изучалась задача оценки количества перекрытий в паттерне – наборе слов в некотором алфавите A, имеющих одну и ту же длину m. Получены теоретические и экспериментальные оценки количества перекрытий для паттернов двух видов. Первый из них – это случайные паттерны, для которых верна равномерная вероятностная модель: все буквы в алфавите A и, соответственно, все слова длины m равновероятны. Доказано, что среднее количество перекрытий P для случайных паттернов, состоящих из n слов длины m, линейно зависит от размера паттерна n и не зависит от длины слов в паттерне. В проведенных компьютерных экспериментах отношение P/n менялось в пределах от 0.33 до 1.06; теоретические оценки этого отношения для тех же паттернов не превосходят 1.67. Вторым видом паттернов, изученных в статье, являются паттерны, заданные матрицами позиционных весов из базы данных HOCOMOCO и пороговыми весами. Для этих паттернов отношение количества перекрытий к количеству слов в экспериментах менялось от 0.004 до 1, для более половины паттернов это отношение меньше 0.1.

Ключевые слова: *перекрытие, паттерн, вхождение паттерна в последовательность.*

ВВЕДЕНИЕ

Работа направлена на изучение паттернов и их перекрытий – комбинаторных объектов, возникающих при анализе текстовых строк в различных областях науки. Паттерном (pattern) называется набор слов в некотором алфавите. Перекрытие (overlap, border, suffix-prefix) между двумя словами в паттерне (возможно, одинаковыми) – это непустое слово, которое является собственным (отличным от самого слова) началом первого и собственным концом второго слова, как это показано на рисунке 1. Перекрытия естественно возникают при поиске вхождений паттернов в текстовую последовательность. Например, использование информации о перекрытиях является основой таких алгоритмов поиска вхождений паттернов, как алгоритм Бойера-Мура [1], алгоритм Кнута-Морриса-Пратта [2], алгоритм Ахо-Корасика [3] и других. Подобные алгоритмы применяются при решении задач бионформатики, лингвистики, сжатия данных, кодирования и др.

В последнее время комбинаторные свойства слов, в том числе, перекрытия между ними и их применение, активно исследуются, обзор литературы по данной теме дан в

[*]furletova@lpm.org.ru
[**]roytberg@lpm.org.ru

работах [4, 5]. Большинство работ по перекрытиям направлено на эффективное построение множеств перекрытий для префиксов слов из данного паттерна (border array) [6], а также на использование перекрытий в алгоритмах поиска вхождений паттернов. Время работы этих алгоритмов часто зависит от количества перекрытий в анализируемом паттерне. Однако, насколько нам известно, исследование количества перекрытий в паттерне до сих пор не проводилось. Наша работа направлена на то, чтобы восполнить этот пробел. При этом рассматриваются только паттерны, состоящие из слов равной длины, такие паттерны возникают, например, во многих задачах молекулярной биологии [7].

Работа имеет следующую структуру. В главе "Основные определения" вводятся основные понятия, они используются в последующих главах. В главе "Среднее количество перекрытий в случайных паттернах" дано доказательство теоремы о среднем количестве перекрытий. Теорема показывает, что при равномерном распределении букв в алфавите среднее количество перекрытий среди паттернов, состоящих из n слов длины m, не превосходит $c \cdot n$, где c – некоторая константа. В главе "Компьютерные эксперименты" приводятся результаты компьютерных экспериментов, которые были проведены для более точной оценки константы c.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Алфавиты и слова

Пусть дан алфавит $A = \{a_1, \ldots, a_l\}$. Через A^k обозначается множество всех слов длины k в алфавите A; A^* – множество всех слов в A. Пусть $w = uvu'$, где $w, u, v, u' \in A^*$. Тогда v – подслово w, u – префикс (начало) w и u' – суффикс (конец) w. Отметим, что само слово является своим подсловом, префиксом и суффиксом. Префикс (суффикс) слова w называется *собственным*, если он отличен от w. Пустое слово ε является и префиксом, и суффиксом любого слова. Префикс и суффикс длины k слова w будем обозначать соответственно через $pref_k(w)$ и $suf_k(w)$. Через $w[i, j]$ будем обозначать подслово w, начинающееся в позиции i и заканчивающееся в позиции j. Мощность множества S будем обозначать через $||S||$. Длину (количество символов) слова w будем обозначать через $|w|$.

Паттерны и перекрытия

В данной работе паттерном мы будем называть набор слов, имеющих одну и ту же длину m. Число m будем называть длиной паттерна H и обозначать через $|H|$. Говоря неформально, в паттерн включаются слова, обладающие некоторым общим (с точки зрения рассматриваемой задачи) свойством. Отметим, что в ряде приложений рассматриваются и паттерны, содержащие слова разной длины. Требование одинаковой длины слов, составляющих паттерн, продиктовано задачами биоинформатики, в частности, задачами анализа сайтов связывания белков [7].

Слово w будем называть перекрытием в паттерне H, если существуют два слова $h_1, h_2 \in H$ такие, что слово w является непустым собственным префиксом h_1 и одновременно непустым собственным суффиксом h_2. Множество всех перекрытий в H будем обозначать через $OV(H)$.

Пример. Пусть задан паттерн H в алфавите {A, C, G, T}, состоящий из 8 слов длины 7, т.е. $||H|| = 8$ и $|H| = 7$:

$$H = \{\text{ACATATA}, \text{AGACACA}, \text{ATACACA}, \text{ATAGATA}, \text{CATTATA}, \text{CTTTCAC}, \text{CTTTGAC}, \text{TACCACA}\}.$$

Тогда множество непустых перекрытий в H будет следующим:

$$OV(H) = \{\text{A}, \text{AC}, \text{ACA}, \text{ATA}, \text{C}, \text{CA}, \text{TA}\}.$$

Математическая биология и биоинформатика. 2016. V. 11. № 1. doi: 10.17537/2016.11.14

На рисунке 1 показано перекрытие ATA между словами ACATATA и ATACACA из *H*.

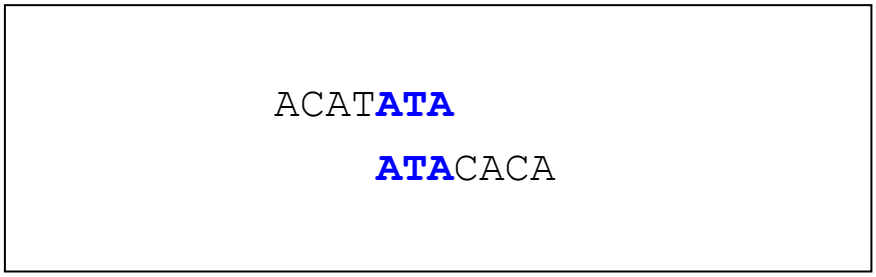

Рис. 1. Перекрытие между словами ACATATA и ATACACA.

Вероятностные распределения

На словах длины m в алфавите A может быть задано вероятностное распределение. Простейшим примером такого распределения является распределение Бернулли [8]. В этом случае каждой букве a из A сопоставляется вероятность p_a; вероятность слова определяется, как произведение вероятностей входящих в него букв (это соответствует предположению о независимости букв в различных позициях слова). Если вероятности всех букв равны, то и все слова будут иметь равную вероятность $1/||A||^m$. Это распределение называется равномерным. В нашей работе мы рассматриваем равномерное распределение. Более сложными вероятностными моделями для множества слов A^m являются марковские и скрытые марковские модели [8], мы планируем рассмотреть их в последующих работах.

Распределение на множестве слов индуцирует распределение на множестве паттернов. При равномерном распределении все паттерны, состоящие из одинакового количества слов одинаковой длины, равновероятны.

СРЕДНЕЕ КОЛИЧЕСТВО ПЕРЕКРЫТИЙ В СЛУЧАЙНЫХ ПАТТЕРНАХ

В данном разделе мы исследуем среднее число перекрытий $overlap\_avg(n, m)$ среди всех паттернов мощности n и длины m при равномерном вероятностном распределении для слов длины m. Множество всех таких паттернов будет обозначаться $Pat(n, m)$. Далее числа n и m будем считать фиксированными. Так как мы рассматриваем паттерны, содержащие фиксированное количество слов, то можно считать, что все рассматриваемые паттерны равновероятны. Таким образом, нашей целью будет вычисление величины

$$overlap\_avg(n,m) = \frac{overlap\_sum(n,m)}{\| Pat(n,m) \|},$$

где $overlap\_sum(n,m) = \sum_{H \in Pat(n,m)} \| OV(H) \|$ – суммарное количество перекрытий во всех паттернах из $Pat(n, m)$.

Теорема. *Пусть на множестве A^m слов длины m в алфавите A задано равномерное вероятностное распределение, причем $||A|| \geq 2$. Тогда среднее количество $overlap\_avg(n, m)$ перекрытий среди всех паттернов размера n и длины m не превосходит cn, где $n \geq 1$; $m \geq 2$ и*

$$c = \frac{\| A \| + 1}{\| A \| - 1} \sqrt{\frac{\| A \|^m}{\| A \|^m - 1}}. \qquad (1)$$

Доказательство. Нам понадобятся следующие обозначения:

– $OV(H, k)$ – количество перекрытий длины k в паттерне H;

– $overlap\_sum(n,m,k) = \sum_{H \in Pat(n,m)} \| OV(H,k) \|$ – суммарное количество перекрытий длины k среди всех паттернов из $Pat(n, m)$;

– $overlap\_avg(n,m,k) = \dfrac{overlap\_sum(n,m,k)}{\| Pat(n,m) \|}$ – среднее количество перекрытий длины k среди всех паттернов из $Pat(n, m)$.

Кроме того, наряду с множеством $Pat(n, m)$ мы рассмотрим множество $Ord(n, m)$, состоящее из упорядоченных наборов из n различных слов длины m. Так как распределение на множестве слов равномерное, то элементы множества $Ord(n, m)$ также будем считать равновероятными. Величины $overlapOrd\_avg(n, m)$, $overlapOrd\_sum(n, m)$, $overlapOrd\_avg(n, m, k)$, $overlapOrd\_sum(n, m, k)$ определяются аналогично величинам $overlap\_avg(n, m)$, $overlap\_sum(n, m)$, $overlap\_avg(n, m, k)$ и $overlap\_sum(n, m, k)$.

Замечание.

1) $\| Pat(n,m) \| = C^n_{\|A\|^m}$, где $C^n_{\|A\|^m}$ – число сочетаний из $\|A\|^m$ по n;

2) $\| Ord(n,m) \| = n! \| Pat(n,m) \| = \| A \|^m (\| A \|^m - 1) \cdot \ldots \cdot (\| A \|^m - n + 1)$.

Доказательство теоремы основано на леммах 1–4.

Лемма 1.

1) $overlap\_avg(n,m) = \sum_{k=1}^{m-1} overlap\_avg(n,m,k)$;

2) $overlapOrd\_avg(n,m) = \sum_{k=1}^{m-1} overlapOrd\_avg(n,m,k)$.

Доказательство. Мы приведем доказательство только первого утверждения, доказательство второго утверждения аналогично.

$$overlap\_avg(n,m) = \frac{overlap\_sum(n,m)}{\| Pat(n,m) \|} = \sum_{k=1}^{m-1} \frac{overlap\_sum(n,m,k)}{\| Pat(n,m) \|} = \sum_{k=1}^{m-1} overlap\_avg(n,m,k).$$

Пусть $H = \{h_1,\ldots,h_n\} \in Ord(n, m)$. Введем дополнительные обозначения:

– $\lambda_{i,j}(H,k) = \begin{cases} 1, & \text{если } pref_k(h_i) = suf_k(h_j), \\ 0, & \text{иначе.} \end{cases}$;

– $\Delta_{i,j}(n,m,k) = \{H \in Ord(n,m) \mid \lambda_{i,j}(H,k) = 1\}$.

Говоря неформально, $\Delta_{i,j}(n, m, k)$ – множество всех упорядоченных наборов $H \in Ord(n, m)$, в которых префикс длины k i-го слова и суффикс длины k j-го слова совпадают.

Отметим, что:

1) $\| \Delta_{i,j}(n,m,k) \| = \sum_{H \in Ord(n,m)} \lambda_{i,j}(H,k)$;

2) $\| OV(H,k) \| \le \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_{i,j}(H,k)$.

Последнее неравенство следует из того, что каждой тройке $< i, j, H >$ такой, что $\lambda_{i,j}(H, k) = 1$, соответствует перекрытие длины k, причем разным тройкам может соответствовать одно и то же перекрытие.

Лемма 2. $overlapOrd\_sum(n,m,k) \le \sum_{i=1}^{n}\sum_{j=1}^{n} \| \Delta_{i,j}(n,m,k) \|$.

Доказательство. $overlapOrd\_sum(n,m,k) = \sum_{H \in Ord(n,m)} \| OV(H,k) \| \le$

$$\le \sum_{H \in Ord(n,m)} \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{i,j}(H,k) = \sum_{i=1}^{n}\sum_{j=1}^{n} \sum_{H \in Ord(n,m)} \lambda_{i,j}(H,k) = \sum_{i=1}^{s}\sum_{j=1}^{s} \| \Delta_{i,j}(n,m,k) \|.$$

Лемма 3. $\| \Delta_{i,j}(n,m,k) \| \le \frac{a \| Ord(n,m) \|}{\| A \|^k}$, где $a = \frac{\| A \|^m}{\| A \|^m - 1}$.

Доказательство. Пусть $H = \{h_1, \cdots, h_n\} \in \Delta_{i,j}(n,m,k)$, то есть префикс длины k слова h_i и суффикс длины k слова h_j совпадают. Рассмотрим отдельно случаи, когда $i = j$ и $i \ne j$.

1. Пусть $i = j$. Тогда слово h_i полностью определяется своим началом длины $m - k$. Поэтому

$$\| \Delta_{i,i}(n,m,k) \| \le \| A \|^{m-k} n_i,$$

где n_i – количество наборов $H \in Ord(n,m)$ с фиксированным элементом h_i, $n_i = \frac{\| Ord(n,m) \|}{\| A \|^m}$. Отсюда

$$\| \Delta_{i,i}(n,m,k) \| \le \frac{\| Ord(n,m) \|}{\| A \|^k}.$$

2. Пусть $i \ne j$ и $i < j$. Пара (h_i, h_j) полностью определяется словом h_i и началом слова h_j длины $m - k$. Тогда существует не более чем $\|A\|^m \cdot \|A\|^{m-k}$ таких пар. Таким образом,

$$\| \Delta_{i,j}(n,m,k) \| \le \| A \|^m \| A \|^{m-k} n_{i,j},$$

где $n_{i,j}$ – количество паттернов $H \in Ord(n,m)$ с фиксированными элементами h_i и h_j. Отметим, что в $Ord(n, m)$ порядок слов в наборах важен, т.е. наборы, которые состоят из одинаковых слов, расположенные в разном порядке, являются разными. Очевидно,

$$n_{i,j} = (\| A \|^m - 2) \cdot \ldots \cdot (\| A \|^m - n + 1) = \frac{\| Ord(n,m) \|}{\| A \|^m (\| A \|^m - 1)},$$

где множитель $(\| A \|^m - i)$ обозначает количество способов выбрать очередное слово при генерации паттерна, при условии, что i слов уже выбрано. Следовательно,

$$\| \Delta_{i,j}(n,m,k) \| \le \frac{\| A \|^{m-k} \| Ord(n,m) \|}{(\| A \|^m - 1)} = \frac{\| A \|^m}{\| A \|^m - 1} \frac{\| Ord(n,m) \|}{\| A \|^k} \le \frac{a \| Ord(n,m) \|}{\| A \|^k}.$$

Следствие. $overlapOrd\_sum(n,m,k) \le \frac{an^2 \| Ord(n,m) \|}{\| A \|^k}$.

Лемма 4. $overlap\_sum(n,m,k) \le \frac{an^2 \| Pat(n,m) \|}{\| A \|^k}$, где $a = \frac{\| A \|^m}{\| A \|^m - 1}$.

Математическая биология и биоинформатика. 2016. V. 11. № 1. doi: 10.17537/2016.11.14

Доказательство. Так как для $Pat(n, m)$ порядок слов в наборах не важен, то каждому набору $H \in Pat(n,m)$ соответствует $n!$ наборов из $Ord(n, m)$. Эти наборы имеют одинаковое число перекрытий длины k, равное $\|OV(H, k)\|$. Отсюда

$$overlap\_sum(n,m,k) = \frac{overlapOrd\_sum(n,m,k)}{n!}.$$

Используя следствие из леммы 3, получаем:

$$overlap\_sum(n,m,k) \le \frac{an^2 \| Ord(n,m) \|}{n! \| A \|^k} = \frac{an^2 \| Pat(n,m) \|}{\| A \|^k}.$$

Следствие. $overlap\_avg(n,m,k) \le \min\left[\| A \|^k, \frac{an^2}{\| A \|^k} \right]$.

Доказательство. Число различных слов длины k над алфавитом A равно $\| A \|^k$. Поэтому $overlap\_avg(n, m, k) \le \| A \|^k$. Используя лемму 4, получаем:

$$overlap\_avg(n,m,k) = \frac{overlap\_sum(n,m,k)}{\| Pat(n,m) \|} \le \frac{an^2}{\| A \|^k}.$$

Окончание доказательства теоремы.

$$overlap\_avg(n,m) = \sum_{k=1}^{m-1} overlap\_avg(n,m,k) \le \sum_{k=1}^{m-1} \min\left[\| A \|^k, \frac{an^2}{\| A \|^k} \right].$$

Решением уравнения $\frac{an^2}{\| A \|^k} = \| A \|^k$ является $t = [\log_{\| A \|}(\sqrt{a} \cdot n)]$. Тогда:

- $\frac{an^2}{\| A \|^k} > \| A \|^k$, при $k < t$;
- $\frac{an^2}{\| A \|^k} < \| A \|^k$, при $k > t$.

Кроме того, $\sqrt{a} \cdot n = \| A \|^t$. Тогда

$$overlap\_avg(n,m) = \sum_{k=1}^{m-1} overlap\_avg(n,m,k) \le$$

$$\le \sum_{k=1}^{m-1} \min\left[\| A \|^k, \frac{an^2}{\| A \|^k} \right] = \sum_{k=1}^{t} \| A \|^k + \sum_{k=t+1}^{m-1} \frac{an^2}{\| A \|^k} = \sum_{k=1}^{t} \| A \|^k + \sqrt{a} \cdot n \sum_{k=t+1}^{m-1} \frac{\| A \|^t}{\| A \|^k}.$$

Найдем суммы геометрических прогрессий:

$$\sum_{k=1}^{t} \| A \|^k = \frac{\| A \| (\| A \|^t - 1)}{\| A \| - 1} \le \frac{\| A \|}{\| A \| - 1} \sqrt{a} \cdot n = c_1 \sqrt{a} \cdot n;$$

$$\sum_{k=t+1}^{m-1} \frac{1}{\| A \|^{k-t}} = \sum_{i=1}^{m-t-1} \frac{1}{\| A \|^i} = \frac{(\| A \|^{m-t-1} - 1)}{\| A \|^{m-t-1} (\| A \| - 1)} \le \frac{1}{\| A \| - 1} = c_2.$$

Отсюда

$$overlap\_avg(U_1) \le (c_1 + c_2) \sqrt{a} \cdot n = cn,$$

где $c = (c_1 + c_2)\sqrt{a} = \frac{\| A \| + 1}{\| A \| - 1} \sqrt{\frac{\| A \|^m}{\| A \|^m - 1}}$.

Замечание. При фиксированном размере алфавита коэффициент c имеет наибольшее значение при $m = 2$. Пусть $m = 2$. Тогда, если $||A|| = 2$, то $c \approx 3.47$; если $||A|| = 4$, то $c \approx 1.73$ и если $||A|| = 20$, то $c \approx 1.11$.

КОМПЬЮТЕРНЫЕ ЭКСПЕРИМЕНТЫ

Равномерно распределенные паттерны

Для оценки значения c в зависимости от m и n мы провели компьютерные эксперименты. Рассматривался четырехбуквенный алфавит, т. е. $||A|| = 4$. Предполагалось, что все буквы в алфавите A равновероятны. Для различных пар m, n мы строили по 100 паттернов из $Pat(n, m)$, считая появление всех слов в выбранном паттерне равновероятным. Для каждого построенного паттерна H затем вычислялось значение $||OV(H)||$, среднее значение этих величин (ниже это значение обозначается $overlap\_avgExp(n, m)$) и отношение

$$cExp(n, m) = \frac{overlap\_avgExp(n, m)}{n}.$$

Для удобства сравнения результатов, полученных при различных значениях m, мы вместо пары (n, m) описываем каждый набор паттернов парой (F, m), где $F = n/4^m$ – доля слов, включаемых в каждый из рассматриваемых паттернов от всех слов длины m. Далее будем обозначать через $cExp(F, m)$ величину $cExp(n, m)$, где $n = F \cdot 4^m$.

Нами рассматривались следующие значения параметра F (всего 175 значений):

1) $F = 0.1+0.01*i$, где $i = 1,\ldots,75$;

2) $F = 0.001*i$, где $i = 1,\ldots,100$.

Для каждого из указанных значений F из диапазона [0.1, 0.85], рассматривалось 4 значения параметра m: 5, 6, 7, 8; для каждого из значений F из диапазона [0.001, 0.1], рассматривалось 6 значений параметра m: 5, 6, 7, 8, 9, 10. Таким образом, всего было рассмотрено $4*75 + 6*100 = 900$ пар параметров (F, m) и для каждой пары было построено по 100 паттернов. Для каждой такой серии паттернов было вычислено отношение $cExp(F, m)$ (см. выше). Более подробные сведения о рассмотренных наборах паттернов приведены в Приложении 1.

На рисунке 2 представлены данные, соответствующие значениям параметра F из диапазона [0.1, 0.85], на рисунке 3 – данные, соответствующие значениям параметра F из диапазона [0.001, 0.1]. Каждая кривая показывает зависимость $cExp(F, m)$ от F для определенного значения m.

На рисунке 2 кривые практически совпадают. Все кривые монотонно убывают, наибольшие значения приближенно равны единице. Если далее продолжить кривые на рисунке 2 для $F = 0.86, \ldots, 1$, то они примут наименьшее значение при $F = 1$. При $F = 1$ паттерн состоит из всех слов длины m и, соответственно, его перекрытия – все непустые слова длины меньшей m. Тогда для всех m отношение количества перекрытий к размеру паттерна равно 0.3(3). Отметим, что экспериментально вычисленное значение $cExp(F, m)$ для всех рассмотренных пар (F, m) не превосходит 1.06. Теоретическое значение c при $||A|| = 4$ и $m \geq 5$ приближенно равно 1.67 (см. формулу (1)).

Линии на рисунке 3 при $m = 8$, 9 и 10 близки друг к другу и хорошо приближаются прямой $y = 1 – 0.8294x$. При меньших значениях m присутствуют заметные колебания в вычисленных величинах.

Математическая биология и биоинформатика. 2016. V. 11. № 1. doi: 10.17537/2016.11.14

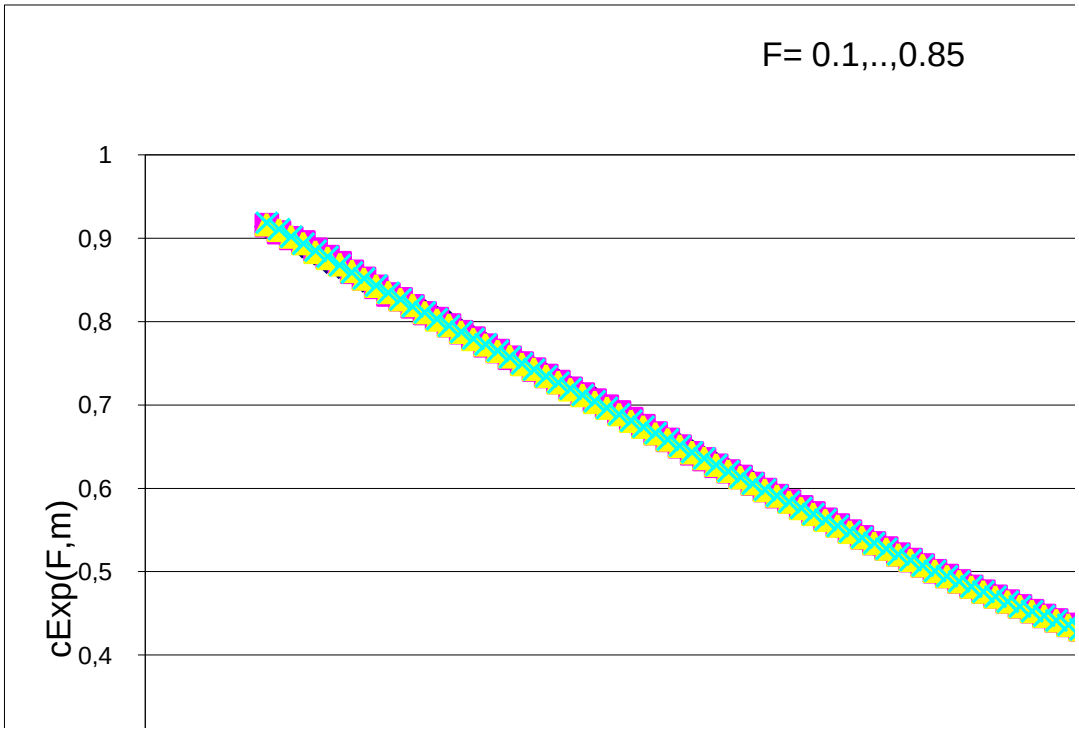

Рис. 2. Зависимость экспериментально полученного среднего значения *cExp*(*F*, *m*) константы *c* от величины *F* = 0.1, …, 0.85 и *m* = 5, 6, 7, 8.

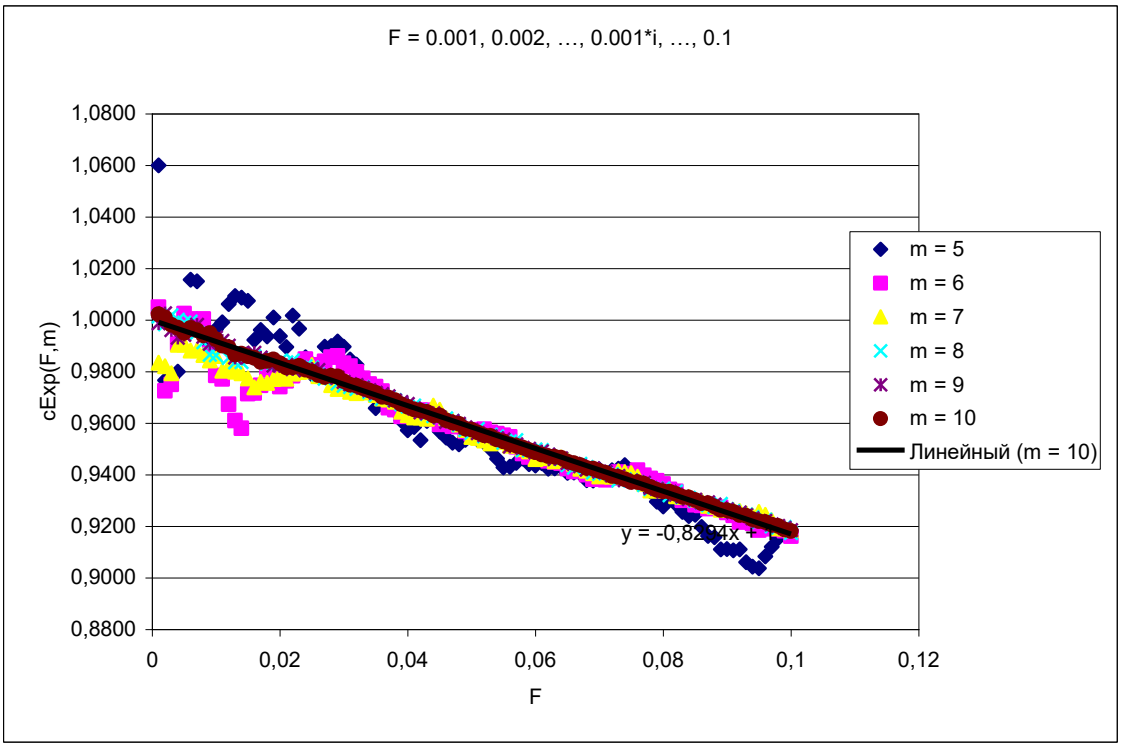

Рис. 3. Зависимость экспериментально полученного среднего значения *cExp*(*F*, *m*) константы *c* от величины *F* = 0.001, 0.002…, 0.1 и *m* = 5, 6, 7, 8, 9, 10.

Паттерны, заданные матрицами позиционных весов

Мы также рассмотрели паттерны, заданные матрицами позиционных весов (МПВ, PWM) из базы данных HOCOMOCO [9]. Данные матрицы описывают факторы

регуляции транскрипции в геноме человека. Было рассмотрено 355 матриц. Напомним, что МПВ размера $m \times \|A\|$ каждому слову длины m в A сопоставляет некоторое число (вес). Пара (МПВ, t), где t – порог, задает паттерн, состоящий из всех слов длины m, веса которых больше или равны порогу t. Для каждой из матриц МПВ был выбран порог t, таким образом, чтобы паттерн, заданный парой (МПВ, t) содержал 0.03 % от всех слов длины m. В итоге, было построено 355 паттернов, длины которых равны 7, …, 15. Для каждого паттерна H было вычислено отношение $cPWM = \|OV(H)\| / \|H\|$. Для более половины паттернов $cPWM < 0.1$. Наименьшее и наибольшее значения $cPWM$ равны 0.004 и 1 соответственно.

Подробная информация о паттернах и результаты вычислений даны в Приложении 2.

ОБСУЖДЕНИЕ

В данной работе рассмотрено равномерное вероятностное распределение букв в алфавите. При этой модели слова длины m равновероятны, и, соответственно, паттерны из n слов длины m равновероятны. В дальнейшем планируется исследовать зависимость количества перекрытий от параметров m и n для вероятностной модели Бернулли. При модели Бернулли паттерны из n слов длины m могут иметь неравные вероятности. Это вносит дополнительную сложность при исследовании.

Данная работа мотивирована анализом сложности алгоритма [10], решающего следующую задачу. Пусть задан паттерн. Требуется найти вероятность (P-значение) встретить паттерн в случайной последовательности заданной длины не менее определенного количества раз. Эта задача имеет применение в биоинформатике. При этом паттерн состоит из слов, характерных для фрагментов генома, обладающих определенной функцией. Например, он может описывать сайты связывания факторов регуляции транскрипции (ССТФ). P-значение используется при поиске функционально-значимых фрагментов в биологических последовательностях [7]; обзор существующих алгоритмов вычисления P-значения дан в работах [10, 11]. Алгоритм SufPref, предложенный в работе [10], использует для вычисления специальный ориентированный граф, вершинами которого являются перекрытия и слова из паттерна. Временная и пространственная сложности алгоритма SufPref линейно зависят от количества вершин в графе. Соответственно, понимание зависимости количества перекрытий от количества слов и их длин паттерна даст понимание о сложностях работы алгоритма и о его применимости к обработке большого объема данных.

ЗАКЛЮЧЕНИЕ

В работе исследуется зависимость количества перекрытий в паттернах от количества слов в паттерне и их длины. Рассматривается равномерное распределение на словах длины m. Для данной вероятностной модели доказано, что среднее количество перекрытий среди паттернов, состоящих из n слов длины m, не превосходит $c \cdot n$, где c – некоторая константа.

Важно отметить, что оценка не зависит от длины слов в паттерне. Кроме того, для получения более точной оценки константы c при равномерном распределении были проведены компьютерные эксперименты. Для четырехбуквенного алфавита и различных значений n и m было сгенерировано по 100 паттернов и было вычислено отношение ($cExp$) среднего числа перекрытий среди этих паттернов к количеству слов n. Для построенных паттернов $cExp \leq 1.06$. Согласно теоретической оценке, для рассмотренных значений n и m константа $c \leq 1.67$.

Кроме того, были рассмотрены паттерны, заданные матрицами позиционных весов из базы данных HOCOMOCO, описывающие сайты связывания факторов регуляции транскрипции. Для этих паттернов отношение количества перекрытий к количеству

слов не превосходит 1, для более половины паттернов это отношение меньше 0.1, что хорошо согласуется с полученными теоретическими и экспериментальными оценками.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (гранты № 14-14-04-32220, № 14-01-93106 и № 16-04-01640).

СПИСОК ЛИТЕРАТУРЫ

1. Boyer R.S., Moore J.S. A fast string searching algorithm. *Communications ACM*. 1977. V. 20. № 10. P. 762–772.
2. Knuth D., Morris J.H., Pratt J.V. Fast pattern matching in strings. *SIAM Journal on Computing*. 1977. V. 6. № 2. P. 323–350.
3. Aho A.V., Corasick M.J. Efficient string matching: An aid to bibliographic search. *Communications of the ACM*. 1975. V. 18. № 6. P. 333–340.
4. Crochemore M., Hancart C., Lecroq T. *Algorithms on strings*. New York: Cambridge University Press, 2007. 353 p.
5. Lothair M. *Combinatorics on Words*. Cambridge: Cambridge University Press, 1997. 260 p.
6. Duval J.P., Lecroq T., Lefebvre A. Border array on bounded alphabet. *Journal of Automata, Languages and Combinatorics*. 2005. V. 10. № 1. P. 51–60.
7. Stormo G.D. DNA binding sites: representation and discovery. *Bioinformatics*. 2000. V. 16. № 1. P. 16–23.
8. Durbin R., Eddy S., Krogh A., Mitchison G. *Biological sequences analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press, 1998.
9. Kulakovskiy I., Medvedeva Y.A, Shaefer U., Kasianov A.S., Vorontsov I.E., Bajic V.B., Makeev V.J. HOCOMOCO: A comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Research*. 2013. V. 41. P. 195–202.
10. Régnier M., Furletova E., Yakovlev V., Roytberg M. Analysis of pattern overlaps and exact computation of P-values of pattern occurrences numbers: case of Hidden Markov Models. *Algorithms for Molecular Biology*. 2014. V. 9. № 1. doi: 10.1186/s13015-014-0025-1.
11. Lothaire M. Statistics on Words with Applications to Biological Sequences. In: *Applied combinatorics on words:* Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press, 2004. P. 610.

Материал поступил в редакцию 19.11.2015, опубликован 27.01.2016.