СИСТЕМНЫЙ АНАЛИЗ

И.В. СЕРГИЕНКО, А.М. ГУПАЛ, А.В. ОСТРОВСКИЙ

УДК 519.217.2

РАСПОЗНАВАНИЕ ФРАГМЕНТОВ ГЕНОВ В ДНК С ПРИМЕНЕНИЕМ МОДЕЛЕЙ МАРКОВА СО СКРЫТЫМИ ПЕРЕМЕННЫМИ

Ключевые слова: *модель Маркова, скрытые переменные, экзоны, интроны, переходные вероятности.*

ВВЕДЕНИЕ

Раскрытие особенностей строения генома живых организмов и, прежде всего, человека в настоящее время является фундаментальной проблемой, решение которой приведет к коренным изменениям во многих областях науки и медицины. Практически все доступные методы исследования генетической структуры используют аппарат математической статистики и компьютерного обучения чаще всего в виде моделей и цепей Маркова. В частности, эти алгоритмы применяются для определения местоположения генов в ДНК, а также для определения их структуры (см., например, Augustus [1], Gene Zilla, Glimmer HMM [2], Twin Scan [3]). При этом предпочтение отдается обобщенным скрытым моделям Маркова (Generalized Hidden Markov Model — GHMM), в которых отдельным скрытым состояниям соответствуют участки генов переменной длины (например, целые интроны или экзоны). В данной работе исследуется более простой подход, при котором скрытые переменные порождают участки небольшой фиксированной длины; с помощью вычислительного эксперимента устанавливается область применимости подобного метода для разбиения генов на функциональные участки. В [4] на основе байесовского подхода и модели цепей Маркова построены простые в вычислительном плане процедуры распознавания известных фрагментов генома организма круглого червя *C. Elegans*.

Работа состоит из четырех разделов. В первом разделе приводятся общие сведения о структуре генома живых организмов, а также выводится алгоритм применительно к модели Маркова со скрытыми переменными (ММСП) для определения экзонов и интронов. Во втором разделе дается обобщение алгоритма для более сложных связей между скрытыми состояниями модели. В третьем разделе обсуждаются эвристики по увеличению скорости работы полученного алгоритма. В четвертом разделе описывается вычислительный эксперимент для проверки применимости предложенного подхода. В Заключении обсуждаются направления возможных дальнейших исследований.

1. МОДЕЛЬ МАРКОВА СО СКРЫТЫМИ ПЕРЕМЕННЫМИ

Как известно, участки ДНК, соответствующие генам эукариот (организмы, клетки которых обладают ядром, в частности растения и животные), имеют достаточно сложную структуру (рис. 1). Основные составляющие генов:

• начальный (5′-UTR) и конечный (3′-UTR) некодирующие участки, непосредственно для кодирования белков они не используются, но могут на них влиять;

• экзоны — участки ДНК, непосредственно кодирующие белки;

• интроны — участки ДНК, расположенные между экзонами, и не участвующие в синтезе белков.

Синтез белков происходит в несколько этапов:

1) ген целиком копируется из ДНК на матричную РНК;

2) в результате так называемого сплайсинга из этой РНК вырезаются интроны, а все экзоны объединяются в кодирующую последовательность (CDS);

3) из CDS с помощью универсального генетического кода, ставящего в соответствие каждой тройке нуклеотидов РНК определенную аминокислоту, формируется последовательность аминокислот белка.

Поскольку UTR-участки не при синтезе белков и не представляют интереса, во многих задачах, связанных с анализом генетической информации, они не рассматриваются. Таким образом, основная задача, которая решается в рамках данной работы с помощью ММСП, — определить разделение участка гена, заключенного между начальным и конечным некодирующими участками, заданного как последовательность нуклеотидов, на экзоны и интроны, т.е. определить для каждого нуклеотида из этого участка, относится он к экзону или интрону.

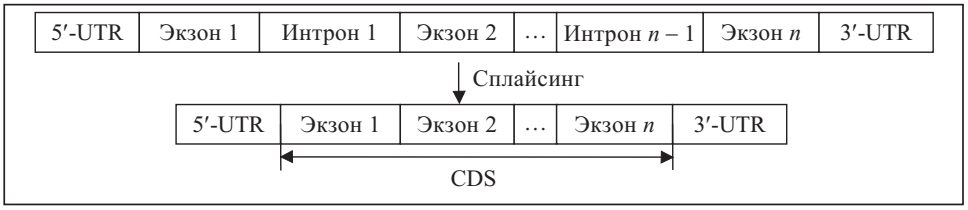

Рис. 1. Структура гена в ДНК и матричной РНК

Более формально: обозначим исследуемый участок гена $S = s_1 s_2 \ldots s_{n-1} s_n$, где $s_i \in \{a, c, g, t\}$ — нуклеотиды. Требуется поставить ему в соответствие участок равной длины $S' = s'_1 s'_2 \ldots s'_{n-1} s'_n$, где $s'_i \in \{A, C, G, T\}$, если s'_i — нуклеотид, принадлежащий экзону, и $s'_i \in \{a, c, g, t\}$ в противном случае. Разумеется, при этом должно быть выполнено требование $\forall i = 1, \ldots, n \;\; lowercase(s'_i) = s_i$, где функция *lowercase* переводит символ в нижний регистр применительно к алфавиту из нуклеотидов $lowercase\colon (A, C, G, T, a, c, g, t) \to (a, c, g, t, a, c, g, t)$.

Предположим, что последовательность нуклеотидов S порождается ММСП, в которой каждая скрытая переменная определяется наблюдаемой последовательностью нуклеотидов постоянной длины l и скрытыми состояниями, указывающими на принадлежность нуклеотидов к экзонам или интронам. Например, в случае $l = 1$ скрытая переменная содержит наблюдаемые значения $\{a, c, g, t\}$ и два скрытых состояния {экзон, интрон}, т.е. с учетом предыдущего замечания скрытая переменная принимает восемь значений: $\{A, C, G, T, a, c, g, t\}$, каждое из которых порождает один нуклеотид — название состояния, приведенное к нижнему регистру. Искомая последовательность S' может быть получена простым склеиванием последовательности скрытых состояний, которые привели к генерации S.

Аналогично будет устроена ММСП и в случае $l \geq 2$. Скрытая переменная принимает 8^l значений, соответствующих последовательности из l символов алфавита $\{A, C, G, T, a, c, g, t\}$, из них 4^l значений соответствуют наблюдаемым нуклеотидам из алфавита $\{a, c, g, t\}$, т.е. состоят из символов, приведенных к нижнему регистру (рис. 2).

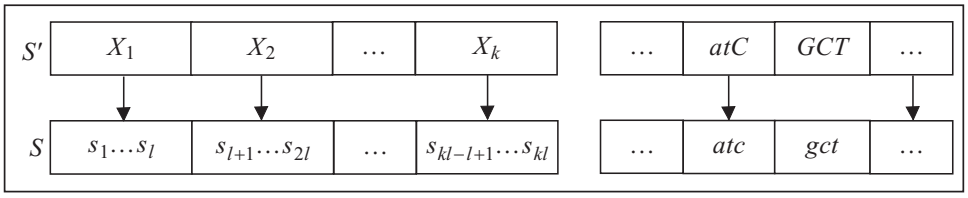

Рис. 2. Общая схема генерации последовательности нуклеотидов с помощью ММСП для $l = 3$

При фиксированных значениях наблюдаемой последовательности нуклеотидов длины l скрытая переменная имеет 2^l скрытых состояний, поэтому $8^l = 4^l \cdot 2^l$. Скрытая переменная имеет $8^l - 4^l$ скрытых значений, т.е. когда в последовательности из l символов имеется хотя бы одна буква из алфавита $\{A, C, G, T\}$.

Поскольку на длину строки S не накладывается никаких ограничений при $l \geq 2$, необходимо, чтобы $|S|$ нацело делилось на l). Однако при достаточно малых значениях l можно полагать, что последние $|S| \, b \bmod l$ символов строки S всегда относятся к последнему экзону, и в дальнейшем полагается, что задача решается для строки, длина которой кратна l: $|S| = kl$. Таким образом, последовательность скрытых переменных $X_1 X_2 \ldots X_k$ формирует последовательность S', возможно, с добавлением нескольких последних символов из S в верхнем регистре.

Итак, задача сводится к нахождению последовательности скрытых переменных $X_1 X_2 \ldots X_k$. Воспользуемся принципом максимума вероятности, т.е. определяем состояния, которые максимизируют вероятность $P(X_1 X_2 \ldots X_k)$ (либо ее логарифм). При этом полагаем, что последовательность переменных удовлетворяет условию Маркова:

$$P(X_1 X_2 \ldots X_{k-1} X_k) = P(X_1) P(X_2 | X_1) \ldots P(X_k | X_{k-1}).$$

В таких условиях задача для строки S может быть решена методом динамического программирования. Действительно, переберем все 2^l скрытых состояний X_k; для каждого из них справедливо равенство

$$\log P(X_1 \ldots X_k | X_k = x_k) = \log P(X_1 \ldots X_{k-1}) + \log P(x_k | X_{k-1}). \qquad (1)$$

Обозначим $M(i, x) = \max \log P(X_1 \ldots X_i | X_i = x)$, где максимум берется по всем возможным значениям (состояниям) $X_1, \ldots, X_{i-1}$. Тогда

$$\max \log P(X_1 \ldots X_k) = \max_{X_k = x} M(k, x). \qquad (2)$$

Из равенств (1) легко выводится рекуррентная формула для $M(i, x)$:

$$M(i, x) = \max_{y = X_{i-1}} (M(i-1, y) + \log P(x | y)), \qquad (3)$$

где максимизация проводится по всем 2^l возможным состояниям X_{i-1}. При

этом $M(i, x)$ зависит исключительно от первых $l(i-1)$ символов строки S, в частности, $M(1, x) = \log P(x)$, откуда следует, что формула (3) для произвольного i вычислима за конечное время.

По формулам (2), (3) можно легко найти не только максимум вероятности $P(X_1 X_2 \dots X_k)$, но и сами состояния X_i: для этого достаточно при каждом вычислении формулы (2) запоминать состояние y, которое максимизирует соответствующее значение M. Предложенный способ определения последовательности скрытых состояний является частным случаем алгоритма Витерби — стандартного средства решения задач подобного рода, который приведен ниже.

Алгоритм 1

Дано: последовательность S длины $|S| = kl$; вероятности для скрытых состояний вида $P(X)$ и $P(Y \mid X)$.

Найти: последовательность $S' = X_1 X_2 \dots X_k$, порождающую S с максимальной вероятностью $P(S')$.

1. **для** всех состояний x: // Инициализация M
2. **если** x порождает $s_1 \dots s_l$, **то**
3. $M(1, x) = \log P(x)$;
4. **иначе** $M(1, x) = -\infty$.
5. **для** всех $i = 2, \dots, k$:
6. **для** всех состояний x:
7. $M(i, x) := -\infty$;
8. **если** x порождает $s_{il-l+1} \dots s_{il}$, **то**
9. **для** всех состояний y:
10. **если** $M(i, x) < \log P(x \mid y) + M(i-1, y)$, **то**
11. $ptr(i, x) := y$;
12. $M(i, x) := \log P(x \mid y) + M(i-1, y)$;
13. $X_k := \arg\max_x M(k, x)$.
14. // Восстановление последовательности состояний с помощью массива *ptr*
15. **для** $i = k, k-1, \dots, 2$:
16. $X_{i-1} := ptr(i, X_i)$.

Легко видеть, что сложность алгоритма за счет трех вложенных циклов (строки 5, 6 и 9) составляет $O(k \cdot 8^{2l})$.

Отметим, что в алгоритме полагаются известными начальные вероятности $P(X)$ и переходные вероятности между скрытыми состояниями модели $P(Y \mid X)$. В то время как задача получения этих вероятностей для обобщенных скрытых моделей Маркова представляет значительную сложность, в данном случае она значительно упрощается за счет специфической структуры модели. Оценки вероятностей могут быть получены посредством сбора статистики по генам родственных организмов либо того же организма, для которых известно разбиение на экзоны и интроны. Если разметить нуклеотиды, входящие в эти гены, так же, как размечены нуклеотиды в строке S', то получим

$$P(X) \cong \frac{n_{st}(X)}{N}; \; P(Y \mid X) \cong g \frac{n(XY)}{n(X)},$$

где N — общее количество генов, $n_{st}(X)$ — число генов, начинающихся с последовательности X, $n(X)$ — число вхождений последовательности X во все гены.

2. ОБОБЩЕННАЯ МОДЕЛЬ

Обобщим рассмотренную в предыдущем разделе модель Маркова со скрытыми переменными таким образом, чтобы она могла учитывать более сложный характер зависимости между состояниями системы. В частности, рассмотрим вариант, когда вероятность перехода в скрытое состояние зависит от m предыдущих скрытых переменных, т.е. когда известны вероятности вида $P(X_i | X_{i-1}, X_{i-2}, \dots, X_{i-m})$. Если вернуться от состояний к нуклеотидам, выдвинутое предположение будет означать, что вероятность встречи последовательности нуклеотидов длины l зависит от предшествующих ей lm нуклеотидов. Считаем, что выполняется следующее предположение.

Вероятность встречи в строке S' последовательности символов длины l $s'_i \dots s'_{i+l-1}$ зависит от предыдущих m символов этой строки $s'_{i-m} \dots s'_{i-1}$ и не зависит от других параметров.

Модель, рассмотренная в предыдущем разделе, получена из данной гипотезы при $m = l$. В дальнейшем изложении полагаем, что $m \geq l$. Аналогично формуле (2) выполняется утверждение

$$\max \log P(S') = \max_x M'(n, x), \tag{4}$$

где n — длина строки S, и введено обозначение

$$M'(i, x) = \max \log P(s'_1 \dots s'_i | s'_{i-m+1} \dots s'_i = x).$$

Максимум в формуле (4) берется по всем последовательностям x длины m, которые могут породить соответствующие символы строки S: $lowercase(x) = s_{n-m+1} \dots s_n$. Для M' можно вывести рекуррентное отношение, подобное (3):

$$M'(i, x) = \max_y (M'(i-l,\ yx_1 \dots x_{m-l}) + \log P(x_{m-l+1} \dots x_m | yx_1 \dots x_{m-l})). \tag{5}$$

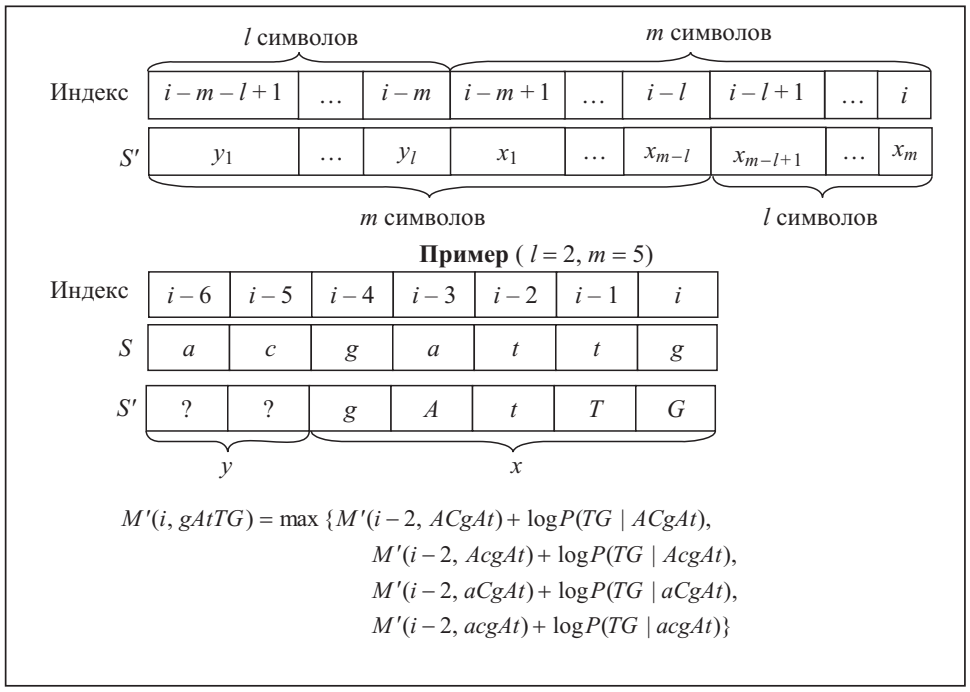

Рис. 3. Общая схема и пример вычисления функции M'

Максимум в данном случае определяется по всем 2^l последовательностям, которые могут породить символы $s_{i-m-l+1} \dots s_{i-m}$ (рис. 3). Формула (5) отличается от (3) тем, что символы из x входят не только в переходную вероятность, но и в функцию M' в правой части равенства. В качестве граничных соотношений для M' можно взять

$$M'(m,x) = \begin{cases} P(x), & \text{если } lowercase(x) = s_1 \dots s_m, \\ -\infty & \text{иначе.} \end{cases} \tag{6}$$

Алгоритм 2 (Обобщение агоритма 1)

Дано: последовательность S длины $|S| = m + kl$; вероятности для скрытых состояний вида $P(X)$ и $P(Y|X)$ при $|X| = m$, $|Y| = l$.

Найти: последовательность $S' = s'_1 \dots s'_n$, порождающую S с максимальной вероятностью $P(S')$.

1. **для** всех состояний x длины m: // Инициализация M'
2. **если** x порождает $s_1 \dots s_m$, **то**
3. $M'(m,x) = \log P(x)$;
4. **иначе** $M'(m,x) = -\infty$.
5. **для** всех $i = 1, \dots, k$:
6. **для** всех состояний x длины m:
7. $M'(m+il,x) := -\infty$;
8. **если** x порождает $s_{il+1} \dots s_{m+il}$, **то**
9. **для** всех состояний y длины l:
10. $head := x_{m-l+1} \dots x_m$; $tail := yx_1 \dots x_{m-l}$;
11. **если** $M'(m+il,x) < \log P(head|tail) + M'(m+(i-1)l, tail)$, **то**
12. $ptr(i,x) := y$;
13. $M'(m+il,x) := \log P(head|tail) + M'(m+(i-1)l, tail)$;
14. $s'_{n-m+1} \dots s'_n := \arg\max\limits_x M'(n,x)$.
15. // Восстановление последовательности состояний с помощью массива *ptr*
16. **для** $i = k, k-1, \dots, 1$:
17. $s'_{il-l+1} \dots s'_{il} := ptr(i, s'_{il+1} \dots s'_{il+m})$.

Для того чтобы формулы (6) были задействованы при вычислениях функции, необходимо, как и для ММСП, ввести ограничения на длину строки: $|S| \equiv n = m + kl$, где k — произвольное целое число (если строка не удовлетворяет этому условию, несколько ее последних символов можно отбросить, считая их принадлежащими последнему экзону).

Таким образом, сформулирован обобщенный алгоритм определения оптимальной последовательности S'. Начальные и переходные вероятности вычисляются так же, как и в предыдущем случае. Сложность алгоритма составляет $O(k \cdot 8^{l+m})$. В граничном случае ($m = l$) алгоритм 2, как легко видеть, переходит в алгоритм 1.

3. ЭВРИСТИКИ АЛГОРИТМА

Главный недостаток алгоритма 2 — работа в цикле со всеми возможными состояниями x и y, в то время как большая часть этих состояний заведомо не может порождать соответствующие символы строки S. В связи с этим естественно перейти от циклов по самим состояниям к циклам по номерам возмож-

ных состояний, упорядоченных в алфавитном порядке. Еще одним аргументом в пользу такого перехода является простота и большая скорость индексирования по целому числу по сравнению с индексированием по строке для большинства языков программирования.

Порядковый номер состояния $r(x)$ легко определить на основании последовательности его символов. Для этого достаточно заменить прописные буквы на нуль, строчные — на единицу и прочитать полученное число в двоичной системе исчисления. Например, $r(aCt) = 101_2 = 5$, т.е. aCt — пятое из восьми возможных состояний, генерирующих символы act (ему предшествуют состояния ACT, ACt, AcT, Act, aCT, после него — acT и act). Несложно реализовать и другие операции, необходимые для алгоритма 2.

• Определение состояния по его номеру и генерируемым им символам (применяется для восстановления последовательности S'). Для этого достаточно представить номер состояния в двоичной форме и перевести порожденные состоянием символы, соответствующие нулям полученного числа, в верхний регистр.

• Определение номеров состояний *head* и *tail*. Несложно убедиться, что $r(head) = r(x)\ b \bmod 2^l$, $r(tail) = [r(x)/2^l] + r(y) \cdot 2^{m-l}$.

• Переход от состояния к его номеру $q(x)$ среди всех состояний, который используется при представлении начальных и переходных вероятностей в виде массивов. Определение q существенно не отличается от определения $r(x)$, только вместо двоичной системы исчисления используется восьмеричная система с соответствием $A \leftrightarrow 0$, $C \leftrightarrow 1, \ldots, t \leftrightarrow 7$, например $q(aCt) = 417_8 = 271$.

С учетом предложенных выше эвристик сложность алгоритма 2 составляет $O(k(l+m) \cdot 2^{l+m})$, так как сложность определения $q(head)$ и $q(tail)$, вычисляемых в строке 11, составляет $O(l)$ и $O(m)$ соответственно. Хотя с точки зрения теории алгоритмов эта оценка не отличается от оценки для исходного алгоритма, на практике модифицированный вариант работает значительно быстрее.

4. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для проверки работоспособности алгоритмов 1 и 2 использовались геномы трех организмов: круглого червя *Caenorhabditis elegans*, растения резуховидки Таля (*Arabidopsis thaliana*) и плодовой мухи *Drosophila melanogaster*, которые хранятся в банке данных NCBI. Сводная информация по ним отображена в табл. 1 (под «плотностью» экзонов понимается отношение количества нуклеотидов, входящих в экзоны, к общему числу нуклеотидов в экзонах и интронах).

Таблица 1

Характеристика генома	*C. elegans*	*A. thaliana*	*D. melanogaster*
Число хромосом	6	5	6
Общая длина генома	100267632	119146348	120381566
Количество генов	23818	35176	20003
Средняя длина гена	3742	2391	11289
Средняя длина гена без UTR	3100	1954	6038
Среднее число экзонов	6,43	5,60	4,83
Средняя длина экзона	206	220	376
Средняя длина интрона	327	157	1101
«Плотность» экзонов, %	42,71	63,13	28,61

Отметим, что в геномах всех исследуемых организмов есть гены, которые могут разбиваться на функциональные части несколькими способами. Исключе-

ние сделано только для генов хромосомы 3R плодовой мухи (ввиду чрезмерно увеличивающегося размера генома за счет дублирующихся генов), для них оставлялся только один вариант разбиения на экзоны и интроны.

Для оценки качества определения экзонов и интронов использовались следующие метрики:

• специфичность и чувствительность по нуклеотидам: $NSp = TP / (TP + FP)$, $NSn = TP / (TP + FN)$, где TP — количество правильно распознанных алгоритмом нуклеотидов из экзонов, FP — количество нуклеотидов из интронов, распознанных алгоритмом как экзонные, FN — число нуклеотидов из экзонов, которые алгоритм отнес к интронам;

• коэффициент корреляции

$$CC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

(дополнительно к введенным обозначениям используется TN — число правильно распознанных алгоритмом нуклеотидов из интронов);

• средняя условная вероятность

$$ACP = \frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN} \right);$$

• специфичность и чувствительность для экзонов:

$$ESp = \frac{TE}{PE}, \; ESn = \frac{TE}{AE},$$

где TE — количество правильно выделенных алгоритмом экзонов, PE — общее число экзонов, предсказанное алгоритмом, AE — суммарное число экзонов в выборке.

Тестирование алгоритмов проводилось для значений $1 \le l \le 4$; значение m при этом варьировалось в пределах от l до $8 - l$. Поскольку количество параметров модели можно оценить количеством переходных вероятностей, т.е. 8^{l+m}, то при $l + m = 8$ число параметров ($8^8 \approx 16 \cdot 10^6$) становится сравнимым с общей длиной генома, поэтому увеличивать значения параметров выше этой границы для исследуемых организмов нецелесообразно. Для сравнения работы алгоритмов при разных комбинациях параметров использовался стандартный в таких случаях метод cross-validation. Гены каждого организма случайным образом разбивались на пять приблизительно равных частей; далее поочередно каждые четыре части использовались для обучения (в данном случае — для определения начальных и переходных вероятностей), а оставшаяся часть — для контроля. Усредненные по пяти запускам алгоритма показатели качества ESp и ESn приведены в табл. 2 (жирным шрифтом выделены наилучшие значения). Результаты относительно чувствительности для экзонов ESn практически аналогичны табл. 2. Максимальные значения метрик и значения параметров, на которых они достигаются, приведены в табл. 3 (в скобках указаны пары значений (l, m), при которых достигается оптимум).

Как видно из таблиц, при увеличении параметров l и m значения производительности алгоритмов на обучающей выборке растут. На контрольной выборке наблюдается существенное переобучение (при $l = 1$, $m = 7$ разность между метриками на обучении и контроле составляет порядка 5 %), за счет которого качество распознавания для *C. elegans* и *A. thaliana* при $m + l = 7$ выше, чем при $m + l = 8$. При этом, как и следует ожидать, качество алгоритма растет при увеличении l при фиксированном m, а также при увеличении m и постоянном l. Для моделей

с одинаковым количеством параметров (т.е. с постоянным значением $m+l$) преимуществом обладает модель с $l=1$. Полученные значения метрик свидетельствуют о достаточно хорошей применимости модели к исследуемым организмам (в [5], например, указывается порядок значений: $NSp \approx NSn \approx 90\,\%$ и $ESp \approx ESn \approx 45\,\% \div 75\,\%$). Также видно, что для более сложно устроенного генома плодовой мухи результаты оказались несколько хуже, чем для двух других видов.

Таблица 2

l	*m*	**Специфичность (в %) по экзонам для алгоритмов с различными параметрами *l* и *m***					
		C. elegans		*A. thaliana*		*D. melanogaster*	
		обучение	контроль	обучение	контроль	обучение	контроль
1	1	16,90	16,91	26,67	26,67	11,76	11,74
1	2	39,84	39,83	52,53	52,54	27,40	27,44
1	3	55,35	55,32	66,71	66,68	38,59	38,48
1	4	65,54	65,44	71,08	71,03	48,37	48,36
1	5	77,99	77,76	76,03	75,77	61,76	61,57
1	6	81,44	**80,48**	77,52	**76,35**	69,62	68,04
1	7	83,46	78,43	**80,12**	75,39	**77,11**	**70,46**
2	2	46,87	46,84	59,43	59,41	32,56	32,55
2	3	60,79	60,69	69,02	68,94	44,06	43,98
2	4	72,02	71,75	73,70	73,32	55,33	54,86
2	5	80,08	78,78	77,04	75,56	66,81	64,79
2	6	**83,70**	76,93	79,82	73,68	76,68	68,39
3	3	67,10	66,68	71,47	71,05	51,40	50,91
3	4	75,76	74,14	75,30	73,65	62,30	60,23
3	5	82,77	75,49	79,30	72,83	74,28	65,88

Таблица 3

Метрики	**Оптимальные значения (в %) метрик по всем алгоритмам**					
	C. elegans		*A. thaliana*		*D. melanogaster*	
	обучение	контроль	обучение	контроль	обучение	контроль
NSp	90,97 (2, 6)	89,45 (1, 7)	97,25 (1, 7)	96,42 (1, 6)	89,53 (2, 6)	87,92 (1, 7)
NSn	97,20 (1, 7)	96,14 (1, 7)	97,54 (1, 7)	96,90 (1, 6)	93,87 (1, 7)	92,60 (1, 7)
CC	89,28 (2, 6)	87,05 (1, 7)	92,91 (1, 7)	90,91 (1, 6)	87,98 (1, 7)	85,87 (1, 7)
ACP	94,64 (2, 6)	93,53 (1, 7)	96,46 (1, 7)	95,45 (1, 6)	93,99 (1, 7)	92,94 (1, 7)
ESp	83,70 (2, 6)	80,48 (1, 6)	80,12 (1, 7)	76,35 (1, 6)	77,11 (1, 7)	70,46 (1, 7)
ESn	85,93 (2, 6)	81,94 (1, 6)	80,93 (1, 7)	75,99 (1, 6)	77,21 (1, 7)	67,80 (1, 7)

ЗАКЛЮЧЕНИЕ

В настоящей работе рассмотрена модель распознавания функциональных участков генов, кодируемых ДНК, на основе моделей Маркова со скрытыми переменными. Приведен алгоритм определения интронов и экзонов в генах и эвристика, позволяющая значительно ускорить его работу. На основе вычислительного эксперимента сделан вывод о возможности применения алгоритма

для генома организмов, имеющих не слишком сложную структуру. В дальнейшем можно проверить применимость работы алгоритма для других сложных организмов, в частности генома человека.

Заметим, что скрытые параметры фигурируют в известной задаче распознавания вторичной структуры белков. Имеется первичная последовательность аминокислот белка, необходимо определить ее вторичную структуру: поставить в соответствие каждой аминокислоте один из двух возможных типов регулярной структуры (α-спираль, β-слой) или ее отсутствие, т.е. нерегулярность (coil). Но даже беглый сравнительный анализ показывает, что эта задача сложнее задачи распознавания функциональных участков генов, поскольку средняя длина участков, состоящих из α-спиралей и β-слоев, гораздо короче средней длины экзонов и интронов в рассмотренных геномах.

Для распознавания вторичной структуры белков в [6, 7] использовались байесовские процедуры на моделях нестационарных цепей Маркова различных порядков. Проведенные численные расчеты на основе информации из базы данных NCBI подтвердили высокую эффективность этих методов, на множестве из 20 тысяч белков средний процент распознавания превысил 80 %. Поэтому целесообразно провести сравнительный анализ применения этих двух подходов как для распознавания фрагментов генов, так и вторичной структуры белков.

СПИСОК ЛИТЕРАТУРЫ

1. Stanke M., Waack S. Gene prediction with a hidden Markov model and a new intron submodel // Bioinformatics. — 2003. — N 19. — P. 215–225.
2. Majoros W.H., Pertea M., Salzberg S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders // Ibid. — 2004. — N 20. — P. 2878–2879.
3. Korf I., Flicek P., Duan D., Brent M.R. Integrating genomic homology into gene structure prediction // Ibid. — 2001. — N 17. — P. 140–148.
4. Андрейчук И.И., Гупал А.М., Рязанов В.В. Байесовская процедура распознавания фрагментов генов в ДНК // Международный научно-технический журнал «Проблемы управления и информатики». — 2011. — № 6. — С. 120–124.
5. Knapp K., Chen Y.-P.P. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy // Nucleic Acids Research. — 2007. — N 35 (1). — P. 317–324.
6. Предсказание вторичной структуры белков на основе байесовских процедур распознавания на цепах Маркова / И.В. Сергиенко, Б.А. Белецкий, С.В. Васильев, А.М. Гупал // Кибернетика и системный анализ. — 2007. — № 2. — С. 59–64.
7. Гупал А.М., Сергиенко И.В. Оптимальные процедуры распознавания. — Киев: Наук. думка, 2008. — 232 с.

Поступила 15.12.2011