УДК 519.217.2

А.В. ОСТРОВСКИЙ

ПРИМЕНЕНИЕ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ ДЛЯ ВЫБОРА ОПТИМАЛЬНЫХ МЕТАПАРАМЕТРОВ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ФРАГМЕНТОВ ГЕНОВ

Аннотация. Рассмотрена задача выбора оптимального порядка скрытой марковской модели для распознавания функциональных фрагментов генов. Предложены четыре статистических критерия определения оптимального порядка на основе отношения правдоподобия, эргодического свойства, марковского свойства и информационного критерия Акаике. Подтверждена эффективность использования для решения рассматриваемой задачи байесовских смесей марковских моделей; с помощью статистических критериев определено оптимальное количество компонент смеси.

Ключевые слова: модель Маркова, распознавание, скрытое состояние, нуклеотид, экзон, интрон, правдоподобие.

ВВЕДЕНИЕ

Задача распознавания функциональных участков генов является одной из основных направлений исследований в биоинформатике. В [1, 2] для ее решения предложена вероятностная модель на основе скрытых моделей Маркова высокого порядка. Качество распознавания алгоритма максимизации правдоподобия на основе этой модели для геномов растений и насекомых сравнимо с актуальными методами на основе более сложных обобщенных скрытых марковских моделей. Для повышения качества распознавания на геномах млекопитающих, обладающих более сложной структурой, предложены линейные (байесовские) смеси моделей, строящиеся с использованием EM-алгоритма [3]. В описанных выше способах построения вероятностной модели возникает проблема выбора оптимального значения ее метапараметров: порядка марковской модели и количества алгоритмов при использовании композиций. В [1–3] этот выбор основан на пятикратной кросс-валидации рассматриваемой выборки. В настоящей статье описывается альтернативный метод, основанный на теории проверки статистических гипотез. В ходе вычислительного эксперимента выявлено, что выводы, сделанные с помощью этой теории, в целом соответствуют полученным ранее результатам.

В статье используются следующие обозначения. Строки помечаются строчными латинскими буквами с чертой сверху, например $\bar{s}$; i-й символ строки $\bar{s}$ обозначается s_i, а последовательность символов строки с i-го по j-й ($j \geq i$) — как s_i^j. Длина строки $\bar{s}$ обозначается $|\bar{s}|$. Через S^*, где S — произвольное конечное множество, обозначено множество конечных строк, составленных из S:

$$S^* \equiv \bigcup_{i=0}^{\infty} S^i.$$

Случайная величина с распределением хи-квадрат и числом степеней свободы df обозначена $\chi^2(df)$. Остальные обозначения вводятся в процессе изложения.

1. ОПИСАНИЕ ВЕРОЯТНОСТНЫХ МОДЕЛЕЙ

Задача распознавания фрагментов генов приводится к следующему общему виду [2]: построить алгоритм распознавания $A: S^* \to H^*$, сопоставляющий каждой строке наблюдаемых состояний последовательность скрытых состояний той же длины,

на основе конечного множества прецедентов (обучающей выборки)

$$X = \{(\bar{s}_i, \bar{h}_i)\}_{i=1}^{n},\ \bar{s}_i \in S^*,\ \bar{h}_i \in H^*.$$

Здесь S — множество наблюдаемых состояний, H — множество скрытых состояний.

Для генов наблюдаемые состояния имеют смысл составляющих элементов генов (нуклеотидов) — аденина, цитозина, гуанина и тимина: $S = \{A, C, G, T\}$. Множество H соответствует двум основным функциональным фрагментам генов — экзонам и интронам: $H = \{ex, in\}$. В [2] описан переход к более широкому множеству скрытых состояний $Q = S \times H$ путем выполнения взаимно-однозначного преобразования

$$(\bar{s}, \bar{h}) \leftrightarrow (\bar{s}, \bar{q}) \leftrightarrow \bar{q},\ \bar{q} = (s_1, h_1)(s_2, h_2)\dots(s_{|s|}, h_{|s|}) \in Q^{|s|}.$$

Во избежание неясности относительно множества H назовем элементы Q полными состояниями. Выборку X будем рассматривать как подмножество строк из Q: $X \subset Q^*$.

Каждое состояние $q \in Q$ порождает строго одно наблюдаемое состояние $P(s \mid q) = [\mathrm{pr}_s(q) = s]$, где функция проекции определена в теории множеств:

$$\mathrm{pr}_s(q) = s \Leftrightarrow \exists\, h \in H,\ q = (s, h).$$

Аналогично определена проекция $\mathrm{pr}_h : Q \to H$. Функции pr_s и pr_h допускают расширение на строки произвольной длины; под проекцией в этом случае подразумевается конкатенация проекций отдельных элементов строки.

Для рассматриваемой задачи распознавания множество Q состоит из восьми элементов, которые для краткости будем обозначать прописными и строчными латинскими буквами:

$$(A, ex) \equiv A;\ (C, ex) \equiv C;\ (G, ex) \equiv G;\ (T, ex) \equiv T;$$

$$(A, in) \equiv a;\ (C, in) \equiv c;\ (G, in) \equiv g;\ (T, in) \equiv t.$$

Определение 1. Сегментом строки $\bar{q} \in Q^*$ называется произвольная ее подстрока q_i^j, удовлетворяющая следующим условиям:

- скрытые состояния всех элементов сегмента определяются как

$$\mathrm{pr}_h(q_i) = \mathrm{pr}_h(q_{i+1}) = \dots = \mathrm{pr}_h(q_{j-1}) = \mathrm{pr}_h(q_j);$$

- сегмент нельзя распространить на соседние элементы строки $\bar{q}$:

$$(i = 1) \vee (\mathrm{pr}_h(q_{i-1}) \neq \mathrm{pr}_h(q_i));\ (j = |\bar{q}|) \vee (\mathrm{pr}_h(q_{j+1}) \neq \mathrm{pr}_h(q_j)).$$

На парах строк $(\bar{s}, \bar{q})$ задается семейство распределений $M(l)$:

$$P(\bar{s}, \bar{q}) = P(\bar{q})P(\bar{q} \mid \bar{s}) = P(\bar{q})[\mathrm{pr}_s(\bar{q}) = \bar{s}];$$

$$P(\bar{q}) = \varphi(|\bar{q}|)\pi(q_1^l) \prod_{i=l+1}^{|\bar{q}|} p(q_i \mid q_{i-l}^{i-1}), \qquad (1)$$

где $l \in N$ — порядок модели; $\varphi : N \to [0, 1]$ — распределение генерируемых строк по длине; $\pi : Q^l \to [0, 1]$ — распределение начальных вероятностей; $p : Q^l \times Q \to [0, 1]$ — распределение переходных вероятностей.

Таким образом, вероятность генерации строки полных состояний $\bar{q}$ описывается марковской цепью l-го порядка. Распределение φ вносит пренебрежимо малый вклад в функцию правдоподобия (1); в дальнейшем, если не оговорено иначе, оно будет игнорироваться.

В [3] рассмотрены байесовские смеси распределений $\mathrm{Mix}(k, l)$

$$P(\bar{q}) = \sum_{j=1}^{k} w_j P_j(\bar{q}),\ w_j \geq 0,\ \sum_{j=1}^{k} w_j = 1, \qquad (2)$$

где компоненты смеси $P_j(\bar{q})$ описываются уравнением (1) с различными начальными и переходными вероятностями.

Определение 2. Согласно распределению (2) j-й компонентой ($1 \le j \le k$) выборки $X \subset Q^*$ называются те строки выборки, для которых апостериорная вероятность компоненты выше заданного порога ε:

$$X^{(j)} = \{\bar{q} \in X : P(j \mid \bar{q}) > \varepsilon\}.$$

2. ОПРЕДЕЛЕНИЕ ОПТИМАЛЬНОГО ПОРЯДКА МОДЕЛИ

Исходя из формулы вероятности строки полных состояний (1) логарифм правдоподобия для выборки имеет вид

$$\log P(X) = \sum_{|\bar{u}|=l+1} N(\bar{u}) \log p(u_{l+1} \mid u_1^l) + \sum_{|\bar{u}|=l} N_{st}(\bar{u}) \log \pi(\bar{u}) + \text{const}(\pi,\, p); \quad (3)$$

$$N(\bar{u}) = \sum_{\bar{q} \in X} \sum_{i=l+1}^{|\bar{q}|} [q_{i-l}^{i} = \bar{u}]; \quad (4)$$

$$N_{st}(\bar{u}) = \sum_{\bar{q} \in X} [q_1^l = \bar{u}]. \quad (5)$$

Согласно (3) достаточной статистикой для распределений переходных вероятностей модели $M(l)$ являются суммарные числа вхождений в строки выборки $X \subset Q^*$ подстрок длины $l+1$ (см. (4)). Для начальных вероятностей модели $M(l)$ достаточная статистика — количество строк в выборке, начинающихся с фиксированной последовательности длиной l состояний (5). В дальнейшем изложении $N(\bar{u})$ будет обозначать аналогичную (4) статистику для строки $\bar{u} \in Q^*$ произвольной длины.

Согласно оценке наибольшего правдоподобия

$$p(q \mid \bar{u}) = \frac{N(\bar{u}q)}{\sum_{w \in Q} N(\bar{u}w)},\ q \in Q,\ \bar{u} \in Q^l, \quad (6)$$

$$\pi(\bar{u}) = \frac{N_{st}(\bar{u})}{|X|},\ \bar{u} \in Q^l. \quad (7)$$

Рассмотрим несколько способов определения оптимального порядка вероятностной модели $M(l)$, которые используют статистики вида (4) и/или (5).

Тестирование отношения правдоподобия. Для определения оптимального порядка модели применим тест отношения правдоподобия для моделей порядков l и $l+1$ [4]. Условия применимости теста выполнены: модель $M(l)$ является частным случаем модели $M(l+1)$, если для параметров последней выполняются отношения

$$\pi(\bar{u}q) = \text{const}(q) = \pi(\bar{u})\ \ \forall \bar{u} \in Q^l,\ q \in Q;$$

$$p(q \mid w\bar{u}) = \text{const}(w) = p(q \mid \bar{u})\ \ \forall \bar{u} \in Q^l,\ q, w \in Q.$$

Пусть $\hat{L}(M(l) \mid X)$ обозначает максимальное правдоподобие модели l-го порядка при заданной выборке X, получаемое подстановкой (6), (7) в (1). Тогда величина

$$x_{lh}^2 = 2 \log \hat{L}(M(l+1) \mid X) - 2 \log \hat{L}(M(l) \mid X) \quad (8)$$

асимптотически распределена согласно закону хи-квадрат с количеством степеней свободы

$$df_{lh} = K_{l+1} - K_l, \quad (9)$$

где K_l — число независимых параметров модели $M(l)$. Общее число независимых параметров модели $M(l)$ составляет

$$K_l^{(\max)} = (|Q|^l - 1) + |Q|^l\ (|Q| - 1) = |Q|^{l+1} - 1, \quad (10)$$

где первое слагаемое соответствует начальным вероятностям модели, второе —

переходным вероятностям. Для рассматриваемой задачи оценка (10) сильно завышена ввиду специфического вида переходных участков между сегментами скрытых состояний. Таким образом, использование этой оценки может привести к завышению оптимального порядка вероятностной модели. В связи с этим будем рассматривать распределение (8) не на всем множестве выборок последовательностей полных состояний, а на выборках из корректных элементов, определенных следующим образом [5, 6].

Определение 3. Строка полных состояний (ген) $\bar{q}$ называется корректной для модели порядка l тогда и только тогда, когда для нее выполнены следующие условия:

- длина строки ограничена снизу: $|\bar{q}| \geq l$;
- первые три состояния строки фиксированы: $q_1^3 = ATG$ или $q_1^3 = ATg$;
- длина всех сегментов $\bar{q}$, кроме, возможно, первого и последнего, не меньше, чем l;
- сегменты q_i^j, соответствующие интронам ($\mathrm{pr}_h(q_i) = in$), начинаются с последовательности нуклеотидов гуанин–тимин: $q_i^{i+1} = gt$ и заканчиваются последовательностью аденин–гуанин: $q_{j-1}^j = ag$.

Определению 3 удовлетворяют 90–95 % генов живых организмов. Условия определения соответствуют дополнительным ограничениям на начальные и переходные вероятности модели $M(l)$. Число независимых параметров модели с учетом этих ограничений приведено в табл. 1. Итак, на множестве выборок корректных строк при росте размера выборки количество степеней свободы случайной величины x_{lh}^2 (8) определяется по формуле (9), причем числа K_l и K_{l+1} берутся из табл. 1. Вероятность генерации выборки моделью $M(l)$ при альтернативной модели $M(l+1)$ соответствует

$$P_{lh} = P(\chi^2(df_{lh}) \geq x_{lh}^2). \tag{11}$$

Критерий отношения правдоподобия применим и для определения оптимального количества компонент в смеси распределений (2). Действительно, смесь распределений с k составляющими является частным случаем смеси с k' составляющими ($k' > k$), в которой дополнительные компоненты имеют нулевой вес. Число независимых параметров смеси с k компонентами, которые описываются моделями l-го порядка, равно $kK_l + k - 1$. Нулевой гипотезой будем считать генерацию выборки смесью (2) с k составляющими, в качестве альтернативной — ее генерацию смесью с $k+1$ составляющими:

$$x_{\text{mix}}^2 = 2\log \hat{L}(\text{Mix}(k+1, l) \mid X) - 2\log \hat{L}(\text{Mix}(k, l) \mid X),$$

Таблица 1. Количество независимых параметров модели $M(l)$

Порядок модели, l	**Число независимых вероятностей**						
	Общее число, $K_l^{(\max)}$	начальных	переходных	Всего, K_l	начальных	переходных	Общее число
		Теоретическая оценка			Для генома человека		
1	63	0	32	32	0	32	32
2	511	0	128	128	0	128	128
3	4095	1	536	537	1	536	537
4	32767	5	2240	2245	5	2234	2239
5	262143	24	9344	9368	20	9085	9105
6	2097151	103	38912	39015	77	35815	35892
7	16777215	431	161792	162223	286	134957	135243
8	134217727	1791	671744	673535	1022	486932	487954
9	1073741823	7426	2785280	2792706	3035	1705565	1708600
10	8589934591	30719	11534336	11565055	6588	5382891	5389479

где $\hat{L}(\mathrm{Mix}(k, l) \mid X)$ — максимальное правдоподобие, получаемое с помощью ЕМ-алгоритма [3]. Величина x^2_{mix} имеет асимптотическое распределение согласно закону хи-квадрат с числом степеней свободы

$$df_{\mathrm{mix}} = (k+1)K_l + k - (kK_l + k - 1) = K_l + 1;$$

уровень значимости нулевой гипотезы составляет

$$P_{\mathrm{mix}} = P(\chi^2(df_{\mathrm{mix}}) \geq x^2_{\mathrm{mix}}). \tag{12}$$

Тестирование эргодического свойства. Ожидаемые значения величин $N(\bar{u})$, $\bar{u} \in Q^{l+1}$, можно получить, исходя из эргодичности вероятностной модели $M(l)$. При переходе от состояний из множества Q к их последовательностям длины l отношения между полными состояниями модели описываются марковской цепью первого порядка. Элементы матрицы переходов B для преобразованной цепи, определяемые как

$$B(\bar{u}, \bar{v}) = P\{q^i_{i-l+1} = \bar{v} \mid q^{i-1}_{i-l} = \bar{u}\},\ \bar{u}, \bar{v} \in Q^l,\ \bar{q} \in X,$$

вычисляются по формуле

$$B(\bar{u}, \bar{v}) = \begin{cases} p(v_l \mid \bar{u}), & \text{если } v^{l-1}_1 = u^l_2; \\ 0, & \text{если } v^{l-1}_1 \neq u^l_2. \end{cases}$$

Пусть для модели $M(l)$ существует эргодическое распределение $\tilde{\pi} : Q^l \to [0, 1]$:

$$\tilde{\pi} = \tilde{\pi}B;\ \tilde{\pi} = \lim_{n \to \infty} \pi_0 B^n$$

($\pi_0 : Q^l \to [0, 1]$ — произвольное начальное распределение). Тогда ожидаемое число вхождений в строки выборки произвольной подстроки u длины $l+1$ составляет

$$\tilde{N}(\bar{u}) = N\tilde{\pi}(u^l_1)\, p(u_{l+1} \mid u^l_1),\ N = \sum_{|u|=l+1} N(\bar{u}), \tag{13}$$

где N — нормировочный множитель, приблизительно равный общей длине строк выборки.

В соответствии с критерием согласия Пирсона [7] для оценки близости чисел $N(\bar{u})$ и $\tilde{N}(\bar{u})$ следует использовать величину

$$x^2_{\mathrm{erg}} = \sum_{\bar{u}} \frac{(N(\bar{u}) - \tilde{N}(\bar{u}))^2}{\tilde{N}(\bar{u})}; \tag{14}$$

суммирование проводится по всем строкам скрытых состояний длины $l+1$ с положительным ожиданием $\tilde{N}(\bar{u}) > 0$; x^2_{erg} имеет асимптотическое распределение хи-квадрат с числом степеней свободы

$$df_{\mathrm{erg}} = V_{\mathrm{erg}} - K_{\mathrm{erg}} - 1;\ V_{\mathrm{erg}} \equiv |\{\bar{u} \in Q^{l+1} : \tilde{N}(\bar{u}) > 0\}|,$$

где K_{erg} — количество независимых параметров модели $M(l)$, оцениваемых с помощью статистики $\{N(\bar{u}) : \bar{u} \in Q^{l+1}\}$, V_{erg} — число слагаемых в сумме (14). Вероятность генерации выборки моделью $M(l)$ согласно предложенной статистике определяется формулой

$$P_{\mathrm{erg}} = P(\chi^2(df_{\mathrm{erg}}) \geq x^2_{\mathrm{erg}}). \tag{15}$$

Критерий на основе эргодического свойства является слабым: эргодическое распределение для строк длины l может существовать и для модели более высокого порядка; в таком случае по-прежнему будет выполняться уравнение (13). Таким образом, описанный критерий подходит для нахождения нижнего предела порядка модели.

Тестирование марковского свойства. Марковское свойство модели $M(l)$ означает, что вероятностное распределение произвольного полного состояния определяется l предыдущими состояниями последовательности полных состояний. Очевидный способ проверки этого свойства — подтверждение условной независимости двух элементов строки полных состояний при условии, что между ними расположены заданные l состояний:

$$P(u_{l+2}\,|\,u_1^{l+1}) = P(u_{l+2}\,|\,u_2^{l+1}) \Leftrightarrow P(u_1, u_{l+2}\,|\,u_2^{l+1}) = P(u_{l+2}\,|\,u_2^{l+1})P(u_1\,|\,u_2^{l+1}),\ \ \bar{u} \in Q^{l+2}.$$

Таким образом, для оценки выполнения марковского свойства можно использовать известный критерий Пирсона для условной независимости [8], однако при этом возникает проблема подсчета количества степеней свободы: вероятности вида $P(u_1|\,u_2^{l+1})$ не входят в параметры модели $M(l)$, но зависят от ее переходных вероятностей. В силу этого ниже описан альтернативный способ проверки марковского свойства модели.

Рассмотрим условные вероятности вида $P(u_{l+1}^{l+2}|\,u_1^l)$ для произвольных строк $\bar{u} \in Q^{l+2}$, входящих в последовательности полных состояний из выборки X. Согласно модели $M(l)$

$$P(u_{l+1}^{l+2}|\,u_1^l) = p(u_{l+1}|\,u_1^l)\,p(u_{l+2}|\,u_2^{l+1}). \tag{16}$$

На основании статистики $\{N(\bar{u}) : \bar{u} \in Q^{l+2}\}$ оценим переходные вероятности модели $M(l)$. Согласно (16) ожидаемое значение количества вхождений в выборку подстроки $\bar{u} \in Q^{l+2}$ определяется формулой

$$\widetilde{N}(\bar{u}) = p(u_{l+1}|\,u_1^l)\,p(u_{l+2}|\,u_2^{l+1}) \sum_{q,\,w \in Q} N(u_1^l q w). \tag{17}$$

Величина

$$x^2_{\text{mar}} = \sum_{\bar{u}} \frac{(N(\bar{u}) - \widetilde{N}(\bar{u}))^2}{\widetilde{N}(\bar{u})}, \tag{18}$$

где сумма берется по всем строкам длины $l+2$ с положительным ожидаемым числом вхождений $\widetilde{N}(\bar{u}) > 0$, в соответствии с критерием согласия Пирсона имеет асимптотическое распределение хи-квадрат с числом степеней свободы

$$df_{\text{mar}} = V_{\text{mar}} - R_{\text{mar}} - K_{\text{mar}}, \tag{19}$$

$$V_{\text{mar}} = |\,\{\bar{u} \in Q^{l+2} : \widetilde{N}(\bar{u}) > 0\}\,|;\ \ R_{\text{mar}} = |\,\{\bar{v} \in Q^{l} : N(\bar{v}) > 0\}\,|.$$

Здесь K_{mar} — количество независимых параметров модели, оцениваемых с помощью достаточной статистики $\{N(\bar{u}) : \bar{u} \in Q^{l+2}\}$; V_{mar} — число суммируемых переменных в выражении (18), R_{mar} — количество различных сумм, вычисляемых в (17). Вероятность порождения выборки моделью $M(l)$ определяется как

$$P_{\text{mar}} = P(\chi^2(df_{\text{mar}}) \geq x^2_{\text{mar}}). \tag{20}$$

Предложенный тест является более сильным по сравнению с тестированием эргодического свойства, поскольку для модели порядка $l+1$ или выше равенство (16) не выполняется. Недостаток теста — использование большого числа переменных в достаточной статистике $\{N(\bar{u}) : \bar{u} \in Q^{l+2}\}$, что делает его ненадежным для меньшего порядка модели l по сравнению с другими рассматриваемыми тестами.

Информационный критерий Акаике. Подход, альтернативный тестированию статистических гипотез, заключается в выборе оптимального порядка модели с помощью информационных критериев. Один из таких критериев, применимый для марковских моделей, — информационный критерий Акаике [9, 10].

Формула вычисления критерия для модели $M(l)$:

$$AIC_l = -2 \log \hat{L}(M(l) \mid X) + 2K_l, \tag{21}$$

где $\hat{L}(M(l) \mid X)$ — максимальное правдоподобие на выборке X при использовании модели l-го порядка; K_l — число независимых параметров модели (табл. 1). Аналогично для смесей распределений (eq:mix) критерий Акаике имеет вид

$$AIC_{k,l} = -2 \log \hat{L}(\mathrm{Mix}\,(k, l) \mid X) + 2kK_l + 2k - 2. \tag{22}$$

Оптимальной вероятностной модели соответствует наименьшее значение критерия.

3. ВЫЧИСЛИТЕЛЬНЫЙ ЭКСПЕРИМЕНТ

Для определения оптимального порядка вероятностной модели использовались реальные и модельные данные. Реальные данные представляли собой геномы двух групп организмов из репозитория NCBI [11]:

- геномы растений с «простой» структурой генов (сравнительно малое число и небольшая длина интронов) — *Oryza sativa* (рис), *Populus trichocarpa* (тополь), *Vitis vinifera* (виноград), *Glycine max* (соя), *Arabidopsis thaliana*, *Medicago truncatula* (модельные виды для генетических исследований);
- геномы млекопитающих, обладающие «сложной» структурой (большое количество и большая длина интронов) — *Homo sapiens* (человек), *Mus musculus* (мышь), *Rattus norvegicus* (крыса), *Papio anubis* (павиан), *Sus scrofa* (свинья), *Bos taurus* (корова), *Equus caballus* (лошадь).

Размер каждого рассматриваемого генома — порядка 10^4 строк суммарной длиной $\sim 10^8$ состояний. Рассматривались гены, не содержащие неизвестных нуклеотидов и удовлетворяющие определению 3 для модели порядка $l_{\max} = 10$.

Модельные данные строились на основе реальных: начальные и переходные вероятности порядка l_{true}, а также распределение строк по длине φ (эмпирическое с окном сглаживания 100) вычислялись, исходя из данных определенного генома, после чего в соответствии с распределением вероятности (1) генерировалось 10000 строк полных состояний. Таким образом, размер модельных данных имел тот же порядок, что и размер реальных данных.

Первым этапом эксперимента стало изучение описанных тестов правдоподобия для моделей до десятого порядка включительно (тест марковского свойства — до девятого порядка) на модельных данных. Рассмотрение моделей более высокого порядка невозможно, поскольку размер выборки становится недостаточным для надежного определения параметров вероятностного распределения (см. табл. 1). Кроме того, малые значения используемых в тестах статистик противоречат общим условиям применимости критерия согласия Пирсона [12]. Пороговая величина значимости для тестов полагалось равным 0,05. Для реализации критериев использовался программный комплекс на языке Java.

Таблица 2. Тест отношения правдоподобия и информационный критерий Акаике для модельных данных, сгенерированных на основе генома сои (*Glycine max*), $l_{\text{true}} = 7$

Порядок модели, l	$\log \hat{L}(M(l) \mid X) \cdot 10^{-7}$	K_l	x^2_{lh}	df_{lh}	P_{lh}	$AIC_l \cdot 10^{-7}$
1	–1,43877	32	1298354	96	0	2,87753
2	–1,43228	128	986694	409	0	2,86455
3	–1,42734	537	546952	1708	0	2,85469
4	–1,42461	2245	526678	7123	0	2,84926
5	–1,42197	9368	809670	29647	0	2,84413
6	–1,41793	39015	784880	123208	0	2,83663
7	–1,41400	162223	444130	511312	1	2,83125
8	–1,41178	673535	1634212	2119171	1	2,83703
9	–1,40361	2792706	—	—	—	2,86307

Таблица 3. Тест марковского свойства для модельных данных, сгенерированных на основе генома человека, $l_{\text{true}} = 8$

Порядок модели, l	x^2_{mar}	K_{mar}	V_{mar}	R_{mar}	df_{mar}	$x^2_{\text{mar}} / df_{\text{mar}}$	P_{mar}
1	2323279	32	212	8	172	13507,43	0
2	1411879	128	720	40	552	2557,75	0
3	1807045	536	3008	168	2304	784,31	0
4	3072082	2240	12435	704	9491	323,68	0
5	4894989	9296	49744	2927	37521	130,46	0
6	5321342	37850	190598	11884	140864	37,78	0
7	6080502	150815	694128	46507	496806	12,24	0
8	1381646	581166	2449399	173852	1694381	0,82	1
9	4798829	2167515	8485295	628378	5689402	0,84	1

Таблица 4. Тест эргодического свойства для генома лошади (*Equus caballus*)

Порядок модели, l	x^2_{erg}	V_{erg}	K_{erg}	df_{erg}	P_{erg}, %
1	3,49	40	32	7	83,63
2	358,42	168	128	39	0,00
3	816,06	704	536	167	0,00
4	1236,11	2944	2240	703	0,00
5	1731,62	12175	9344	2830	100,00
6	2416,23	48475	38624	9850	100,00
7	3032,34	184255	155402	28852	100,00

Точность предложенных критериев была проверена на модельных данных. Результаты вычислительного эксперимента показывают, что тест отношения правдоподобия (11), информационный критерий Акаике (21) и тест марковского свойства (20) дают точные оценки порядка модели, если $l_{\text{true}} \geq 6$ (табл. 2, 3). При этом тест эргодичности не позволяет определить порядок модели ввиду причин, описанных в разд. 2, он выполняется для произвольного порядка модели $l \geq 1$.

Вторым этапом эксперимента стало применение статистических тестов, рассмотренных в разд. 2, для геномов биологических видов. В отличие от модельных данных тест эргодичности (15) не является тривиальным; для большинства рассмотренных геномов он выполняется для модели первого порядка и моделей порядка $l \geq l_{\text{erg}}$, где $l_{\text{erg}} > 2$ (табл. 4). Число l_{erg}, таким образом, является нижней границей порядка модели.

Тест марковского свойства не выполняется ни для одного исследованного генома, кроме тополя (*Populus trichocarpa*); тем не менее, отношение $x^2_{\text{mar}} / df_{\text{mar}}$ уменьшается при увеличении порядка модели. Тест отношения правдоподобия также не позволяет определить оптимальный порядок модели l_{lh} для большинства геномов, однако отношение x^2_{lh} / df_{lh} для больших значений l близко к единице. Критерий Акаике (21) для реальных данных является унимодальной функцией от параметра l; точка его минимума l_{AIC} надежно определена для всех геномов, кроме генома человека. Итоги определения оптимального порядка модели для геномов приведены в табл. 5.

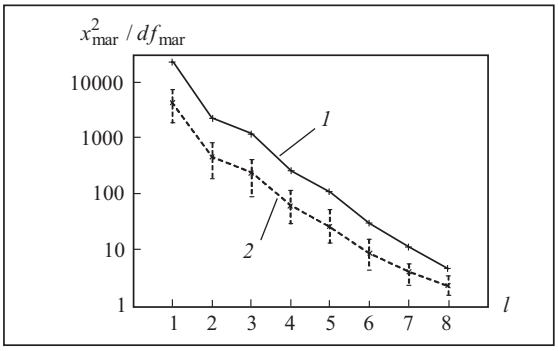

Рис. 1. График теста марковского свойства для генома мыши: *1* — для распределения (1); *2* — для смеси (2) с пятью компонентами (среднее значение и диапазон значений)

Таблица 5. Оптимальный порядок модели для исследованных геномов

Геном	l_{lh}	$x^2_{lh} / df_{lh},\ l = 10$	l_{erg}	$x^2_{\text{mar}} / df_{\text{mar}},\ l = 9$	l_{AIC}
A. thaliana	≥10	1,007	6	1,421	7
Glycine max	≥10	1,172	6	1,473	8
M. truncatula	9	0,853	6	1,194	7
Oryza sativa	9	0,837	7	1,019	8
Populus trichocarpa	8	0,735	6	0,981	7
Vitis vinifera	9	0,919	5	1,235	8
Homo sapiens	≥10	2,328	6	3,427	10
Mus musculus	≥10	1,698	6	2,338	9
Rattus norvegicus	≥10	1,151	5	1,619	9
Papio anubis	≥10	1,520	5	2,322	9
Sus scrofa	≥10	1,562	5	2,340	9
Bos taurus	≥10	1,762	5	2,520	9
Equus caballus	≥10	1,117	5	1,557	9

Таблица 6. Оптимальный порядок модели для компонент генома свиньи (*Sus scrofa*)

Составляющие смеси	l_{lh}	$x^2_{lh} / df_{lh},\ l = 10$	l_{erg}	$x^2_{\text{mar}} / df_{\text{mar}},\ l = 8$	l_{AIC}
Геном в целом	≥10	1,562	5	5,208	9
Компонента 1	9	0,861	4	3,012	8
Компонента 2	9	0,661	5	1,710	8
Компонента 3	9	0,683	3	2,280	8
Компонента 4	9	0,726	2	2,381	8
Компонента 5	9	0,593	3	1,762	8

Таблица 7. Тест отношения правдоподобия и информационный критерий Акаике для смесей распределений на основе генома риса (*Oryza sativa*)

Порядок модели, l	Число компонент, k	$\log \hat{L}(\text{Mix}(k,l) \mid X) \cdot 10^{-8}$	x^2_{lh}	df_{lh}	P_{lh}	$AIC_{l,k} \cdot 10^{-8}$
7	1	– 0,99927	1260142	162224	0	2,00179
	2	– 0,99297	336768	162224	0	1,99243
	3	– 0,99129	260320	162224	0	1,99231
	4	– 0,98999	223424	162224	0	1,99295
	5	– 0,98887	142600	162224	1	1,99396
	6	– 0,98816	—	—	—	1,99578
8	1	– 0,99401	1668498	673536	0	2,00149
	2	– 0,98567	746206	673536	0	1,99828
	3	– 0,98194	663322	673536	1	2,00429
	4	– 0,97862	617600	673536	1	2,01112
	5	– 0,97553	—	—	—	2,01842

Третьим этапом вычислительного эксперимента стало исследование смесей вероятностных распределений (2). Было выяснено, что для компонент выборки, найденных согласно определению 2 с пороговым значением апостериорной вероятности $\varepsilon = 0{,}95$, предложенные тесты оценивают порядок модели ниже, чем при рассмотрении одного вероятностного распределения (1) для выборки в целом (рис. 1, табл. 6). В отличие от выборок в целом по меньшей мере для одной из компонент всех геномов растений и большинства геномов млекопитающих тест отношения правдоподобия (11) и тест марковского свойства (20) дают фактические оценки порядка модели $l_{lh} \le 9$, $l_{\text{mar}} \le 9$.

Тест отношения правдоподобия (12) и критерий Акаике для смесей распределений (22) показывают, что оптимальное количество компонент смеси составляет около пяти для моделей седьмого порядка и уменьшается при увеличении l — порядка модели. Для большинства геномов, в частности для риса, смеси с точки зрения критерия Акаике эффективнее отдельных моделей (табл. 7).

ЗАКЛЮЧЕНИЕ

Рассмотрены четыре способа определения оптимального порядка вероятностной модели на основе аппарата тестирования статистических гипотез и информационного критерия Акаике. Результаты вычислительных экспериментов, проведенных на геномах биологических видов, в целом согласуются с приведенным в [1] методом выбора порядка модели на основе пятикратной кросс-валидации. Для геномов растений оптимальный порядок модели составляет 7 или 8, для геномов млекопитающих — 9.

Рассмотрены два способа определения оптимального количества компонент k в байесовских смесях распределений (2). Результаты согласуются с приведенной в [3] оценкой $k^* \approx 5$, сделанной с применением кросс-валидации. Показано, что существенное повышение качества распознавания при использовании смесей распределений связано с тем, что подмножества выборки, описываемые отдельными компонентами смеси, лучше оцениваются марковскими цепями, чем выборка в целом.

Отклонение от марковского свойства в задачах распознавания фрагментов генов побуждает к рассмотрению дискриминантных моделей, таких как условные марковские сети (conditional random fields) [13], марковские цепи с максимальной энтропией (maximum entropy Markov models) [14] и метод опорных векторов для скрытых марковских моделей (hidden Markov support vector machines) [15].

СПИСОК ЛИТЕРАТУРЫ

1. 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.
2. Островский А.В. Определение вторичной структуры белков с помощью моделей Маркова // Международный научно-технический журнал «Проблемы управления и информатики». — 2013. — № 2. — С. 140–147.
3. Сергиенко И.В., Гупал А.М., Островский А.В. Использование EM-алгоритма для классификации генов // Кибернетика и системный анализ. — 2015. — **51**, № 1. — С. 48–58.
4. Wilks S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses // The Annals of Mathematical Statistics. — 1938. — **9**, N 1. — P. 60–62.
5. Ридли М. Геном: автобиография вида в 23 главах. — М.: Эксмо, 2008. — 432 с.
6. Splicing of messenger RNA precursors / R.A. Padgett, P.J. Grabowski, M.M. Konarska, S. Seiler, P.A. Sharp // Annual Review of Biochemistry. — 1986. — **55.** — P. 1119–1150.
7. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling // Philosophical Magazine. Series 5. — 1900. — **50** (302). — P. 157–175.
8. Справочник по прикладной статистике. В 2-х т. Т. 1. Пер. с англ. / Под ред. Э. Ллойда, У. Ледермана, Ю.Н. Тюрина. — М.: Финансы и статистика, 1989. — 510 с.
9. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. — 1974. — **19**, N 6. — P. 716–723.
10. Tong H. Determination of the order of a Markov chain by Akaike's information criterion // Journal of Applied Probability. — 1975. — **12.** — P. 488–497.
11. Genbank / D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman et al. // Nucleic Acids Research. — 2013. — **41** (Database issue). — P. D36–D42.
12. Cochran W.G. The χ^2 test of goodness of fit // The Annals of Mathematical Statistics. — 1952. — **23.** — P. 315–345.
13. Lafferty J., McCallum A., Pereira F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // Proc. 18th Intern. Conf. on Machine Learning, Williamstown (MA), USA, 2001. — P. 282–289.
14. McCallum A., Freitag D., Pereira F.C. Maximum entropy Markov models for information extraction and segmentation // Proc. 17th Intern. Conf. on Machine Learning, Stanford (CA), USA, 2000. — P. 591–598.
15. Altun Y., Tsochantaridis I., Hofmann T. Hidden Markov support vector machines // Proc. 20th Intern. Conf. on Machine Learning, Washington (DC), USA, 2003. — **3.** — P. 3–10.

Поступила 16.03.2015