УДК 519.217.2

**А.М. ГУПАЛ, А.В. ОСТРОВСКИЙ**

# ИСПОЛЬЗОВАНИЕ КОМПОЗИЦИЙ МОДЕЛЕЙ МАРКОВА ДЛЯ ОПРЕДЕЛЕНИЯ ФУНКЦИОНАЛЬНЫХ УЧАСТКОВ ГЕНОВ

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

## ВВЕДЕНИЕ

Определение внутренней структуры последовательностей нуклеотидов из ДНК, составляющих геном человека и других биологических организмов, представляет значительный теоретический и прикладной интерес для многих областей науки. Одной из весомых подзадач при определении структуры генома является распознавание экзонов (участков ДНК, которые кодируют белки) и интронов (некодирующих участков, расположенных между экзонами). В настоящее время наиболее распространенный способ решения этой задачи — использование обобщенных моделей Маркова со скрытыми параметрами (generalized hidden Markov models) [1, 2].

Более простой подход, использующий модели на основе обычных цепей Маркова со скрытыми переменными, рассмотрен в [3]. Такой подход позволяет с теоретической точки зрения обосновать некоторые эмпирические закономерности, которые в других алгоритмах принимаются a priori (например, специфический вид переходных участков между экзонами и интронами). Как показано в [3], полученную модель можно использовать для эффективного распознавания участков генов в организмах с относительно просто устроенным геномом, например, в большинстве растений или насекомых. В то же время при попытке применить модель для более развитых видов (млекопитающих или птиц) качество определения экзонов и интронов снижается по сравнению с известными алгоритмами. Для повышения качества классификации предлагается использовать композиции алгоритмов на основе моделей Маркова.

В первом разделе настоящей статьи сформулирована задача распознавания функциональных участков генов и описана в общих чертах модель на основе композиции алгоритмов, предлагаемая для ее решения. Во втором разделе выведен критерий оптимальности композиции, который в следующем разделе связан с понятием информационной энтропии. Полученные теоретические результаты подытожены в четвертом разделе, где приведен алгоритм построения оптимальной композиции. В пятом разделе статьи описан вычислительный эксперимент, цель которого — выяснить эффективность нового подхода, и проанализированы его результаты. В заключении приведены направления для возможных дальнейших исследований.

## 1. ПОСТАНОВКА ЗАДАЧИ

Как известно из биохимии, белки всех живых организмов кодируются генами — последовательностями нуклеотидов: аденина (A), цитозина (C), гуанина (G) и тимина (T). Участок гена между началом первого и концом последнего экзона, представляющий наибольший интерес, состоит из чередующихся между собой экзонов и интронов. Таким образом, задача сводится к определе-

нию для каждого нуклеотида его принадлежности к экзону или интрону. Более формально, требуется по известной последовательности $S \in O^*$, где $O = \{\mathrm{A, C, G, T}\}$ — множество наблюдаемых состояний, восстановить последовательность скрытых состояний $S' \in Q^* \equiv \{\mathrm{A, C, G, T}, a, c, g, t\}^*$, где прописными буквами обозначены нуклеотиды, принадлежащие экзонам, а строчными — интроны. При этом скрытые состояния должны соответствовать наблюдаемым, т.е. для алгоритма распознавания $\mathsf{A}: O^* \to Q^*$ необходимо выполнение условия $\forall S \in O^*\ \Pr(\mathsf{A}(S)) = S$, где $\Pr: Q^* \to O^*$ — проекция скрытых состояний $\Pr: \langle \mathrm{A, C, G, T}, a, c, g, t\rangle \to \langle \mathrm{A, C, G, T, A, C, G, T}\rangle$. Кроме того, алгоритм должен иметь свойство оптимальности, т.е. максимизировать условную вероятность цепочки скрытых состояний

$$\mathsf{A}(S) = \arg\max_{S'} P(S' \mid S) = \arg\max_{S'} (P(S')[\Pr(S') = S]). \tag{1}$$

В работе [3] рассмотрены модели на основе цепей Маркова произвольного $m$-го порядка, для которых вероятность $P(S')$ вычисляется следующим образом:

$$P(S') = \pi(s'_1 \ldots s'_m)\, p(s'_{m+1} \mid s'_1 \ldots s'_m) \times \ \ldots \times p(s'_n \mid s'_{n-m} \ldots s'_{n-1}), \tag{2}$$

где $\pi(x)$ обозначена вероятность появления подстроки $x$ в начале строки $S'$, а $p(y \mid x)$ — вероятность появления в ней состояния $y$ при известной последовательности предыдущих состояний $x$. Оценки начальных и переходных вероятностей модели взяты из обучающей выборки $T = \{S'_i\} \subset Q^*$, содержащей гены с известным разбиением на экзоны и интроны

$$\hat{\pi}(x) = N_{st}(T, x) / |T|, \ \ \hat{p}(x \mid y) = N(T, yx) / N(T, y). \tag{3}$$

Здесь $N_{st}(T, x)$ — количество последовательностей из выборки, которые начинаются $x$, а $N(T, x)$ — число вхождений строки $x$ во все последовательности $T$. Алгоритм вида (1), который использует для подсчета вероятностей выражения (2), (3) и получен при обучении на выборке $T$, далее будем обозначать $\mathsf{A}[T]$.

Исследуем композицию алгоритмов

$$\mathsf{A}(S) = \begin{cases} \mathsf{A}_1(S), & S \in G_1, \\ \mathsf{A}_2(S), & S \in G_2, \\ \vdots & \\ \mathsf{A}_l(S), & S \in G_l, \end{cases} \tag{4}$$

где области $G_1, G_2, \ldots, G_l$ образуют покрытие множества $O^*$, т.е. $\bigcup\limits_i G_i = O^*$, $G_i \cap G_j = \{\}$ при $i \neq j$.

Фактически подобные композиции являются частным случаем смесей алгоритмов $\mathsf{A}(S) = C\left(\sum\limits_i g_i(S)\, \mathsf{A}_i(S)\right)$, где $g_i: O^* \to [0, 1]$ — весовые функции, $C$ — решающее правило. В данном случае каждый составляющий алгоритм эксклюзивно компетентен в соответствующей ему области $g_i(S) = [S \in G_i]$. Использование таких весовых функций позволяет не определять пространство для проведения промежуточных алгебраических операций сложения и умножения, а также делает возможным использование тривиального решающего правила $C(S) \equiv S$.

Построение композиции вида (4), в которой базовые алгоритмы используют цепи Маркова, на основе обучающей выборки $T$, как и для смесей алгоритмов в общем случае, включает следующие шаги.

**Шаг 1.** Определение каким-либо способом областей $G_1, G_2, \ldots, G_l$.

**Шаг 2.** Обучение составляющих композицию алгоритмов на соответствующих им частях выборки $\mathsf{A}_k = \mathsf{A}[T_k]$, где $T_k = \{S' \in T \mid \Pr(S') \in G_i\}$.

Второй шаг рассмотрен в [3]. Таким образом, основной интерес представляет нахождение оптимального покрытия множества $O^*$, для чего предлагается использовать вероятностный подход.

## 2. КРИТЕРИЙ ОПТИМАЛЬНОГО РАЗБИЕНИЯ

Обозначим $P(T_k \mid \mathsf{A}[T_k])$, $k = 1, \ldots, l$, совместную вероятность порождения строк из множества $T_k$ моделью, обученной на этой части выборки. Согласно (2), (3)

$$P(T_k \mid \mathsf{A}[T_k]) = \prod_{S' \in T_k} \left[ \hat{\pi}(s'_1 \ldots s'_m) \prod_{i=m+1}^{|S'|} \hat{p}(s'_i \mid s'_{i-m} \ldots s'_{i-1}) \right]$$

или после перехода к логарифмам

$$\log P(T_k \mid \mathsf{A}[T_k]) = \sum_{S' \in T_k} \log \frac{N_{st}(T_k, s'_1 \ldots s'_m)}{|T_k|} + \sum_{S' \in T_k} \sum_{i=m+1}^{|S'|} \log \frac{N(T_k, s'_{i-m} \ldots s'_i)}{N(T_k, s'_{i-m} \ldots s'_{i-1})} =$$

$$= \sum_{S' \in T_k} \log N_{st}(T_k, s'_1 \ldots s'_m) - |T_k| \log |T_k| +$$

$$+ \sum_{S' \in T_k} \sum_{i=m+1}^{|S'|} \log N(T_k, s'_{i-m} \ldots s'_i) - \sum_{S' \in T_k} \sum_{i=m+1}^{|S'|} \log N(T_k, s'_{i-m} \ldots s'_{i-1}).$$

Суммирование во всех получившихся суммах выполняется по коротким фрагментам строк, входящих во множество $T_k$: префиксам длины $m$ в первой сумме, подстрокам длины $m$ и $m+1$ — во второй и третьей соответственно. Таким образом, выражение для вероятности можно упростить, если перейти к непосредственному суммированию по строкам фиксированной длины

$$\log P(T_k \mid \mathsf{A}[T_k]) = \sum_{|y|=m} N_{st}(T_k, y) \log N_{st}(T_k, y) - |T_k| \log |T_k| +$$

$$+ \sum_{|x|=m+1} N(T_k, x) \log N(T_k, x) - \sum_{|y|=m} N(T_k, y) \log N(T_k, y). \qquad (5)$$

При этом подразумевается, что при $x = 0$ выполняется тождество $x \log x = 0$.

Логарифм вероятности генерации обучающей выборки набором моделей, использующихся в алгоритмах $\mathsf{A}[T_1], \ldots, \mathsf{A}[T_l]$, в силу непересекаемости множеств $T_1, \ldots, T_l$ равен

$$\log P(T \mid \mathsf{A}[T_1], \ldots, \mathsf{A}[T_l]) = \sum_{k=1}^{l} \log P(T_k \mid \mathsf{A}[T_k]). \qquad (6)$$

Непосредственная максимизация (6) по возможным разбиениям $G_1, \ldots, G_l$ и соответствующим им $T_1, \ldots, T_l$ представляет значительную сложность. Для облегчения этой задачи построим разбиение в виде дерева, т.е. вначале найдем оптимальное разбиение множества наблюдаемых строк $O^*$ на два подмножества, затем одного из полученных множеств — еще на два подмножества, и т.д. В качестве меры разделительной способности разбиения множества $T_k$ на части $T_k^+$ и $T_k^-$ по аналогии с (6) используем функцию

$$\Delta(T_k, T_k^+, T_k^-) = \log P(T_k^+ \mid \mathsf{A}[T_k^+]) + \log P(T_k^- \mid \mathsf{A}[T_k^-]) - \log P(T_k \mid \mathsf{A}[T_k]). \quad (7)$$

Несложно заметить, что, складывая выражения (7) для разделяемого множества на каждом из $l-1$ этапов описанного выше алгоритма, получим (6) с точностью до не зависящего от полученного разбиения слагаемого $\log P(T \mid \mathsf{A}[T])$. Таким образом, алгоритм построения дерева является поэтапным жадным способом максимизации вероятности (6).

### 3. СВЯЗЬ С ИНФОРМАЦИОННОЙ ЭНТРОПИЕЙ

Выражение (5) содержит члены, напоминающие информационную энтропию для эмпирических вероятностных распределений $\hat{\pi}$ и $\hat{p}$, полученных на основании выборки $T_k$. Действительно, энтропия для начального распределения имеет вид

$$H(\hat{\pi}) = -\sum_{|y|=m} \hat{\pi}(y) \log \hat{\pi}(y) = -\frac{1}{|T_k|} \sum_{|y|=m} N_{st}(T_k, y)(\log N_{st}(T_k, y) - \log |T_k|) =$$

$$= \log |T_k| - \frac{1}{|T_k|} \sum_{|y|=m} N_{st}(T_k, y) \log N_{st}(T_k, y). \qquad (8)$$

Энтропию для условного распределения $\hat{p}$ можно найти следующим образом:

$$H(\hat{p}) = H(\hat{x} \mid \hat{y}) = H(\hat{x}, \hat{y}) - H(\hat{y}) = H(\hat{x}) - H(\hat{y}), \qquad (9)$$

$$H(\hat{y}) = -\sum_{|y|=m} \frac{N(T_k, y)}{N_m(T_k)} \log \frac{N(T_k, y)}{N_m(T_k)} =$$

$$= \log N_m(T_k) - \frac{1}{N_m(T_k)} \sum_{|y|=m} N(T_k, y) \log N(T_k, y), \qquad (10)$$

где $\hat{x}$ — эмпирическое распределение для последовательностей из $m+1$ скрытых состояний; $\hat{y}$ — аналогичное распределение для последовательностей длины $m$; $N_m(T_k)$ — общее число последовательностей длины $m$ в выборке. Подставив (10) и аналогичную ей формулу для $H(\hat{x})$ в (9) и воспользовавшись тем, что

$$N_m(T_k) \approx N_{m+1}(T_k) \approx N_1(T_k) \equiv \sum_{S' \in T_k} |S'|,$$

имеем

$$H(\hat{p}) \approx -\frac{1}{N_1(T_k)} \times$$

$$\times \left( \sum_{|x|=m+1} N(T_k, x) \log N(T_k, x) - \sum_{|y|=m} N(T_k, y) \log N(T_k, y) \right). \qquad (11)$$

Сравнив (8) и (11) с (5), получим

$$\log P(T_k \mid \mathsf{A}[T_k]) \approx -H(\hat{\pi}) |T_k| - H(\hat{p}) \sum_{S' \in T_k} |S'|.$$

Таким образом, нахождение максимума выражения (7) приблизительно соответствует минимизации энтропии эмпирически получаемых распределений начальных и переходных вероятностей для частей выборки по сравнению с распределениями в целом. Это с теоретической точки зрения является дополнительным обоснованием выбора меры разделительной способности (7).

### 4. НАХОЖДЕНИЕ ОПТИМАЛЬНЫХ РАЗБИЕНИЙ

Рассмотрим бинарные разбиения из (7) на основе предикатов $I : O^* \to \{0, 1\}$, в которых используются концентрации определенных нуклеотидов в последовательности скрытых состояний

$$I_{X,\theta}(S) = \left[ \sum_{x \in X} n(S,x) < \theta \right],\ X \subset O,\ \theta \in (0,1),\ n(S,x) \equiv \frac{N(S,x)}{|S|}. \qquad (12)$$

Преимущество подобных предикатов заключается в легкости их вычисления и простоте интерпретации. Для такого предиката мера качества разделения (7) примет вид

$$\Delta(T_k, I) \equiv \Delta(T_k, \{S' \in T_k | I(S')\}, \{S' \in T_k | \neg I(S')\}).$$

Из всех $2^4 = 16$ возможных подмножеств множества нуклеотидов имеет смысл рассматривать семь: $X \in \mathbf{X} = \{\{A\}, \{C\}, \{G\}, \{T\}, \{A, C\}, \{A, G\}, \{A, T\}\}$. В самом деле, подмножества {} и {A, C, G, T} дают тривиальный результат. В силу тождества $\forall S \in O^* \sum_{x \in O} n(S,x) = 1$ имеет место $I_{X,\theta} \sim I_{O \setminus X, 1-\theta}$, следовательно, не имеет смысла использовать и остальные семь подмножеств, дополнения которых находятся в $\mathbf{X}$.

Предикат, разбивающий набор строк $T_s$ на две части оптимальным образом, будем искать среди предикатов всех семи типов (по $q$ кандидатов каждого типа), генерируемых приведенным далее алгоритмом.

**Алгоритм 1.** $\mathbf{I} = \text{Predicates}(T_s, q)$.

**Дано:** Набор последовательностей скрытых состояний $T_s$, число предикатов каждого типа $q$.

**Найти:** Набор потенциально оптимальных предикатов $\mathbf{I} = \{I_i\}_{i=1}^{7q}$.

1. $\mathbf{I} = \{\}$;
2. **для всех** наборов нуклеотидов $X \in \mathbf{X}$:
3. $n_X := \left\{ \sum_{x \in X} n(\Pr(S'), x) \right\}$, $S' \in T_s$ — статистика по концентрации нуклеотидов;
4. Упорядочить элементы $n_X$ по возрастанию;
5. **для** $i = 1, \ldots, q$:
6. $\theta := n_X[i \cdot |T_s| / (q+1)]$ — предикат $I_{X,\theta}$ выполняется для $i/(q+1)$-й части выборки;
7. $\mathbf{I} := \mathbf{I} \cup \{I_{X,\theta}\}$.

С учетом изложенного выше схема поиска оптимального разбиения, описанная в разд. 2 статьи, примет следующий вид.

**Алгоритм 2.**

**Дано:** Обучающая выборка $T$; количество областей $l$; $q$.

**Найти:** Разбиение $G_1, \ldots, G_l$, алгоритмы $\mathsf{A}_1, \ldots, \mathsf{A}_l$.

1. $\mathbf{T} := \{T\}$ — текущее разбиение обучающей выборки;
2. $\mathbf{G} := \{O^*\}$ — разбиение пространства наблюдаемых строк;
3. **для** $i = 1, \ldots, l-1$:
4. **для** $k = 1, \ldots, |\mathbf{T}|$:
5. $\mathbf{I} := \text{Predicates}(T_k, q)$;
6. $I_k^{\max} := \arg\max_{I \in \mathbf{I}} \Delta(T_k, I)$ — оптимальный предикат для текущей части;
7. $r := \arg\max_{k=1,\ldots,|\mathbf{T}|} \Delta(T_k, I_k^{\max})$ — часть выборки, которую надо разделить;
8. Выделить новую часть обучающей выборки:

$$T_{i+1} := \{S' \in T_r | I_r^{\max}(S')\};\ T_r := \{S' \in T_r | \neg I_r^{\max}(S')\};$$

9. Аналогично для пространства $O^*$:

$$G_{i+1} := \{S' \in G_r \mid I_r^{\max}(S')\}; \; G_r := \{S' \in G_r \mid \neg I_r^{\max}(S')\};$$

10. **для** $i = 1, \ldots, l$:

11. $\mathsf{A}_i := \mathsf{A}[T_i]$

За счет выполнения вычислений меры $\Delta(T_k, I)$ на шестом шаге алгоритма в несколько параллельных потоков можно добиться существенного ускорения работы на многоядерных системах.

## 5. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для оценки эффективности использования комбинаций алгоритмов из банка данных NCBI взяли геномы шести биологических видов: Homo sapiens (человека), Gallus gallus (курицы), Suc scrofa (свиньи), Rattus norvegicus (крысы), Mus musculus (мыши), Papio anubis (павиана). При этом рассматривались только гены с полностью известной нуклеотидной записью, для генов человека было введено верхнее ограничение на длину $|S'| \le 40000$. Для оценки качества использовалась пятикратная кросс-валидация: выборка случайным образом разбивалась на пять приблизительно равных частей, каждая из которых по очереди использовалась в качестве контроля, а остальные четыре части — в качестве обучающей выборки. На обучающей и контрольной частях выборки замерялись меры качества [4]: четыре, связанные с отдельными нуклеотидами, — специфичность *NSp*, чувствительность *NSn*, коэффициент корреляции *CC* и средняя условная вероятность *ACP*, а также две, показывающие качество распознавания границ между экзонами и интронами, — экзонные специфичность *ESp* и чувствительность *ESn*.

В ходе эксперимента выяснилось, что для разумных значений порядка цепей Маркова $m \in \{5, 6, 7\}$ деревья разбиения, построенные на обучающей выборке, практически не отличаются от дерева для всего набора генов, а также один от другого при разных значениях *m.* Эти свойства объясняются относительно простым видом предикатов (12). В связи с этим для всех тестов использовались разбиения на основе полных геномов, полученные для порядка цепей Маркова $m = 6$. Что касается других параметров алгоритма, то число предикатов каждого типа в алгоритме 1 принималось равным $q = 10$, а количество алгоритмов в композиции $l$ варьировалось от одного до шести.

Схемы деревьев для геномов человека и курицы, полученные в результате работы алгоритма 2 с описанными выше параметрами, приведены на рис. 1, 2, из которых видно, что большая часть предикатов в оптимальной композиции алгоритмов основана на суммарной концентрации нуклеотидов A и T (или, что тоже самое, C и G). Значимость этой концентрации отмечена в [5, 6].

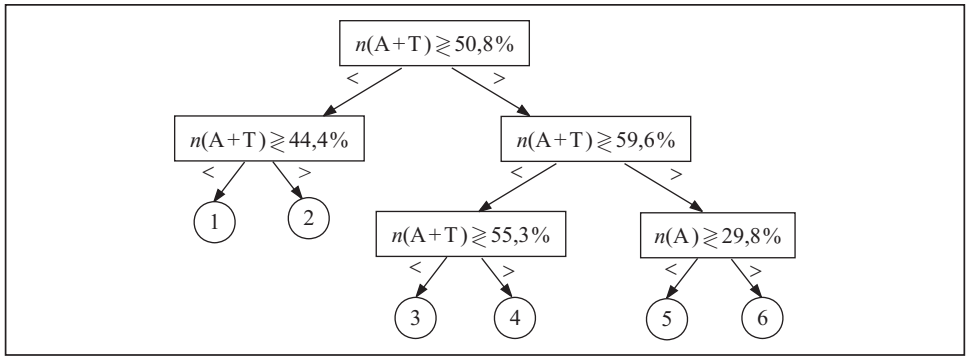


*Рис. 1.* Дерево, построенное для генов человека при $m = 6$

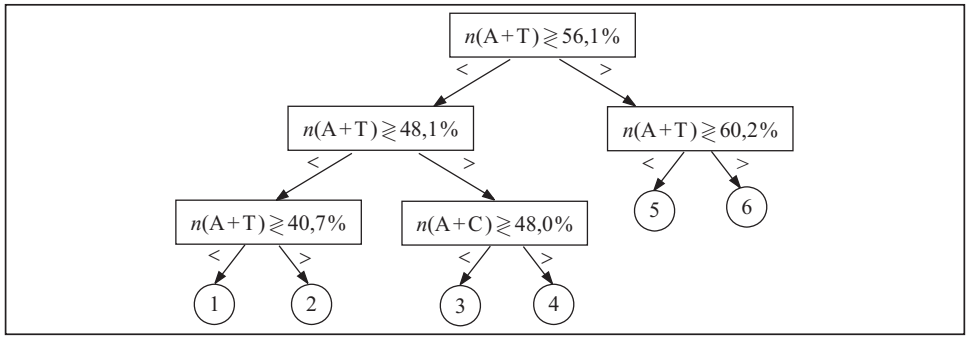


Рис. 2. Дерево, построенное для генов курицы при $m = 6$

**Таблица 1**

| Мера | Выборка | Результаты классификации для генома человека при $m = 6$ с числом алгоритмов | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| *NSp* | обучающая | 34,39 | 37,51 | 39,14 | 39,02 | 40,57 | 41,70 |
| | контрольная | 34,16 | 37,02 | 38,39 | 38,07 | 39,38 | 40,24 |
| *NSn* | обучающая | 88,72 | 91,70 | 92,24 | 92,32 | 92,55 | 92,77 |
| | контрольная | 88,22 | 90,78 | 90,81 | 90,49 | 90,43 | 90,07 |
| *CC* | обучающая | 47,96 | 52,18 | 53,95 | 53,87 | 55,42 | 56,53 |
| | контрольная | 47,55 | 51,36 | 52,70 | 52,27 | 53,49 | 54,15 |
| *ACP* | обучающая | 75,71 | 77,80 | 78,62 | 78,59 | 79,29 | 79,79 |
| | контрольная | 75,49 | 77,37 | 77,97 | 77,76 | 78,29 | 78,56 |
| *ESp* | обучающая | 25,70 | 31,59 | 33,09 | 34,45 | 35,64 | 36,79 |
| | контрольная | 24,91 | 29,92 | 30,58 | 31,26 | 31,92 | 32,35 |
| *ESn* | обучающая | 27,93 | 39,93 | 41,78 | 44,33 | 45,94 | 47,12 |
| | контрольная | 26,93 | 37,49 | 38,14 | 39,56 | 40,37 | 40,49 |

**Таблица 2**

| Мера | Выборка | Результаты классификации для генома человека при $m = 7$ с числом алгоритмов | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| *NSp* | обучающая | 36,46 | 43,24 | 48,76 | 50,75 | 57,93 | 60,64 |
| | контрольная | 35,58 | 41,33 | 45,75 | 46,72 | 52,63 | 54,43 |
| *NSn* | обучающая | 91,32 | 94,52 | 95,26 | 95,48 | 95,75 | 96,03 |
| | контрольная | 89,56 | 91,42 | 90,57 | 89,02 | 87,62 | 87,53 |
| *CC* | обучающая | 51,01 | 58,62 | 63,68 | 65,40 | 71,10 | 73,23 |
| | контрольная | 49,48 | 55,67 | 59,17 | 59,38 | 63,49 | 64,71 |
| *ACP* | обучающая | 77,25 | 80,83 | 83,09 | 83,85 | 86,37 | 87,32 |
| | контрольная | 76,45 | 79,32 | 80,78 | 80,77 | 82,49 | 83,02 |
| *ESp* | обучающая | 30,68 | 40,57 | 45,21 | 48,94 | 53,95 | 56,36 |
| | контрольная | 27,64 | 34,81 | 37,29 | 38,87 | 42,27 | 43,84 |
| *ESn* | обучающая | 35,23 | 52,16 | 56,78 | 61,82 | 66,43 | 67,99 |
| | контрольная | 31,11 | 43,46 | 44,93 | 46,64 | 48,83 | 49,71 |

В табл. 1, 2 приведены результаты качества распознавания экзонов и интронов в геноме человека в зависимости от количества алгоритмов в композиции при использовании цепей Маркова шестого и седьмого порядка соответственно. Как видно, использование композиций позволяет существенно повысить все меры качества распознавания, прежде всего нуклеотидную специфичность *NSp* и экзонные меры *ESp* и *ESn*. Полученные данные также свидетельствуют, что при превышении опре-

деленного предела количества алгоритмов в композиции она становится менее продуктивной из-за усиления эффекта переобучения — базовым алгоритмам композиции начинает соответствовать слишком мало генов из обучающей выборки. В табл. 3 представлены результаты применения классификации базового алгоритма и композиции на контрольной выборке при использовании цепей Маркова седьмого порядка. Сравнение данных позволяет сделать вывод о том, что применение композиций дает возможность повысить меры качества распознавания на 10–15 %.

**Таблица 3**

| Вид | Число алгоритмов | Результаты использования алгоритмов с мерами качества | | | | | |
|---|---|---|---|---|---|---|---|
| | | *NSp* | *NSn* | *CC* | *ACP* | *ESp* | *ESn* |
| Homo sapiens (человек) | 1 | 35,58 | 89,56 | 49,48 | 76,45 | 27,64 | 31,11 |
| | 6 | 54,43 | 87,53 | 64,71 | 83,02 | 43,84 | 49,71 |
| Gallus gallus (курица) | 1 | 54,41 | 64,65 | 56,06 | 78,12 | 47,39 | 32,03 |
| | 3 | 68,29 | 64,14 | 63,81 | 81,92 | 52,90 | 37,70 |
| Sus scrofa (свинья) | 1 | 33,24 | 85,66 | 47,60 | 75,75 | 24,87 | 26,64 |
| | 4 | 47,54 | 81,02 | 58,04 | 79,83 | 36,21 | 40,86 |
| Mus musculus (мышь) | 1 | 59,97 | 85,32 | 67,20 | 83,97 | 42,22 | 40,39 |
| | 4 | 71,96 | 83,03 | 74,08 | 87,11 | 50,90 | 47,73 |
| Rattus norvegicus (крыса) | 1 | 61,73 | 83,59 | 67,47 | 84,01 | 40,75 | 36,49 |
| | 6 | 76,23 | 75,32 | 72,22 | 86,11 | 47,34 | 38,78 |
| Papio anubis (павиан) | 1 | 39,91 | 86,65 | 52,28 | 77,45 | 30,41 | 31,42 |
| | 6 | 65,00 | 79,75 | 68,56 | 84,42 | 50,54 | 51,18 |

## ЗАКЛЮЧЕНИЕ

Рассмотрена задача определения фрагментов генов высших организмов (млекопитающих и птиц). Предложен метод решения этой задачи на основе определенного вида композиций алгоритмов, использующих модели Маркова со скрытыми состояниями. Исследованные композиции позволяют существенно повысить качество классификации фрагментов генов по сравнению с отдельными алгоритмами и достичь уровня известных алгоритмов на основе обобщенных моделей Маркова, решающих поставленную задачу. Эффективность композиций, таким образом, косвенно свидетельствует о применимости математического аппарата, введенного в [3], для определения экзонов и интронов в генах.

В качестве направлений для дальнейших исследований выделим изучение более общего вида предикатов, используемых для создания оптимального разбиения в алгоритме 1, а также обобщение результатов для применения в схожих задачах, например, для определения вторичной структуры белков [7].

СПИСОК ЛИТЕРАТУРЫ


1. Stanke M., Waack S. Gene prediction with a hidden Markov model and a new intron submodel // Bioinformatics. — 2003. — **19**, Suppl. 2. — 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. — **20**, N 16. — P. 2878–2879.
3. Сергиенко И.В., Гупал А.М., Островский А.В. Распознавание фрагментов генов в ДНК с применением моделей Маркова со скрытыми переменными // Кибернетика и системный анализ. — 2012. — № 3. — С. 58–67.
4. Knapp K., Chen Y.-P.P. An evaluation of contemporary hidden Markov model genefinders with a predicted exon taxonomy // Nucleic Acids Research. — 2007. — **35**. — P. 317–324.
5. Sumner A.T., de la Torre J., Stuppia L. The distribution of genes on chromosomes: a cytological approach // J. Mol. Evol. — 1993. — **37**, N 2. — P. 117–122.
6. Aïssani B., Bernardi G. CpG islands, genes and isochores in the genomes of vertebrates // Gene. — 1991. — **106**, N 2. — P. 185–195.
7. Предсказание вторичной структуры белков на основе байесовских процедур распознавания на цепях Маркова / И.В. Сергиенко, Б.А. Белецкий, С.В. Васильев, А.М. Гупал // Кибернетика и системный анализ. — 2007. — № 2. — С. 59–64.


*Поступила 23.01.2012*