Распознавание на основе скрытых марковских моделей

ОСТРОВСКИЙ Алексей Викторович

Институт кибернетики имени В. М. Глушкова НАН Украины

9 декабря 2014 г.

План доклада

1. Постановка задач распознавания скрытых последовательностей.
2. Обзор существующих методов решения задач распознавания.
3. Применение обобщений марковских моделей для распознавания.
4. Композиции марковских моделей.

Биоинформатика

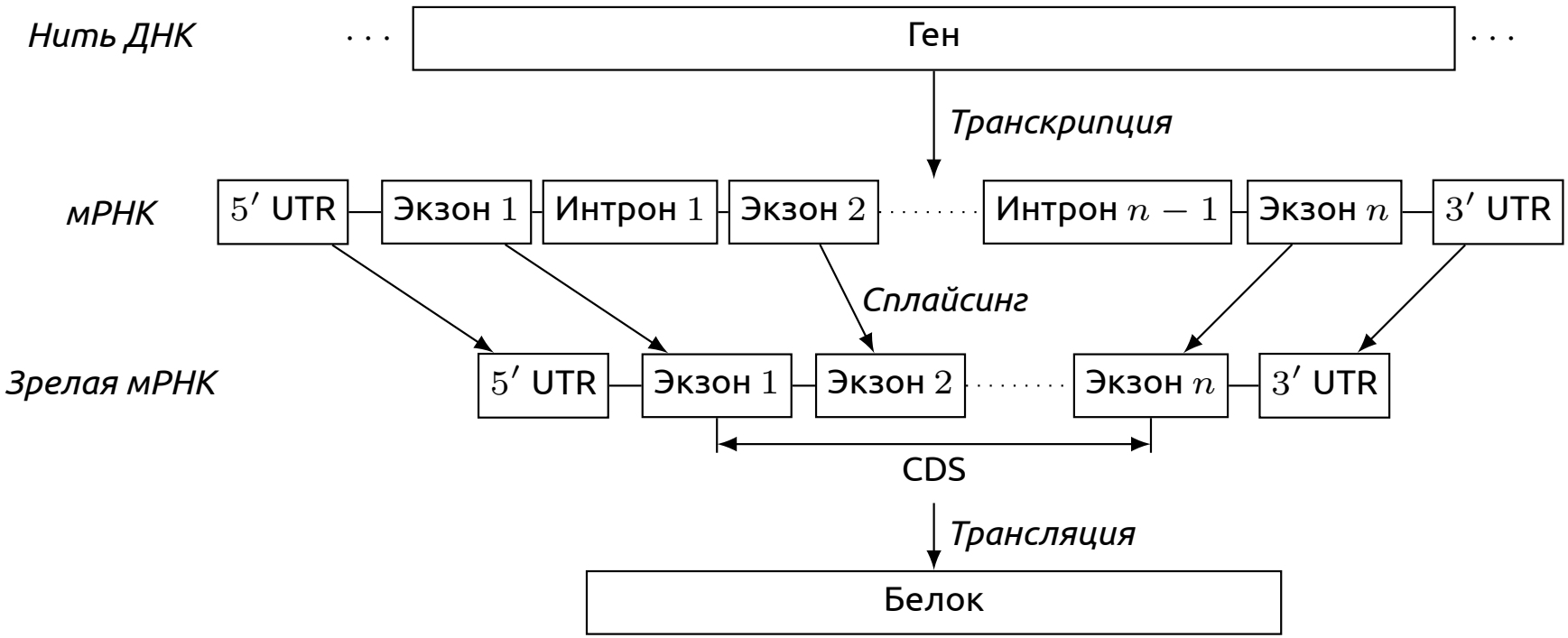
Определение

Биоинформатика — применение методов математической статистики и информатики для анализа и обработки биологических данных: последовательностей нуклеотидов (ДНК) и аминокислот (белки).

Цели исследований:

- раскрытие процесса эволюции;
- повышение эффективности селекции;
- создание лекарств и белков с заданными свойствами;
- диагностирование генетических заболеваний.

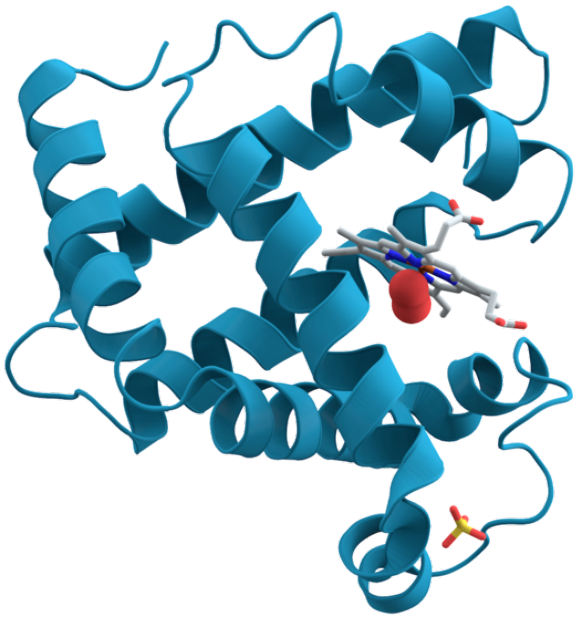
Задача анализа ДНК



Задача

Определить разбиение участка гена, заключенного между начальным и конечным некодирующими участками, который задан как последовательность известных нуклеотидов, на экзоны и интроны.

Задача распознавания вторичной структуры белка



Задача

На основании заданной последовательности аминокислот белка определить для каждой аминокислоты вторичную структуру (спираль, β-лист либо отсутствие структуры), в которую она входит.

Общая задача распознавания

Ген:

Нуклеотиды	ATGG…GCAA	GTAA…TCAG	ATTT…AAAG	$\cdots$	GTAA…CCAG	GGTG…ATAA
Структура	Экзон	Интрон	Экзон	$\cdots$	Интрон	Экзон

Белок:

Аминокислоты	NL	KLGFLV	KQPEE	PWFQTEWKFADKAGKDL	GF	EVIKIA	VPD	$\cdots$
Структура	–	β	–	α	–	β	–	$\cdots$

Состояния	ДНК	Белки
наблюдаемые (S)	нуклеотиды	аминокислоты
скрытые (H)	экзоны, интроны	α-спирали, β-листы, нерегулярные структуры

Вероятностная постановка задачи

Задача

Найти алгоритм $\mathcal{A}$, который для произвольной строки наблюдаемых состояний $\bar{s} \in S^*$ определяет строку скрытых состояний $\bar{h} \in H^*$, максимизирующую критерий качества $\mathcal{L}(\bar{s}, \bar{h})$.

Принцип максимума правдоподобия: поиск наиболее вероятной строки скрытых состояний.

$$\mathcal{A}(\bar{s}) = \arg\max_{\bar{h}} P(\bar{h} \mid \bar{s}, \Theta) = \arg\max_{\bar{h}} P(\bar{h}, \bar{s} \mid \Theta);$$

Θ — параметры модели.

Задача (обучение параметров модели)

По заданной обучающей выборке $\{(\bar{s}_j, \bar{h}_j)\}_{j=1}^N$ найти оптимальные параметры вероятностной модели Θ^*:

$$\Theta^* = \arg\max_{\Theta} \prod_{j=1}^{N} P(\bar{s}_j, \bar{h}_j \mid \Theta).$$

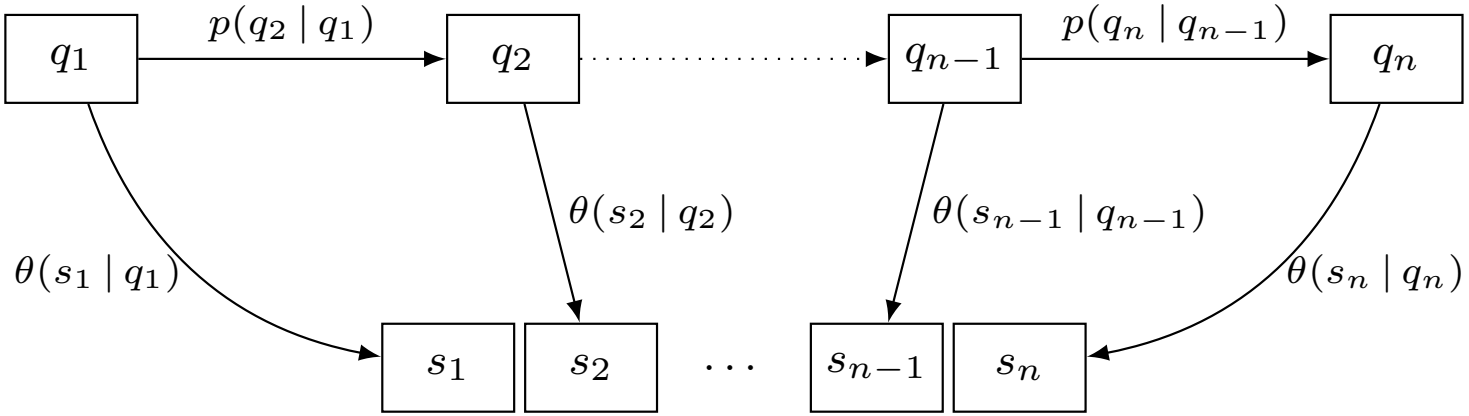
Скрытая марковская модель

Определение

Однородная скрытая марковская модель — дискретный случайный процесс, характеризуемый следующим набором параметров и переменных:

- Q — конечное множество скрытых состояний;
- S — конечное множество наблюдаемых состояний модели;
- $\varphi(d)$ — распределение строк скрытых состояний по длине;
- $\pi(q)$ — начальная вероятность скрытого состояния $q \in Q$;
- $p(q_j \mid q_i)$ — вероятность перехода из $q_i \in Q$ в $q_j \in Q$ за единицу времени;
- $\theta(s \mid q)$ — вероятность наблюдения состояния $s \in S$ при скрытом состоянии $q \in Q$.

Скрытая марковская модель



Скрытые параметры модели: $\bar{q} \equiv q_1 q_2 \ldots q_n$.

Вычисление совместной вероятности:

$$P(\bar{s}, \bar{q}) = \varphi(n)\pi(q_1) \prod_{i=2}^{n} p(q_i \mid q_{i-1}) \prod_{i=1}^{n} \theta(s_i \mid q_i);$$

$$n = |\bar{s}|.$$

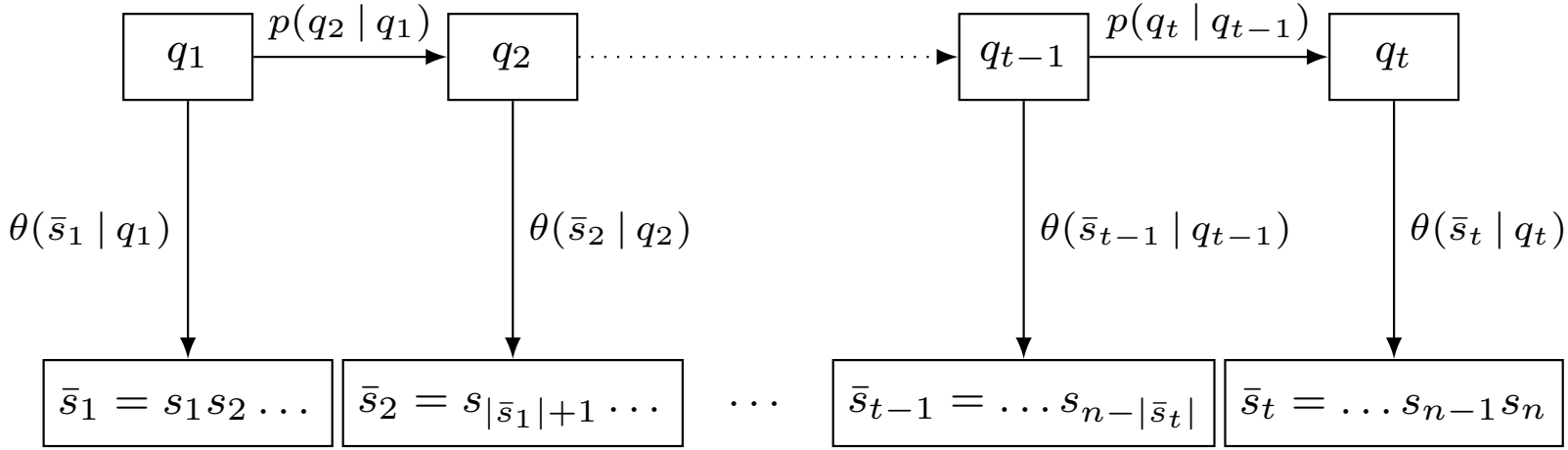
Обобщенная скрытая марковская модель

Определение

Однородная обобщенная скрытая марковская модель — дискретный случайный процесс, характеризуемый следующими параметрами:

- Q — конечное множество скрытых состояний;
- S — конечное множество наблюдаемых состояний системы;
- $\varphi(d)$ — распределение строк скрытых состояний по длине;
- $\pi(q)$ — начальная вероятность скрытого состояния $q \in Q$;
- $p(q_j \mid q_i)$ — вероятность перехода из $q_i \in Q$ в $q_j \in Q$ за единицу времени;
- $\theta(\bar{s} \mid q_i)$ — вероятность генерации скрытым состоянием $q_i \in Q$ строки $\bar{s} \in S^*$ в произвольный момент времени.

Обобщенная скрытая марковская модель



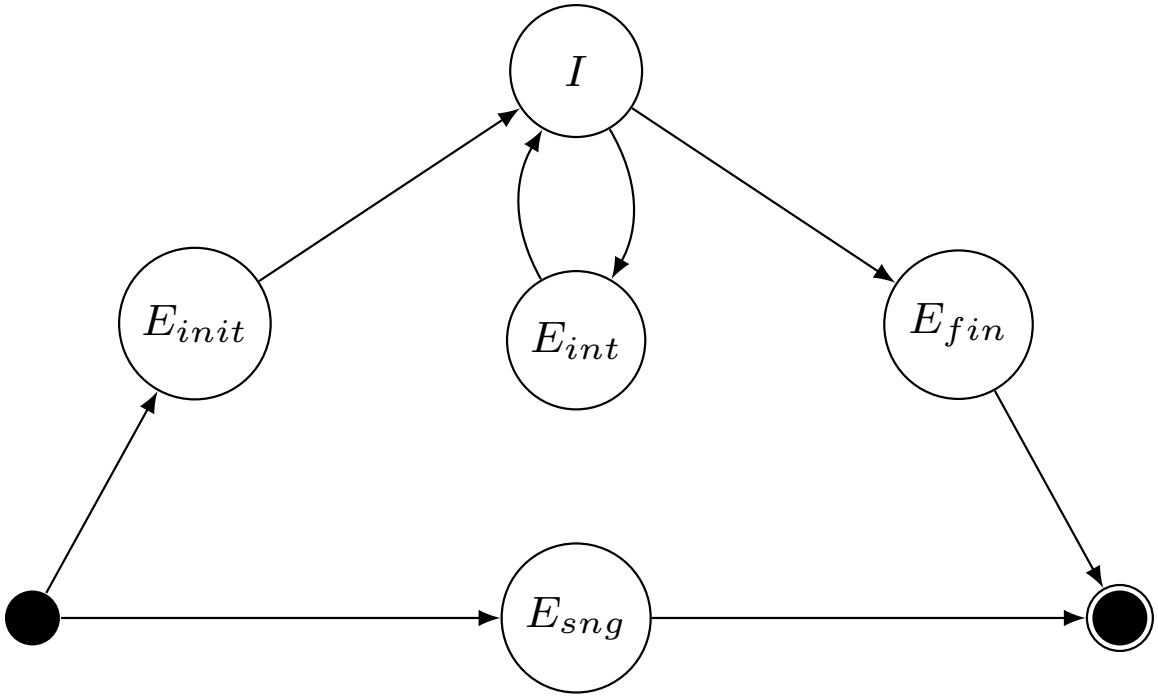
Скрытые параметры модели: $(\bar{q}, |\bar{s}_1|, \ldots, |\bar{s}_t|)$, $t = |\bar{q}|$.

$$P(\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_t, \bar{q}) \longrightarrow \max, \quad \text{s.t.} \quad \bar{s}_1 \ldots \bar{s}_t = \bar{s}.$$

Вычисление совместной вероятности:

$$P(\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_t, \bar{q}) = \varphi(t)\pi(q_1) \prod_{i=2}^{t} p(q_i \mid q_{i-1}) \prod_{i=1}^{t} \theta(\bar{s}_i \mid q_i).$$

Пример OCMM для распознавания фрагментов генов



$$Gene = E_{sng} | (E_{init}(I E_{int}) + E_{fin})$$

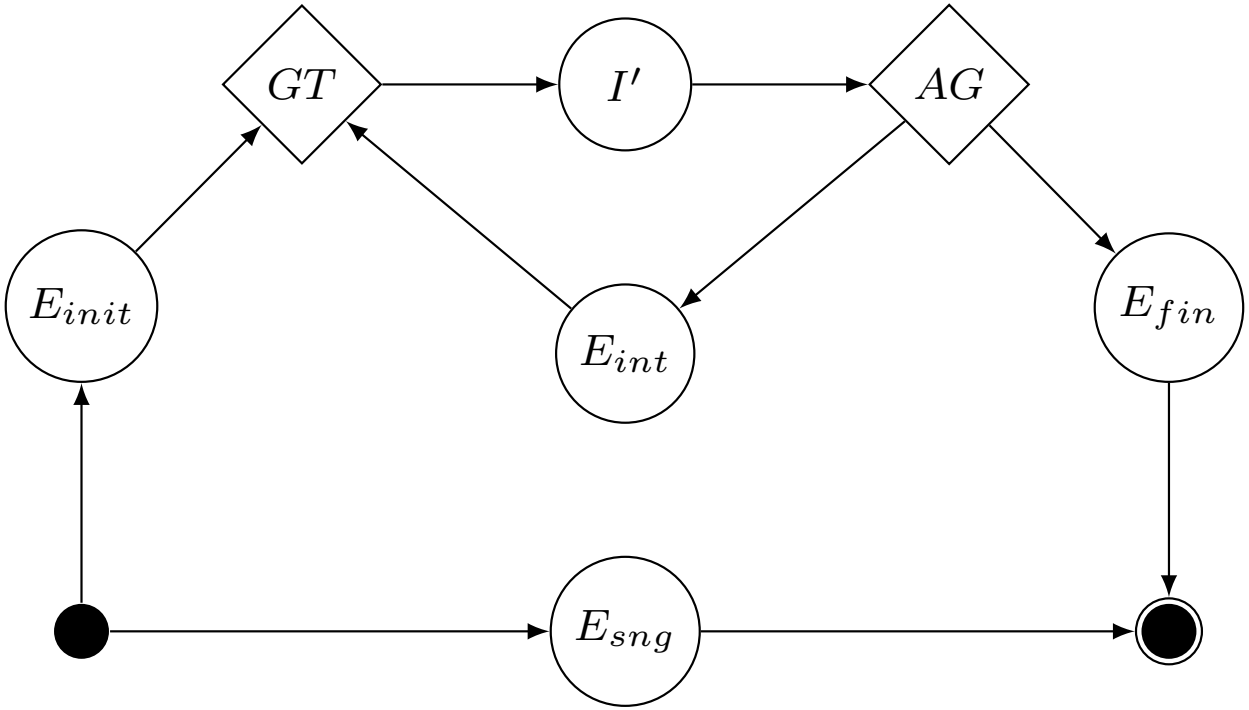
E_{sng} — одиночный экзон; E_{fin} — конечный экзон; I — интрон.
E_{init} — начальный экзон; E_{int} — промежуточный экзон;

$\theta(\bar{s} \mid \cdot)$ — марковские цепи 5-го порядка; распределение длин эмпирическое со сглаживанием для экзонов, гипергеом. для интронов.

Модификация OCMM

Проблема: низкая скорость нахождения максимума правдоподобия.

Решение: использование *сигналов* (состояний с фиксированной длиной).



$$Gene = E_{sng} | (E_{init} (\text{GT}\ I'\ \text{AG}\ E_{int}) + E_{fin})$$

Распознавание сигнала GT**:** …CAT CGAGTCGATA TAC… $\longrightarrow$ SVM, нейросеть, …

Недостатки OCMM

- Большая сложность вычислений;
- требуется, чтобы наблюдаемой строке соответствовало малое количество скрытых состояний;
- использование сигналов невозможно для многих задач (напр., для распознавания структуры белков);
- сложный вид модели затрудняет использование в качестве составляющей (напр., в композициях).

Зависимости между состояниями

		влияющие состояния					
Наблюдаемое состояние	D	L	G	F	E	V	I
Структура	α	α	–	–	β	β	?
		влияющие состояния					

$$P(h_i, s_i \mid h_1 \dots h_{i-1}, s_1 \dots s_{i-1}) = P(h_i, s_i \mid h_{i-l} \dots h_{i-1}, s_{i-l} \dots s_{i-1}).$$

Предложение

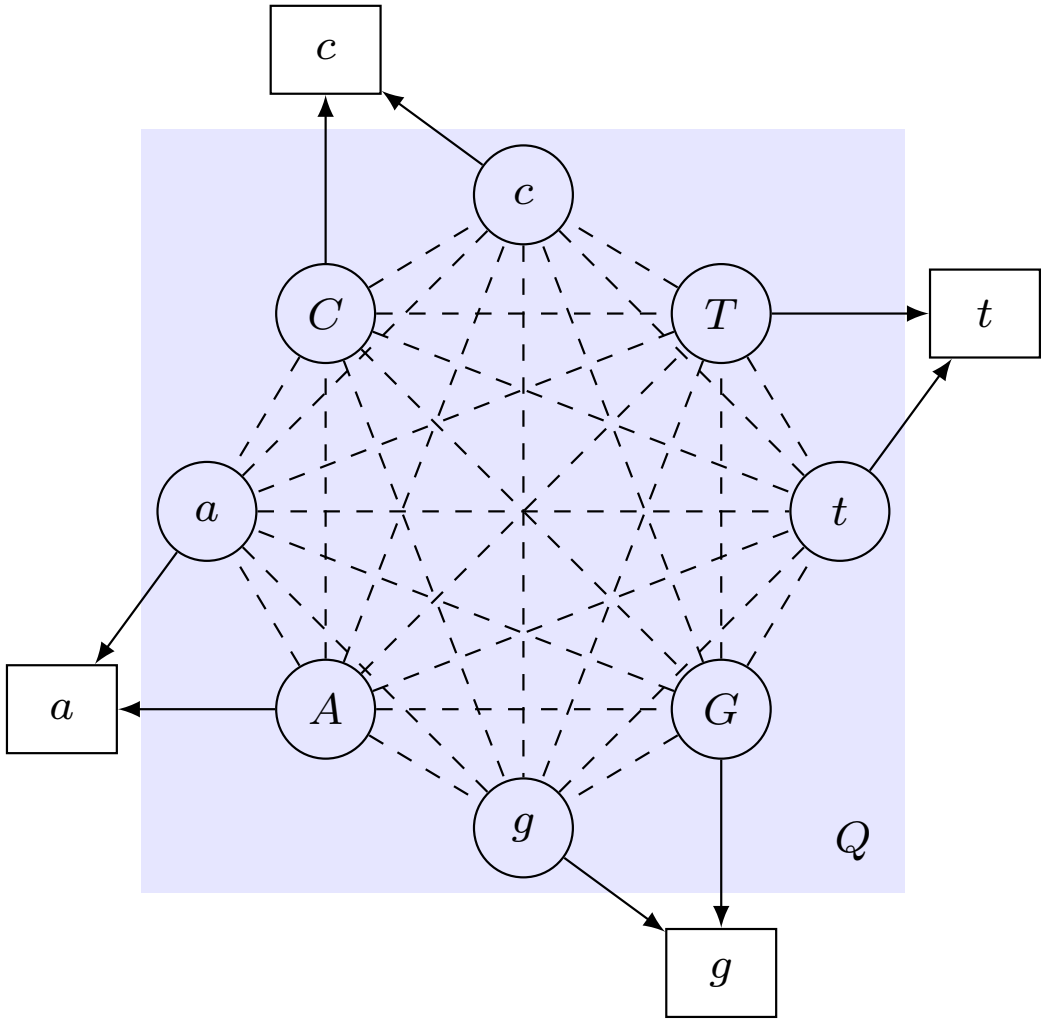
Скрытые состояния модели должны содержать информацию о:

1. наблюдаемом состоянии;
2. вхождении наблюдаемого состояния в структуру.

$$q_i = (s_i, h_i); \quad P(q_i \mid q_1 \dots q_{i-1}, \bar{s}) = P(q_i, s_i \mid q_{i-l} \dots q_{i-1}).$$

Множество скрытых состояний: $Q = S \times H$.

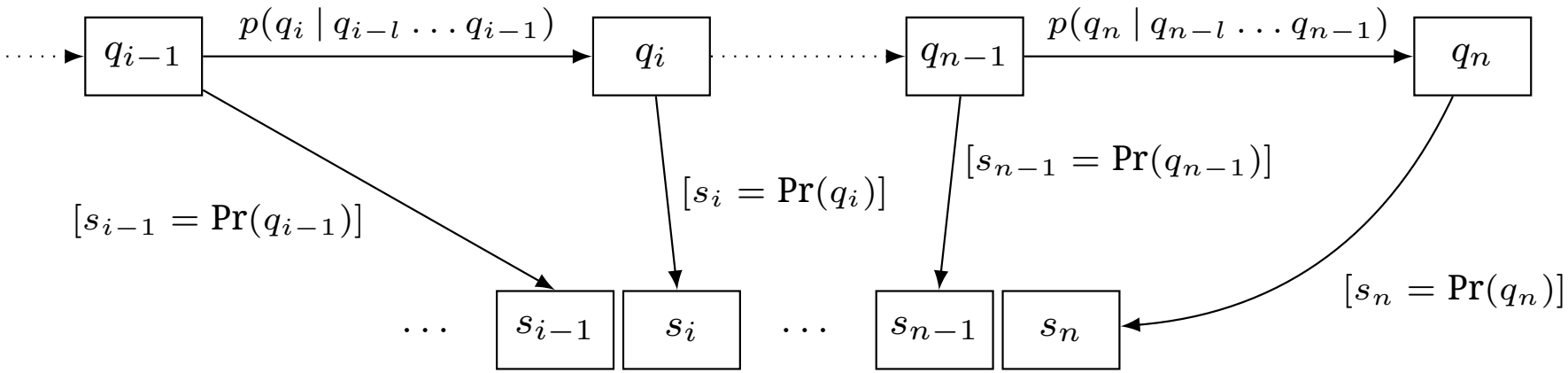
Модель $\mathcal{M}(l)$ для генов



Скрытые состояния: A, C, G, T — нуклеотиды в экзонах, a, c, g, t — в интронах.

Зависимости между состояниями — модель $\mathcal{M}(l)$

Функция проекции: $\Pr : Q^* \to S^*, \quad \Pr((s_1, h_1)(s_2, h_2) \dots (s_n, h_n)) = s_1 s_2 \dots s_n.$



Совместная вероятность:

$$P(\bar{s}, \bar{q}) = P(\bar{q})[\Pr(\bar{q}) = \bar{s}];$$

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

Параметры модели

Параметры модели $\mathcal{M}(l)$:

- начальные вероятности $\{\pi(x) \,|\, x \in Q^l\}$.
- переходные вероятности $\{p(y \,|\, x) \,|\, x \in Q^l, y \in Q\}$.

Обучающая выборка: $\{(\bar{s}_j, \bar{h}_j)\}_{j=1}^N \sim \{\bar{q}_j\}_{j=1}^N \subset Q^*$.

Оценка параметров: $\pi^*(x) = N_{st}(x)/N$; $p^*(y \,|\, x) = N(xy)/N(x)$.

x	$N(x)$	x	$N(x)$
ttagA	2611	ttaga	250037
ttagC	1277	ttagc	220061
ttagG	5388	ttagg	206132
ttagT	1267	ttagt	267565
		Всего	954338

$$p^*(C \,|\, ttag) = 1277/954338 = 0{,}0013;$$
$$p^*(g \,|\, ttag) = 206132/954338 = 0{,}2160.$$

Максимизация правдоподобия

Задача: $P(\bar{s}, \bar{q}) = P(\bar{q})[\Pr(\bar{q}) = \bar{s}] \longrightarrow \max_{\bar{q}}$.

Функция $F : \mathbb{N} \times Q^l \to \mathbb{R}$**:**

$$F(i, x) = \max_{q_1 \dots q_i} \log P(s_1 \dots s_i, q_1 \dots q_i \mid q_{i-l+1} \dots q_i = x).$$

- F определяет решение задачи максимизации:

$$\max \log P(\bar{s}, \bar{q}) = \max_{x \in Q^l} F(n, x);$$

- для вычисления F существует рекуррентная формула:

$$F(i, x) = \max_{y \in Q} \big(F(i-1, yx_1 \dots x_{l-1}) + \log p(x_l \mid yx_1 \dots x_{l-1}) + \log[x_l = s_i]\big).$$

Граничные равенства: $F(l, x) = \log \pi(x) + \log[x = s_1 \dots s_l]$.

Вычислительная сложность алгоритма: $\mathcal{O}(|\bar{s}| \cdot |H|^{l+1})$.

Пример вычисления функции F

Модель $\mathcal{M}(1)$, $S = \{A, C, G, T\}$, $Q = \{A, C, G, T, a, c, g, t\}$.

Позиция i	1	2	3	$\cdots$
Наблюдаемое состояние s_i	A	T	G	$\cdots$
Скрытое состояние q_i	A или a	T или t	G или g	$\cdots$

$$F(1, A) = \log \pi(A); \quad F(1, a) = \log \pi(a);$$

$$F(2, T) = \max\{F(1, A) + \log p(T \mid A), \quad F(2, t) = \max\{F(1, A) + \log p(t \mid A),$$
$$F(1, a) + \log p(T \mid a)\}; \qquad F(1, a) + \log p(t \mid a)\};$$

$$F(3, G) = \max\{F(2, T) + \log p(G \mid T), \quad F(3, g) = \max\{F(2, T) + \log p(g \mid T),$$
$$F(2, t) + \log p(G \mid t)\}. \qquad F(2, t) + \log p(g \mid t)\}.$$

$$\vdots$$

Вычислительный эксперимент

Метрики качества:

- качество распознавания *отдельных* скрытых состояний: символьная специфичность, символьная чувствительность, коэффициент корреляции, средняя условная вероятность, доля правильно распознанных состояний;
- качество распознавания *границ* между состояниями: фрагментная специфичность, фрагментная чувствительность.

Выборки:

- Национальный центр биотехнологической информации США (`http://ncbi.nlm.nih.gov/`);
- Европейский центр молекулярной и биоинформации (`ftp://ftp.cmbi.ru.nl/`);

Методика: кросс-валидация.

Вычислительный эксперимент

Результаты эксперимента:

- Оптимальный порядок модели $l \in \{6, 7\}$ (для генов), $l \in \{4, 5\}$ (для белков).
- Для геномов простых организмов (растения, насекомые) качество распознавания соответствует алгоритмам на основе ОСММ ($\sim$90 % для символьных метрик, 50–75 % для фрагментных).
- Для белков качество распознавание соответствует современным алгоритмам (75–80 % для символьных метрик, 30–40 % для фрагментных).
- Модель, обученную на одном виде, можно использовать для распознавания генов родственных видов.
- Для геномов сложных организмов (птицы, млекопитающие) качество распознавания низкое.

Композиции моделей

Наблюдение

Свойства гена зависят от концентрации в нем нуклеотидов цитозина (C) и гуанина (G). Во многих алгоритмах распознавания применяются отдельные модели для разных диапазонов концентраций.

Гипотеза

Строки наблюдаемых состояний порождаются смесью вероятностных моделей $\mathcal{M}_1$, $\mathcal{M}_2$, ..., $\mathcal{M}_k$:

$$P(\bar{q}) = \sum_{j=1}^{k} w_j(\bar{q}) P(\bar{q} \,|\, \mathcal{M}_j), \quad w_j(\bar{q}) \geqslant 0, \sum_{j=1}^{k} w_j(\bar{q}) = 1.$$

$w_j : Q^* \to [0, 1]$ — функция компетентности модели $\mathcal{M}_j$.

Определение функций компетентности

Способы определения $w_j(\bar{q})$:

- $w_j(\bar{q}) \equiv w_j = \text{const}(\bar{q})$
  Смесь моделей. $\{w_j\}_{j=1}^k$ и параметры $\mathcal{M}_1$, ..., $\mathcal{M}_k$ определяются с помощью EM-алгоритма.

- $w_j(\bar{q}) = [\Pr(\bar{q}) \in G_j]$
  Каждая строка порождается одной моделью. Выбор модели зависит от строки наблюдаемых состояний $\bar{s}$.

$G_j \subset S^*$ — *область компетентности* модели $\mathcal{M}_j$.

Алгоритмическая композиция

Совместная вероятность:

$$P(\bar{s}, \bar{q}) = P(\bar{q} \,|\, \mathcal{M}_j)[\Pr(\bar{q}) = \bar{s}], \quad j = \arg[\bar{s} \in G_j].$$

Алгоритмическая композиция:

$$\mathcal{A}_c(\bar{s}) = \begin{cases} \mathcal{A}_1(\bar{s}), & \text{если } \bar{s} \in G_1, \\ \vdots & \\ \mathcal{A}_k(\bar{s}), & \text{если } \bar{s} \in G_k; \end{cases}$$

$\mathcal{A}_j$ — алгоритм распознавания на основе $\mathcal{M}_j$.

Это — частный вид общей алгоритмической композиции [Журавлёв]:

$$\mathcal{A}_c(\bar{s}) = C(F(\mathcal{A}_1(\bar{s}), \mathcal{A}_2(\bar{s}), \ldots, \mathcal{A}_k(\bar{s}))),$$

F — корректирующая операция, C — решающее правило.

Обучение параметров композиции

Параметры композиции:

- области компетентности $G_1, \ldots, G_k$;
- параметры моделей $\mathcal{M}_1, \ldots, \mathcal{M}_k$.

Обучающая выборка $T = \{\bar{q}_i\}_{i=1}^N \subset Q^*$.

$$\log P(T) = \sum_{i=1}^{N} \log\Big(\sum_{j=1}^{k} w_j(\bar{q}_i) P(\bar{q}_i \,|\, \mathcal{M}_j)\Big) =$$

$$= \sum_{\bar{q} \in T_1} \log P(\bar{q} \,|\, \mathcal{M}_1) + \cdots + \sum_{\bar{q} \in T_k} \log P(\bar{q} \,|\, \mathcal{M}_k) \longrightarrow \max_{\mathcal{M}_1, \ldots, \mathcal{M}_k};$$

Утверждение

Параметры составляющей модели $\mathcal{M}_j$, $1 \leqslant j \leqslant k$ находятся с помощью обучения на выборке

$$T_j = \{\bar{q} \in T \,|\, \Pr(\bar{q}) \in G_j\}.$$

Определение оптимального разбиения

Предложение

Разбиение на области компетентности $\{G_j\}_{j=1}^k$ производится с помощью последовательности предикатов, зависящих от наблюдаемых состояний.

$$I : S^* \to \mathbb{R}^\alpha \to \{0,1\}, \mathbb{R}^\alpha \text{ — пространство признаков.}$$

Предикаты на основе концентраций состояний:

$$I(\bar{s}; X, \omega) = \Big[\sum_{x \in X} n(\bar{s}, x) \geqslant \omega\Big], \quad X \subset S, \omega \in (0,1).$$

1. вычислить суммарную концентрацию состояний из X в строке $\bar{s}$;
2. сравнить концентрацию с порогом ω.

Пример областей компетентности

Предикат: $I(\bar{s}) = [n(\bar{s}, \{C, G\}) \geqslant 0{,}492]$.

Области компетентности:

$$G_1 = [n(C) + n(G) < 0{,}492], \quad G_2 = S^* \setminus G_1 = [n(C) + n(G) \geqslant 0{,}492].$$

Ген	Содержание нуклеотидов, %					Часть разбиения
	$n(A)$	$n(C)$	$n(G)$	$n(T)$	$n(\{C, G\})$	
s_1	26,18	17,52	21,01	35,30	38,53	G_1
s_2	21,24	27,77	25,37	25,63	53,14	G_2
s_3	23,09	23,75	19,06	34,10	42,81	G_1
s_4	19,59	29,24	30,69	20,48	59,93	G_2
s_5	20,98	25,77	20,23	33,01	46,00	G_1
s_6	32,16	16,34	17,68	33,81	34,02	G_1

Алгоритм построения разбиения

$$\log P(T) \longrightarrow \max_{G_1,\ldots,G_k} \sim \max_{T_1,\ldots,T_k}.$$

1. Начальное разбиение состоит из всей выборки: $T_1 := T$, $k := 1$.
2. Определить *оптимальный* предикат:

$$(T^*, I^*) = \arg\max_{T_j, I} \Delta(T_j, I), \quad j = 1, \ldots, k.$$

3. Разбить часть выборки T^* на две согласно предикату I^*:

$$k := k + 1;$$
$$T_k := \{\bar{q} \in T^* \mid I(\mathrm{Pr}(\bar{q}))\};$$
$$T^* := \{\bar{q} \in T^* \mid \neg I(\mathrm{Pr}(\bar{q}))\}.$$

4. Если не выполнен критерий останова, вернуться к шагу 2.

Функционал качества предиката

Вычисление качества предиката I на выборке T:

1. $T^+ = \{\bar{q} \in T \mid I(\Pr(\bar{q}))\}, T^- = T \setminus T^+$;

   (части, на которые предикат делит выборку);

2. $\mathcal{M}[T] = \arg\max_{\mathcal{M}} P(T \mid \mathcal{M}), \mathcal{M}[T^+] = \ldots, \mathcal{M}[T^-] = \ldots$

   (модели, обученные на выборке и ее частях);

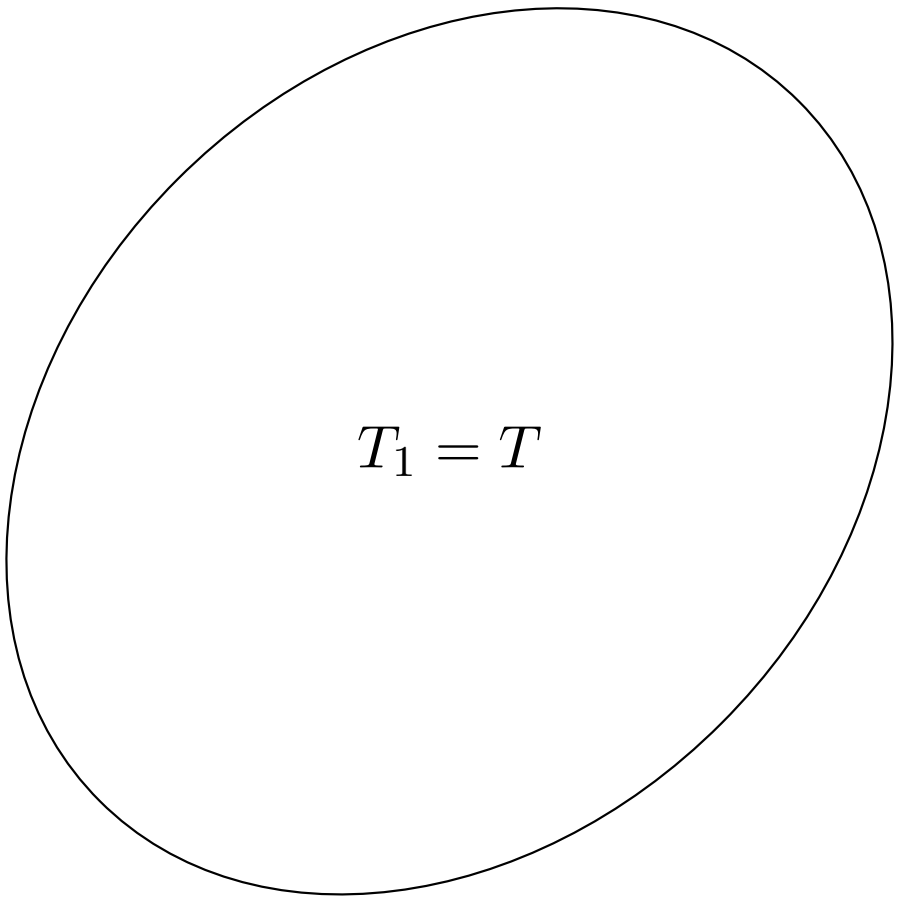
3. $\Delta(T, I) = \log P(T^+ \mid \mathcal{M}[T^+]) + \log P(T^- \mid \mathcal{M}[T^-]) - \log P(T \mid \mathcal{M}[T])$

   (функционал качества).

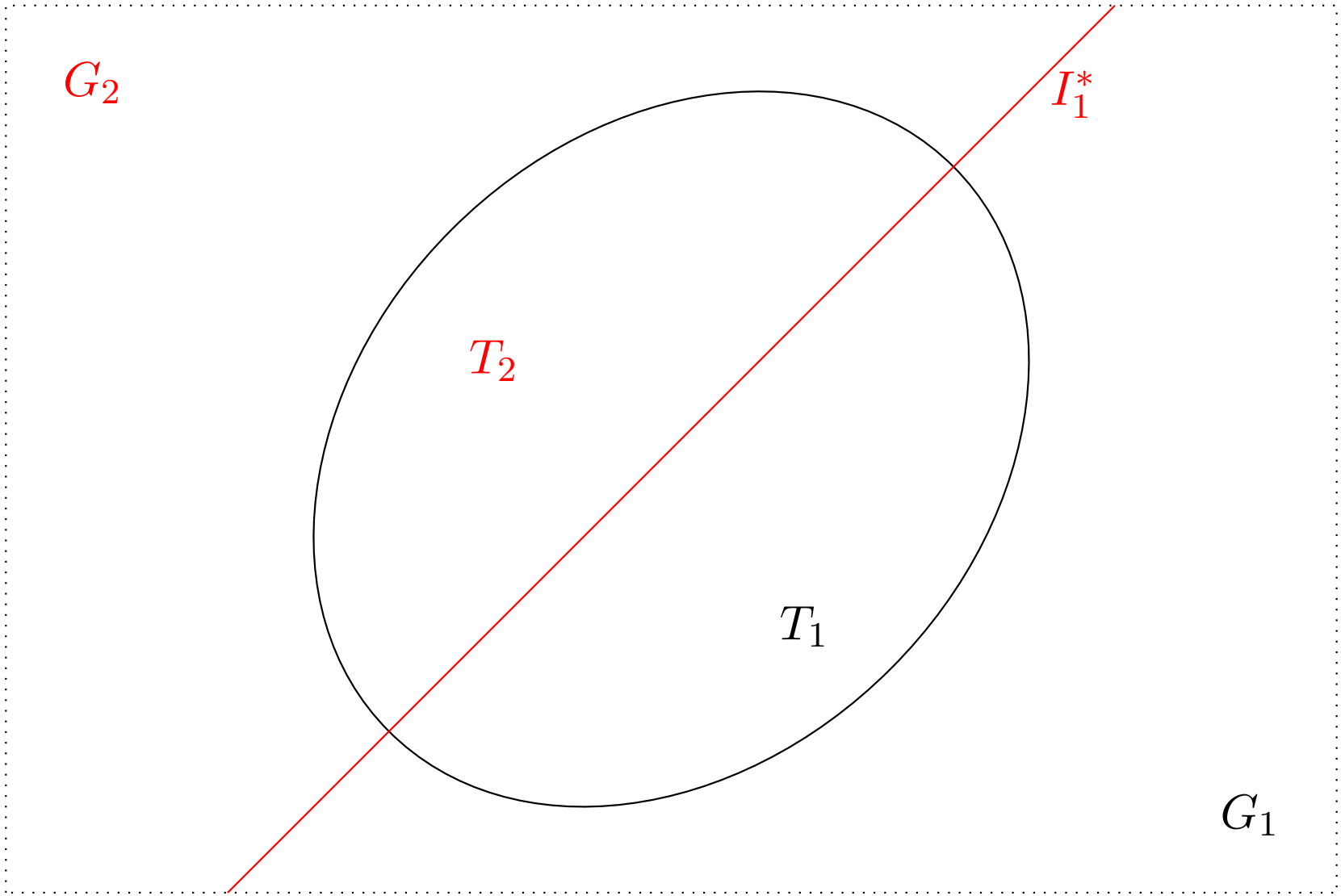
Наблюдение

Если просуммировать $\Delta(T^*, I^*)$, полученные на каждой итерации алгоритма построения разбиения, получим $\log P(T \mid T_1, \ldots, T_k)$ с точностью до постоянной.

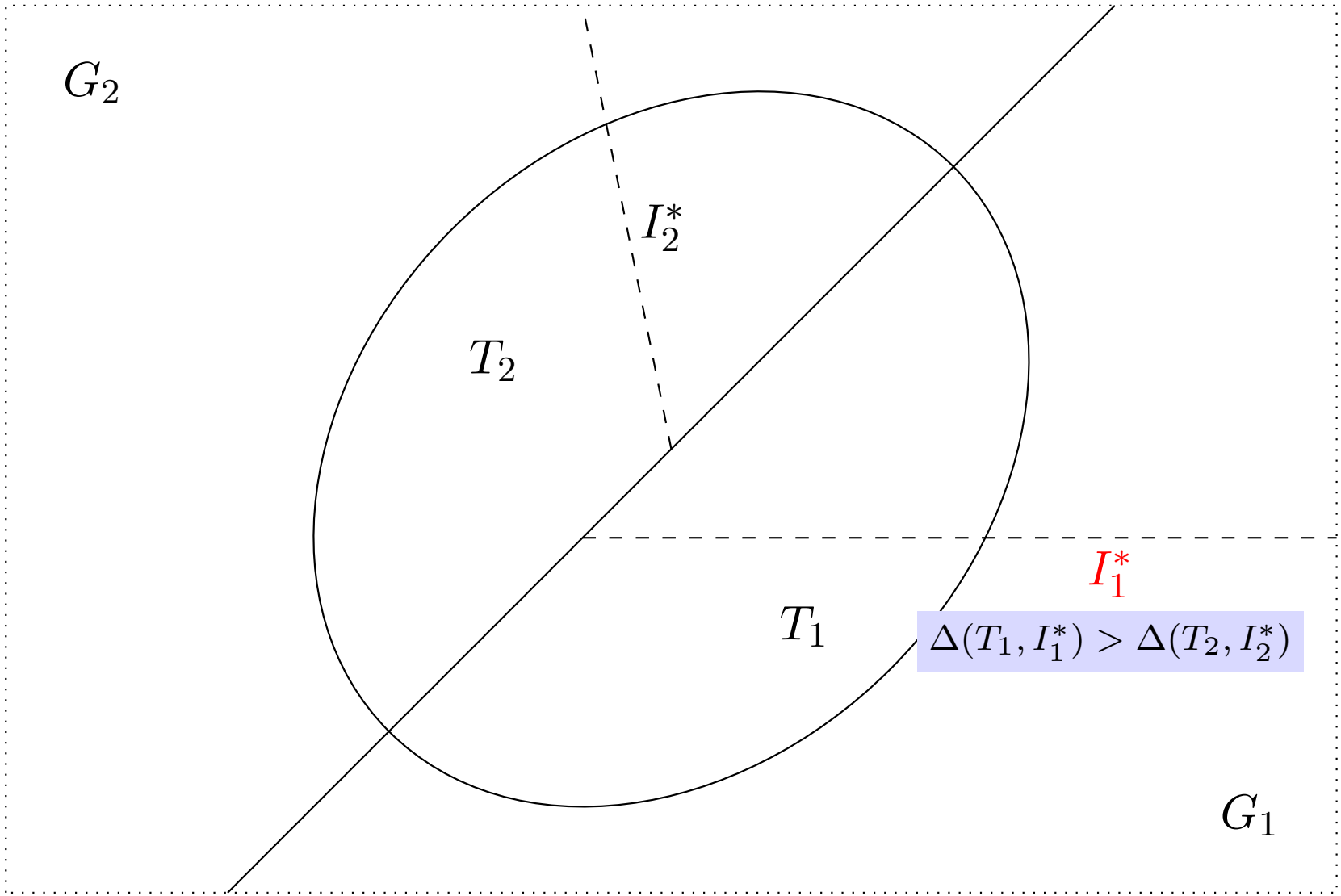
Пример построения разбиения



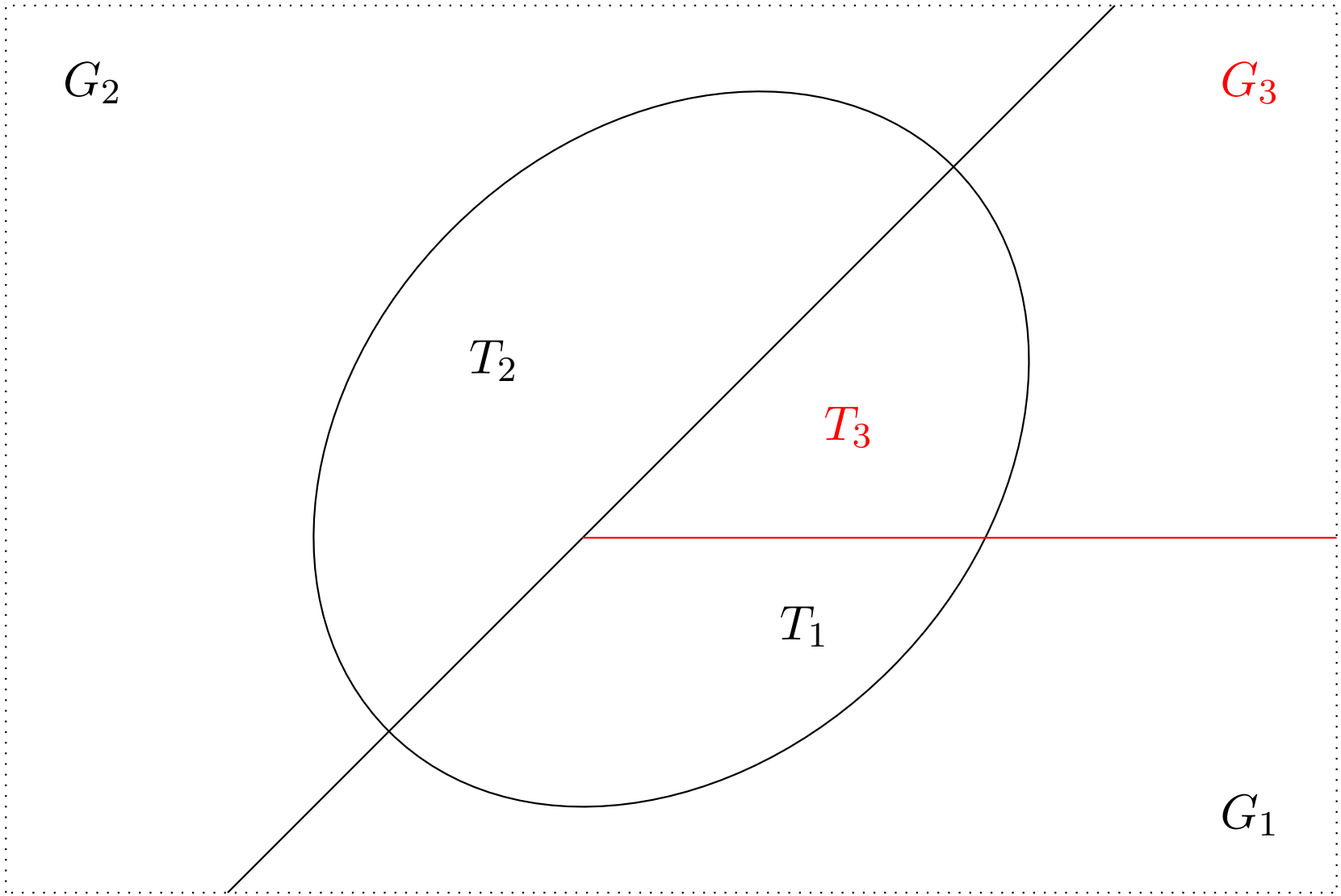
Пример построения разбиения



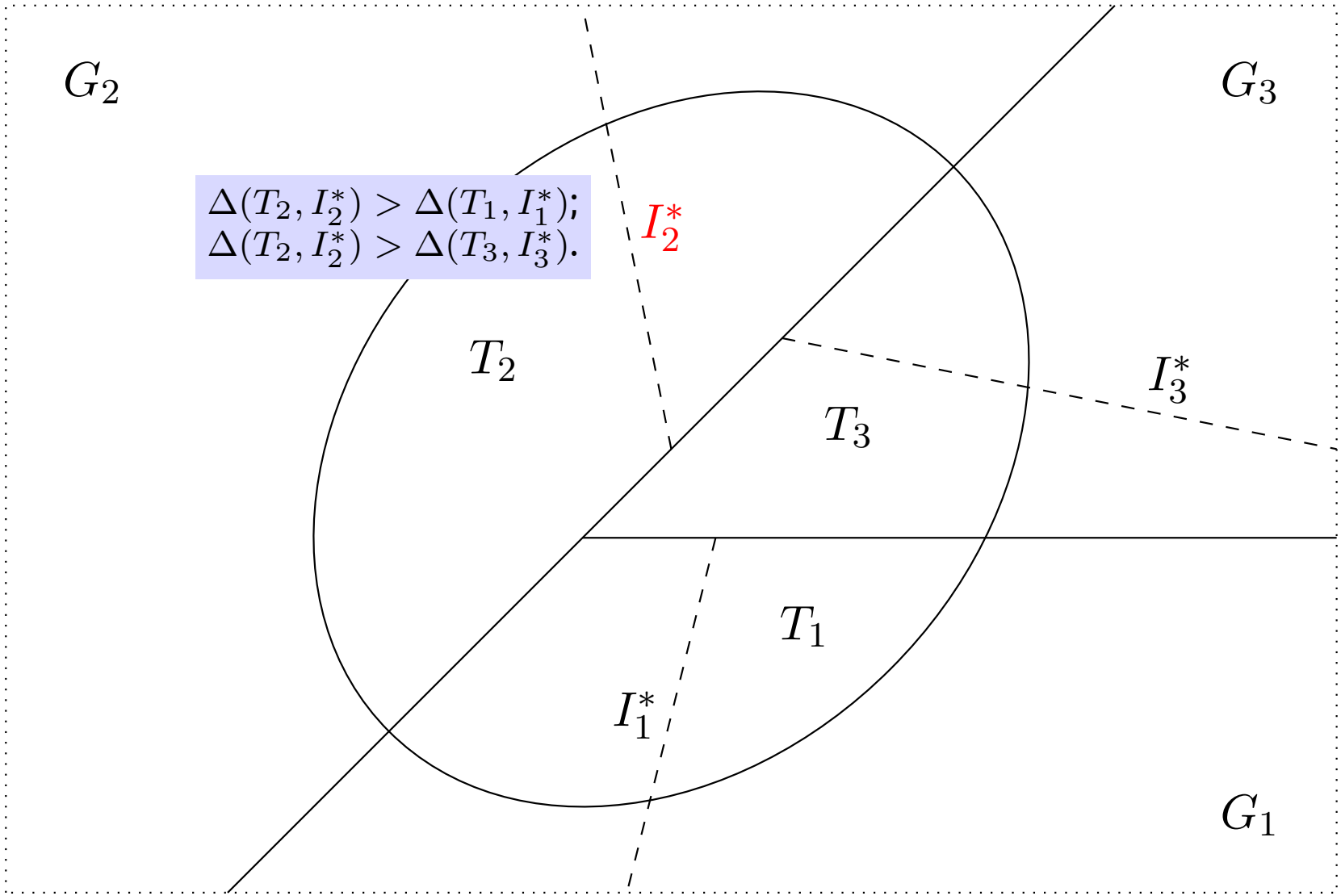
Пример построения разбиения



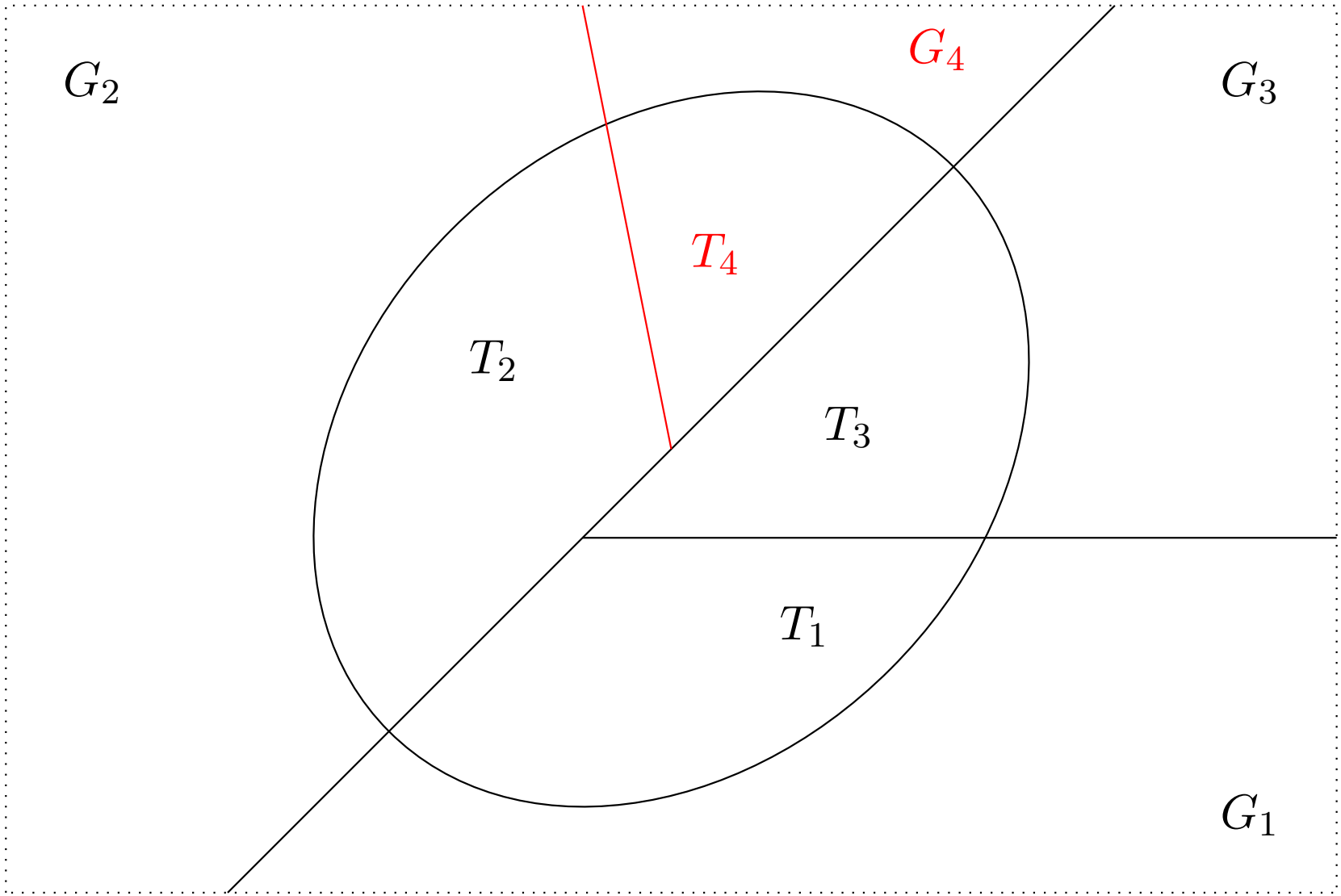
Пример построения разбиения



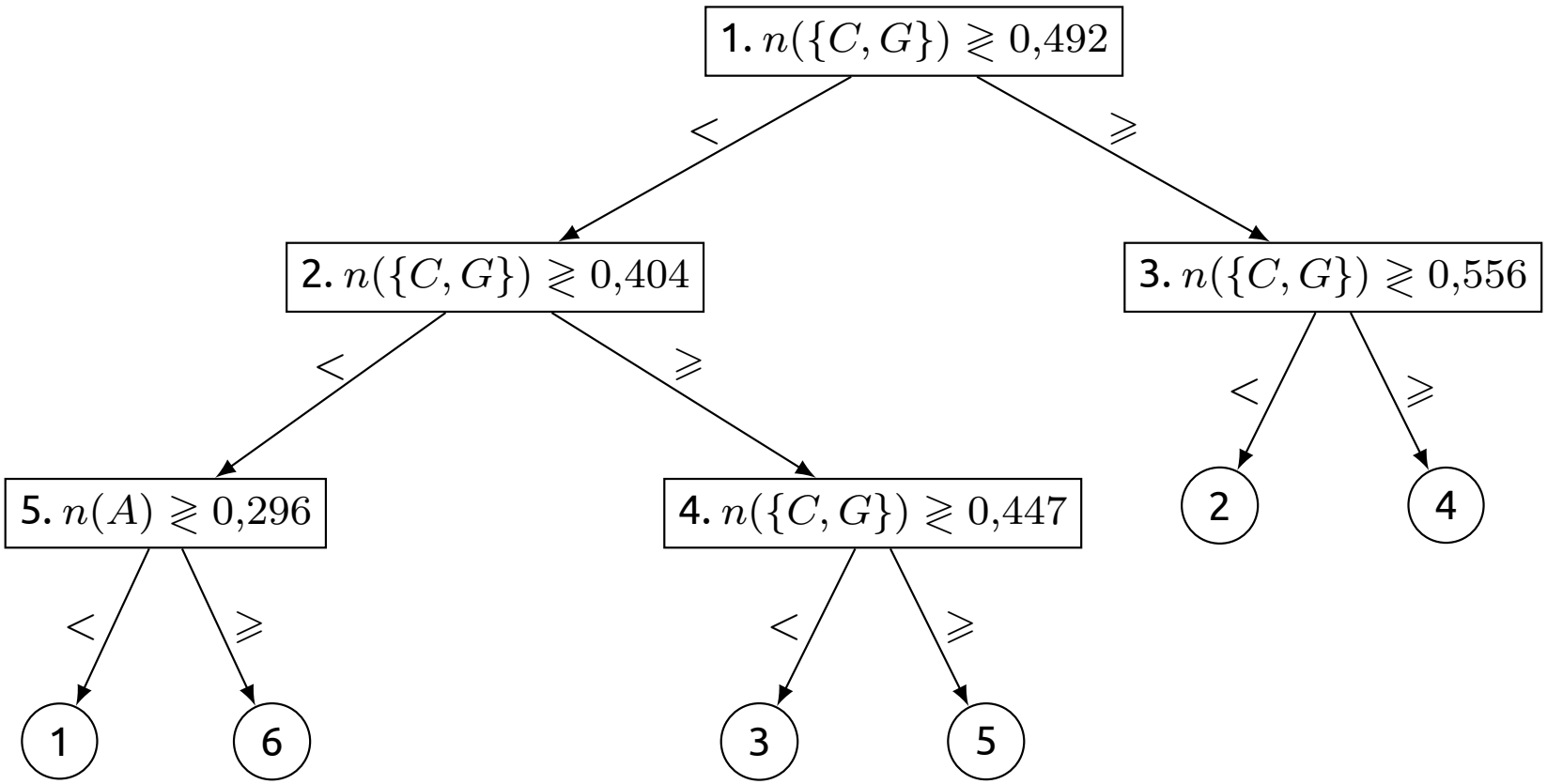
Пример построения разбиения



Пример построения разбиения



Пример разбиения



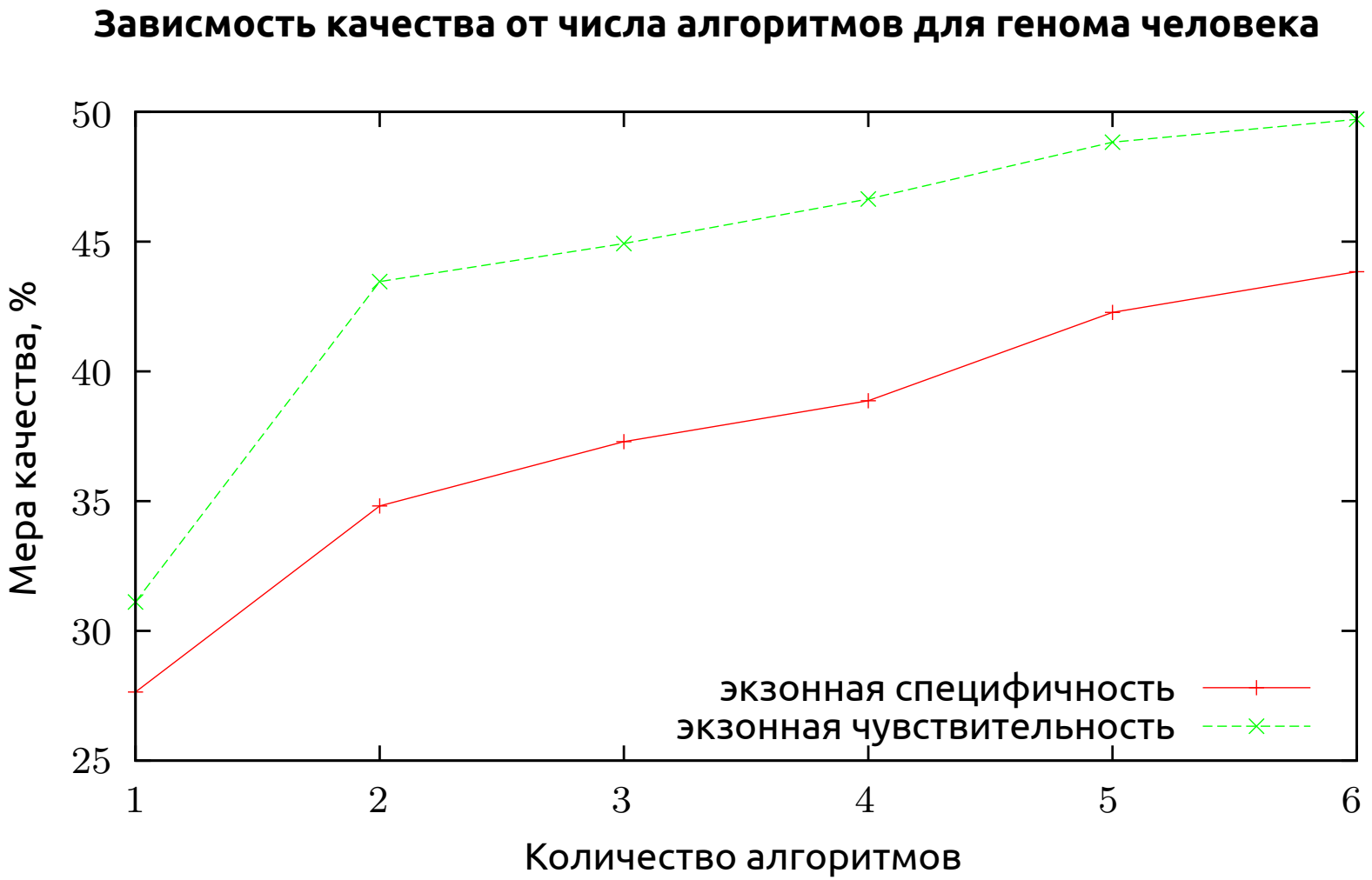
Дерево разбиения для генома человека

Пример разбиения

Разбиение для генома человека:

Часть разбиения	Описание	Кол-во генов
G_1	$n(\{C,G\}) < 0{,}404; n(A) < 0{,}296$	1727
G_2	$0{,}492 \leqslant n(\{C,G\}) < 0{,}556$	5549
G_3	$0{,}404 \leqslant n(\{C,G\}) < 0{,}447$	3731
G_4	$n(\{C,G\}) > 0{,}556$	5525
G_5	$0{,}447 \leqslant n(\{C,G\}) < 0{,}492$	3641
G_6	$n(\{C,G\}) < 0{,}404; n(A) \geqslant 0{,}296$	2077

Качество композиций



Качество композиций

Вид	k	Мера качества, %					
		отдельные состояния				границы	
		SSp	SSn	CC	ACP	ESp	ESn
Homo sapiens	1	35,58	89,56	49,48	76,45	27,64	31,11
(человек)	6	54,43	87,53	64,71	83,02	43,84	49,71
Sus scrofa	1	33,24	85,66	47,60	75,75	24,87	26,64
(свинья)	4	47,54	81,02	58,04	79,83	36,21	40,86
Mus musculus	1	59,97	85,32	67,20	83,97	42,22	40,39
(мышь)	4	71,96	83,03	74,08	87,11	50,90	47,73
Rattus norvegicus	1	61,73	83,59	67,47	84,01	40,75	36,49
(крыса)	6	76,23	75,32	72,22	86,11	47,34	38,78
Papio anubis	1	39,91	86,65	52,28	77,45	30,41	31,42
(павиан)	6	65,00	79,75	68,56	84,42	50,54	51,18

Итоги вычислительного эксперимента

- Качество при использовании композиций повышается на 10–15 %.
- Оптимальное количество алгоритмов в композиции — $k \in \{4, 5, 6\}$.
- Большое количество предикатов на основе концентрации $n(\{C, G\})$.
- Близким видам соответствуют похожие деревья предикатов.
- Для задачи распознавания вторичной структуры белков композиции неэффективны.

Перспективы

Направления дальнейших исследований:

- Более сложный вид функций компетентности (напр., $w_i(\bar{s}) = \sigma(\cdot)$).
- Поиск последовательности наиболее вероятных скрытых состояний (другие функции потерь).
- Применение скрытого размещения Дирихле для кластеризации последовательностей и фрагментов.

Спасибо за внимание!

Сайт: http://sestudy.edu-ua.net/ru/bio

E-mail: ostrovski.alex@gmail.com