

НММ. Витерби для поиска СрG остановок.

Алгоритмы в биоинформатике

Антон Елисеев

eliseevantoncoo@gmail.com

Что было на прошлой лекции?

- Локальное выравнивание.
- Эффективный по памяти ($O(n)$) алгоритм выравнивания.

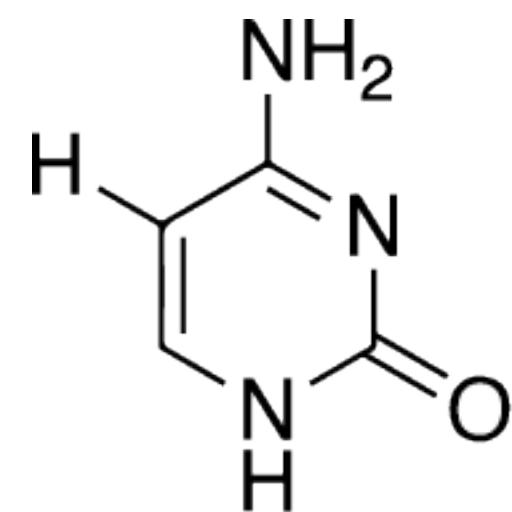
Что будет на этой лекции?

- Обсудим задачи разметки в биоинформатике.
- Научимся использовать скрытые марковские модели для симуляции размеченных последовательностей.
- Научимся находить наиболее правдоподобную последовательность скрытых состояний.

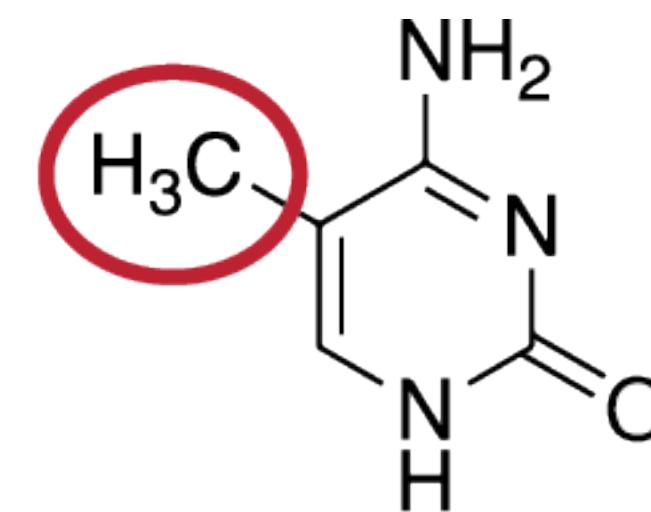
CpG островки.



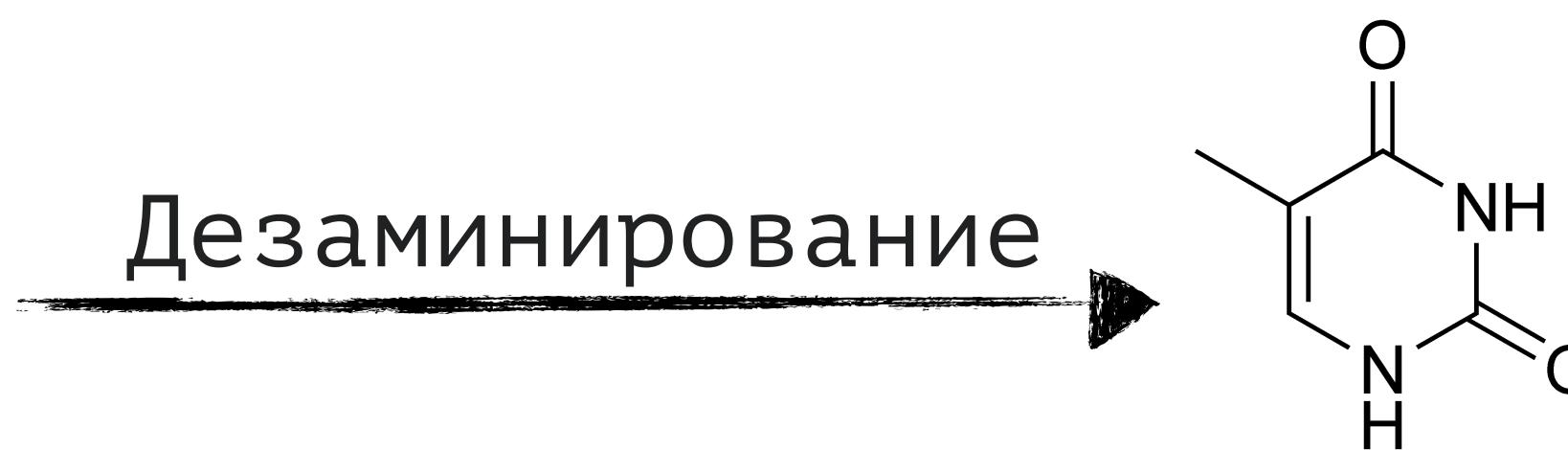
CpG островки.



Цитозин



Метилированный
цитозин



Тимин

$$P(CG) < P(C)P(G)$$

CpG островки.

- В геноме млекопитающих около 70% CpG метилировано
- В промоторных областях метилирование подавляется
- Маркер онтогенеза (инактивация генов-супрессоров опухолевого роста)

CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

- Где CpG островки в последовательности?

CTTCATGTGAAAGCAGACGTAAGTCA

CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

- Где CpG островки в последовательности?

CTTCATGTGAAAGCAGACGTAAAGTCA

-----+-----

CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

- Где CpG островки в последовательности?
- С какой вероятностью некоторая позиция принадлежит CpG островку?

CTTCATGTGAAAGCAGACGTAAAGTCA

-----+-----

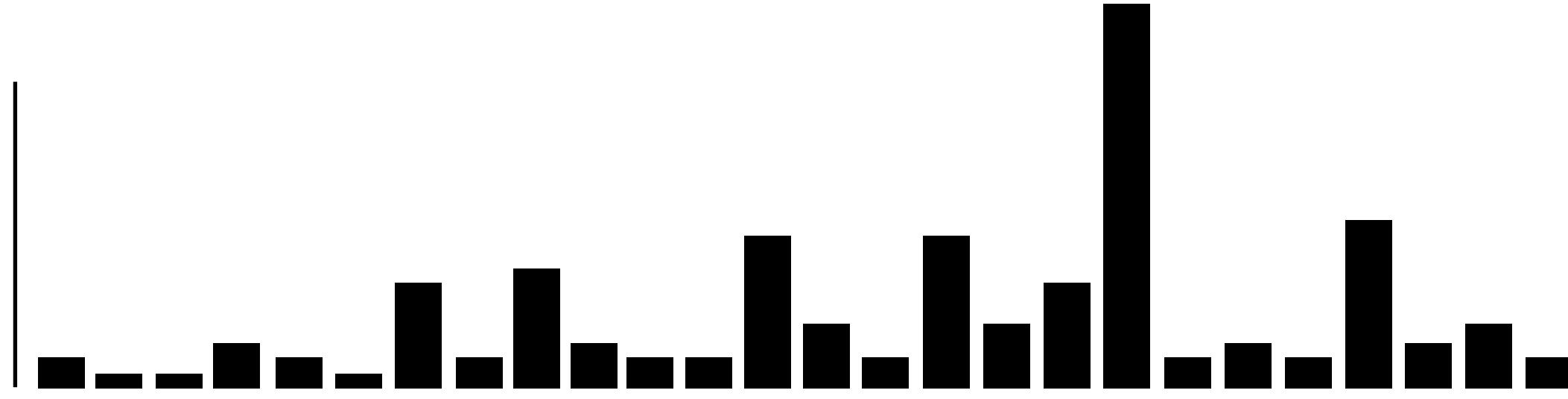
CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

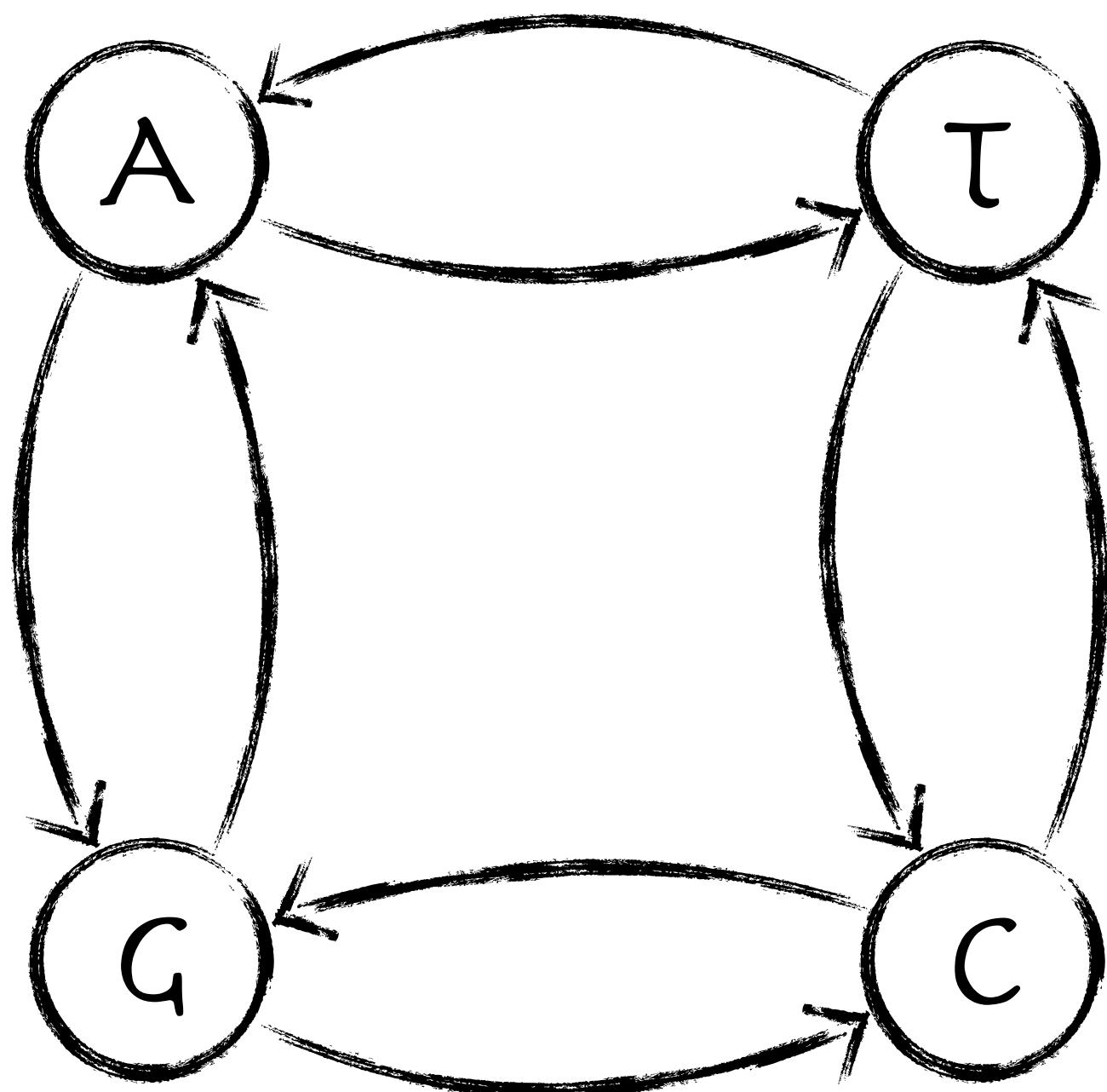
- Где CpG островки в последовательности?
- С какой вероятностью некоторая позиция принадлежит CpG островку?

CTTCATGTGAAAGCAGACGTAAGTCA

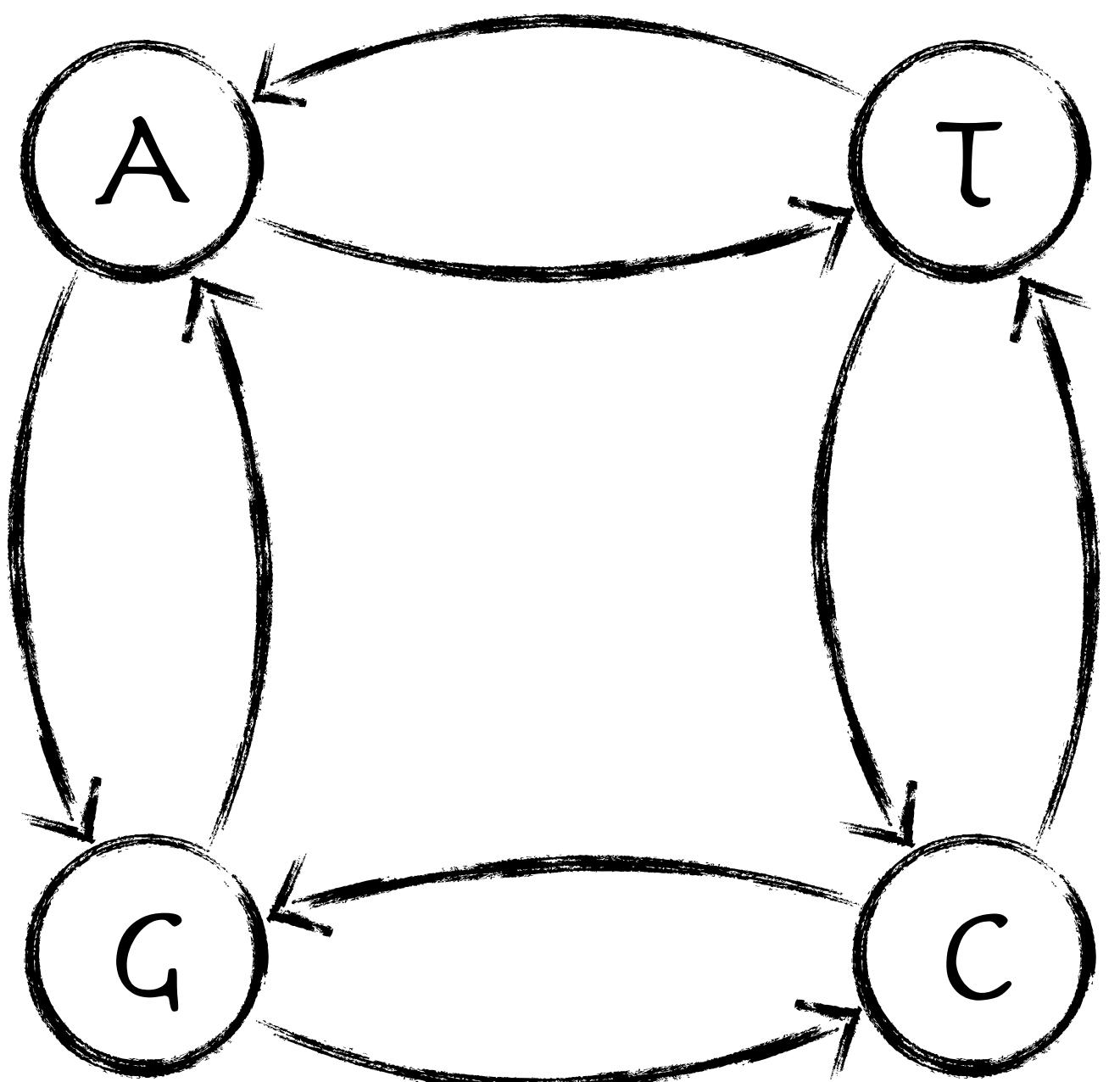
-----+-----



СрG островки. Вероятностная модель.



CpG островки. Вероятностная модель.



A	C	G	T
A			
C			
G			
T			

CpG островки. Веса модели.

Участки из CpG

.....GAATTCTTCTGTCTAGTTTATAGGAAGATGTTCC
TTTCAGCGTATGCATCAAAGAGCTCCAAGTTCCACTACAGAGTC
TTCAAAAAGAATGTTCAAAACTGCTCTATGAAAAGGAATGTTCAC
CTCTGTGAGTAGAATGCAAGCATCACAAAAAGTTCTGGGAATGC
TTCTGTCTAGTTTATGTGAAGACATTCCCGTTCCAACGAAAGC
CTAAAAAGCTATCCAATATCCACTTGCAGATTCTACAAAAAGAGTG
TTTCAAAACTGCAGTATCAACAGAAAGGTTCAACTCTGTGAGCTGA
GTACACACATCACAGAGAAGTTCTGGGAATGCTCTGTCTAGTT
TTATGTGAAGATATTCCCTTTTCAGCATAGGCCTCAATGGGTTCC
AAATGTCCTTTCCAGGTACTACAAAAAGAGTGTTCAAAAGCT
CTATGAAAGGGAATGTTCAAACCTCTGTGAGTTGAATGCAAACATCAT
GAAGAAGTTCTGAGAATACTCTGACTAGTTTATGTGAAGATA
TTCCCATTCCAATGAAAGCCTCAAAGCTGCCAAATATTCCCTTG
CAGATCCTACAAAGAGAGTGTTCAAAACTACTCTAAAAAGAAA
TGTTCAAACCTGTGAGTTGAGTACACATATCACAAAGAAGTTCTT
AGCATGTTCTGTCCTGTTTATTGTAGATCTTCCGGTTCCCG
TGAAGGCCTCAAAGCTGTCCAA.....

$$P(ab) = \frac{\#ab}{\sum_c \#ac}$$

CpG островки. Веса модели.

Участки из CpG

.....GAATTCTTCTGTCTAGTTTTATAGGAAGATGTTCCCTT
TTTCAGCGTATGCATCAAAGAGCTCCAAGTTCCACTACAGAGTC
TTCAAAAAGAATGTTCAAAACTGCTCTATGAAAAGGAATGTTCAC
CTCTGTGAGTAGAATGCAAGCATCACAAAAAGTTCTGGGAATGC
TTCTGTCTAGTTTTATGTGAAGACATTCCCGTTCCAACGAAAGC
CTAAAAAGCTATCCAAATATCCACTTGCAGATTCTACAAAAAGAGTG
TTTCAAAAACGTATCAACAGAAAGGTTCAACTCTGTGAGCTGA
GTACACACATCACAGAGAAGTTCTGGGAATGCTCTGTCTAGTT
TTATGTGAAGATATTCCCTTTTCAGCATAGGCCTCAATGGGTTCC
AAATGTCCTTTCCAGGTACTACAAAAAGAGTGTTCACAAACTGCT
CTATGAAAGGGAATGTTCAAACCTGTGAGTTGAATGCAAACATCAT
GAAGAAGTTCTGAGAATACTCTGACTAGTTTATGTGAAGATA
TTCCCATTTCCAATGAAAGCCTCAAAGCTGTCCAAATATTCCCTTG
CAGATCCTACAAAGAGAGTGTTCACAAACTACTCTAAAAAGAAA
TGTTCAAACCTGTGAGTTGAGTACACATATCACAAAGAGTTCTT
AGCATGTTCTGTCCTGTTTATTGTAGATCTTCCGGTTCCCG
TGAAGGCCTCAAAGCTGTCCAA.....

	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.339	.375	.125
T	.079	.355	.384	.182

CpG островки. Веса модели.

	A	C	G	τ
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.339	.375	.125
τ	.079	.355	.384	.182

CpG островки

	A	C	G	τ
A	.300	.205	.285	.210
C	.322	.298	.078	0.302
G	.248	.246	.298	.208
τ	.177	.239	.292	.292

Остальные участки генома

CpG островки. Веса модели.

	A	C	G	τ
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.339	.375	.125
τ	.079	.355	.384	.182

CpG островки

	A	C	G	τ
A	.300	.205	.285	.210
C	.322	.298	.078	.302
G	.248	.246	.298	.208
τ	.177	.239	.292	.292

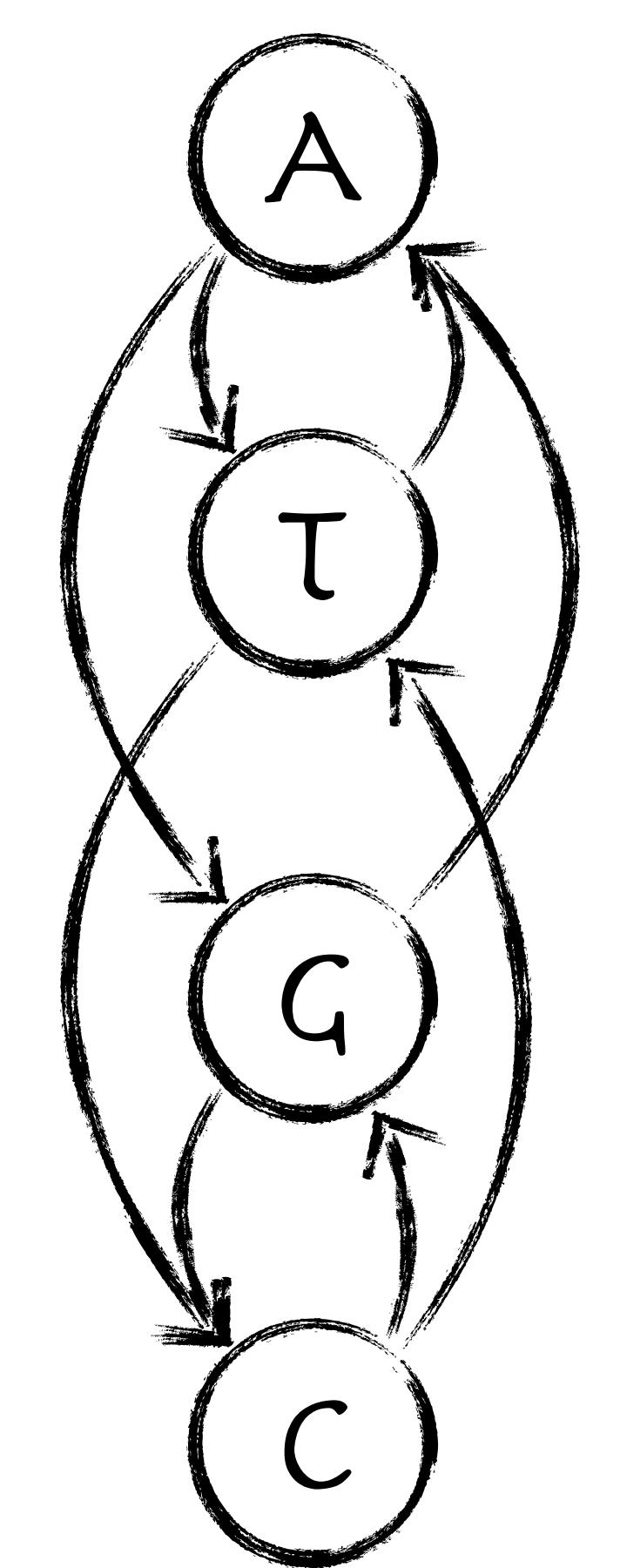
Остальные участки генома

CpG островки.

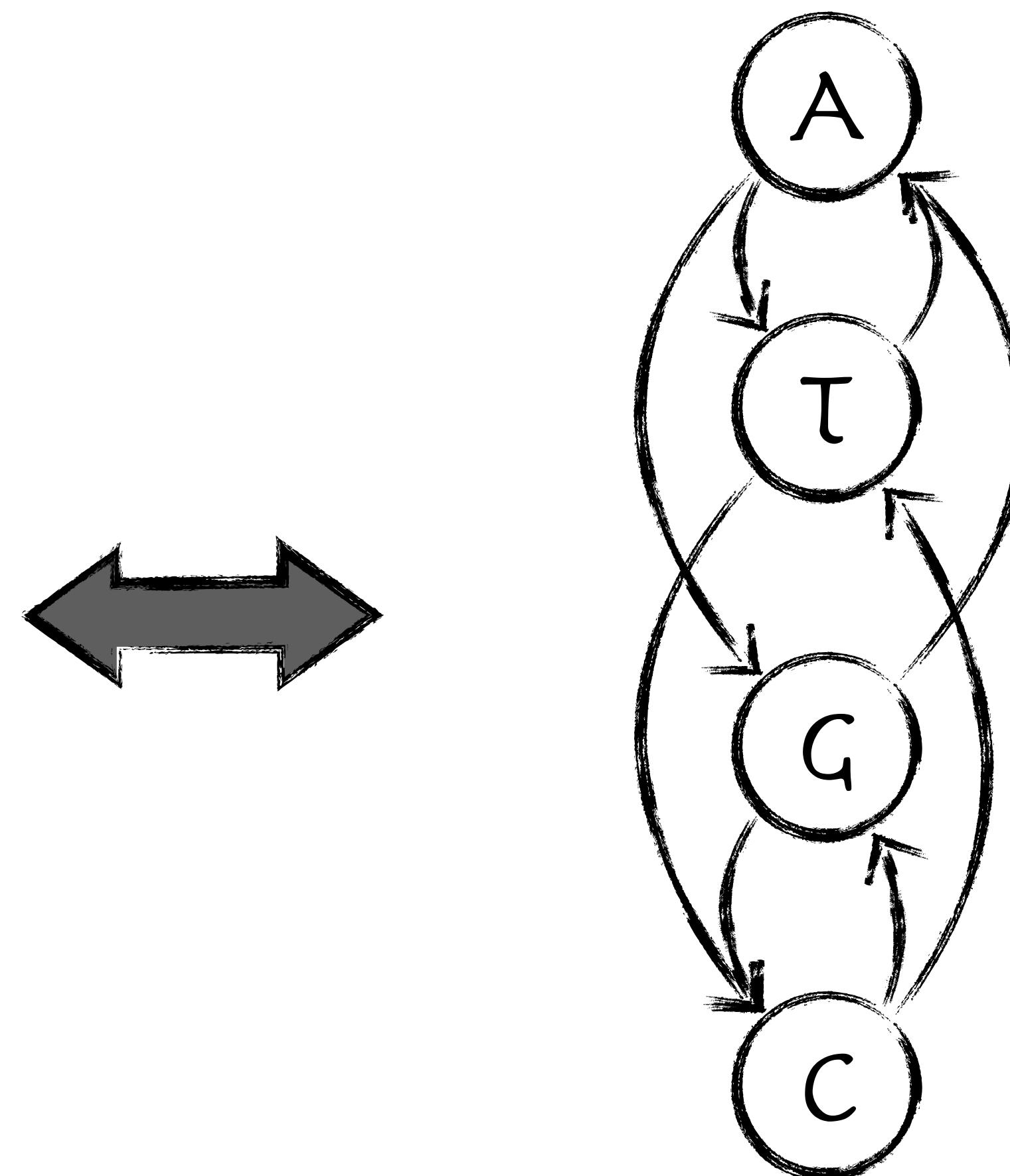
- Можем генерировать разные последовательности!
- Но как размечать?

CpG островки. НММ.

CpG островки



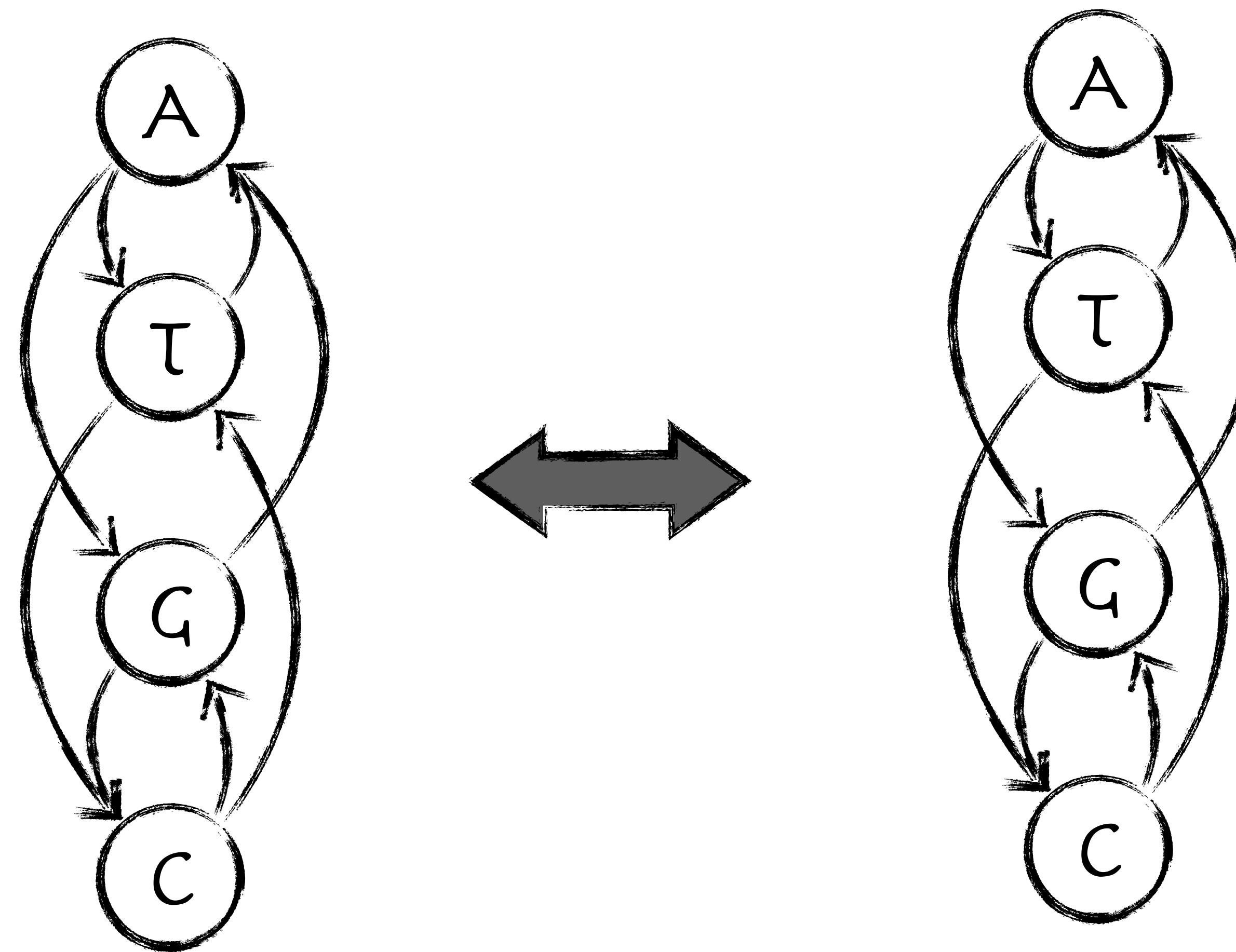
Остальные участки генома



CpG островки. НММ.

CpG островки

Остальные участки генома



Мы видим только последовательность символов, а их происхождение (состояние модели) для нас скрыто!

CpG островки. Параметры НММ.

π , x – последовательности разметки CpG и nonCpG и подстрока генома.

$$\pi_i \in \{ +, - \}, x_i \in \{A, T, G, C\}$$

$a_{kl} = P(\pi_i = l | \pi_{i-1} = k)$ - вероятность начала CpG островка или его окончания. А также вероятность продолжить CpG и nonCpG участки.

$e_k(b) = P(x_i = b | \pi_i = k)$ - вероятность появления нуклеотида b в CpG островке или вне его.

СрG островки. Параметры НММ.

π, x – последовательности скрытых состояний и символов

$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$ - вероятность перехода между скрытыми состояниями l и k .

$e_k(b) = P(x_i = b \mid \pi_i = k)$ - вероятность сгенерировать символ b находясь в состоянии k

СрG островки. Вероятность.

Допустим мы знаем π , x

Тогда чему равна вероятность $P(x, \pi)$?

СрG островки. Вероятность.

Допустим мы знаем π, x

Тогда чему равна вероятность $P(x, \pi)$?

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^L e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$$

0 – начальное состояние

CpG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома x

CpG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома x

Как определить последовательность состояний π ?

CpG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома x

Как определить последовательность состояний π ?

Найдем наиболее вероятную последовательность π , при которой можно получить x !

НММ. Максимальное правдоподобие.

$$\pi^* = \operatorname{argmax}_{\pi} P(x, \pi)$$

Перебор? :)

НММ. Максимальное правдоподобие.

$$\pi^* = \operatorname{argmax}_{\pi} P(x, \pi)$$

Перебор? :)

Для нашей модели СрG островков всевозможных путей 2^L

НММ. Максимальное правдоподобие.

$$\pi^* = \operatorname{argmax}_{\pi} P(x, \pi)$$

Перебор? :)

Для нашей модели СрG островков всевозможных путей 2^L

Если бы мы знали $v_k(i)$ – вероятность, что максимально правдоподобный путь заканчивается в состоянии k при i -том наблюдении, то как нам посчитать $v_k(i + 1)$?

НММ. Максимальное правдоподобие.

$$\pi^* = \operatorname{argmax}_{\pi} P(x, \pi)$$

Перебор? :)

Для нашей модели СрГ островков всевозможных путей 2^L

Если бы мы знали $v_k(i)$ – вероятность, что максимально правдоподобный путь заканчивается в состоянии k при i -том наблюдении, то как нам посчитать $v_k(i + 1)$?

$$v_l(i + 1) = e_l(x_{i+1}) \max_k (v_k(i) a_{kl})$$

НММ. Алгоритм Витерби.

Динамика!

- Инициализация:

$$\nu_0(0) = 1, \nu_k(0) = 0$$

- Рекурсия:

$$\nu_l(i) = e_l(x_i) \max_k (\nu_k(i-1) a_{kl})$$

- Чтобы найти последовательность скрытых состояний найдем $P(x, \pi^*) = \max_k (\nu_k(L) a_{k0})$ и восстановим последовательность обратным ходом

Витерби. Пример.

Пусть $x = CGCG$

Этот участок мог бы быть порожден состояниями + - + - или - - - -, найдем наиболее правдоподобную последовательность состояний

	C	G	C	G
1	0	0	0	0
A+	0			
C+	0			
G+	0			
T+	0			
A-	0			
C-	0			
G-	0			
T-	0			

Витерби. Пример.

Пусть $x = CGCG$

Этот участок мог бы быть порожден состояниями + - + - или - - - -, найдем наиболее правдоподобную последовательность состояний

	C	G	C	G
1	0	0	0	0
A+	0	0	0	0
C+	0			
G+	0			
T+	0			
A-	0			
C-	0			
G-	0			
T-	0			

Витерби. Пример.

Пусть $x = CGCG$

Этот участок мог бы быть порожден состояниями + - + - или - - - -, найдем наиболее правдоподобную последовательность состояний

	C	G	C	G
1	0	0	0	0
A+	0	0	0	0
C+	0	0.13	0	0.012
G+	0	0	0.034	0
T+	0	0	0	0
A-	0	0	0	0
C-	0	0.13	0	0.0026
G-	0	0	0.010	0
T-	0	0	0	0

Витерби. Пример.

Пусть $x = CGCG$

Этот участок мог бы быть порожден состояниями + - + - или - - - -, найдем наиболее правдоподобную последовательность состояний

	C	G	C	G	
1	0	0	0	0	
A+	0	0	0	0	
C+	0	0.13	0	0.012	0
G+	0	0	0.034	0	0.0032
T+	0	0	0	0	0
A-	0	0	0	0	0
C-	0	0.13	0	0.0026	0
G-	0	0	0.010	0	0.00021
T-	0	0	0	0	0

Витерби. Сложность.

- Оценка сложности по времени

Витерби. Сложность.

- Оценка сложности по времени
 $O(ns^2)$
- По памяти?

Витерби. Сложность.

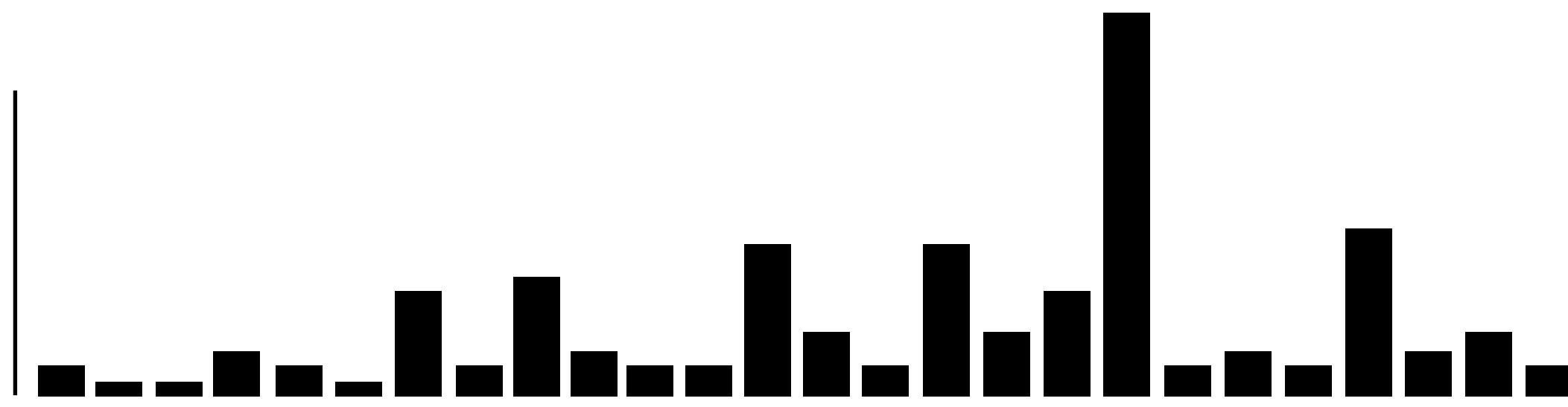
- Оценка сложности по времени
 $O(ns^2)$
- По памяти
 $O(ns)$

Просмотр вперёд и назад

С какой вероятностью некоторая позиция принадлежит CpG островку?

Будем искать ответ для всех позиций $i \in [1, L]$ В результате получим некоторое распределение для состояния +

CTTCATGTGAAAGCAGACGTAAGTCA



Просмотр вперёд и назад

x – последовательность наблюдений

a_{ij} – матрица переходов

$e_k(b)$ – вероятность генерации b в состоянии k

$f_k(i) = P(x_1, \dots, x_i, \pi_i = k)$ – вероятность наблюдать

подпоследовательность $x[\dots i]$ и попасть в состояние π_i

$b_k(i) = P(x_{i+1}, \dots, x_L, \pi_i = k)$ – вероятность на i -том наблюдении быть в состоянии π_i , а после этого наблюдать подпоследовательность $x[i \dots]$.

Тогда $P(\pi_i = k, x) = \frac{f_k(i)b_k(i)}{P(x)}$, где $P(x)$ – вероятность наблюдать x

Просмотр вперёд и назад

Найдем $f_k(i) = P(x_1, \dots, x_i, \pi_i = k)$ — вероятность наблюдать подпоследовательность $x[\dots i]$ и попасть в состояние π_i

Просмотр вперёд и назад. Вперед!

Найдем $f_k(i) = P(x_1, \dots, x_i, \pi_i = k)$ — вероятность наблюдать подпоследовательность $x[\dots i]$ и попасть в состояние π_i

Нужно просто просуммировать вероятности с предыдущего шага!

- Инициализация:

$$f_0(0) = 1, f_k(0) = 0$$

- Рекурсия:

$$f_l(i) = e_l(x_i) \sum_k (f_k(i-1) a_{kl})$$

Просмотр вперёд и назад. Назад!

Найдем $b_k(i) = P(x_{i+1}, \dots, x_L, \pi_i = k)$

- Инициализация:

$$b_k(L) = a_{k0}$$

- Рекурсия:

$$b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i + 1)$$

- Посчитаем в конце $P(x) = \sum_L a_{0l} e_l(x_1) b_l(1)$

Просмотр вперёд и назад.

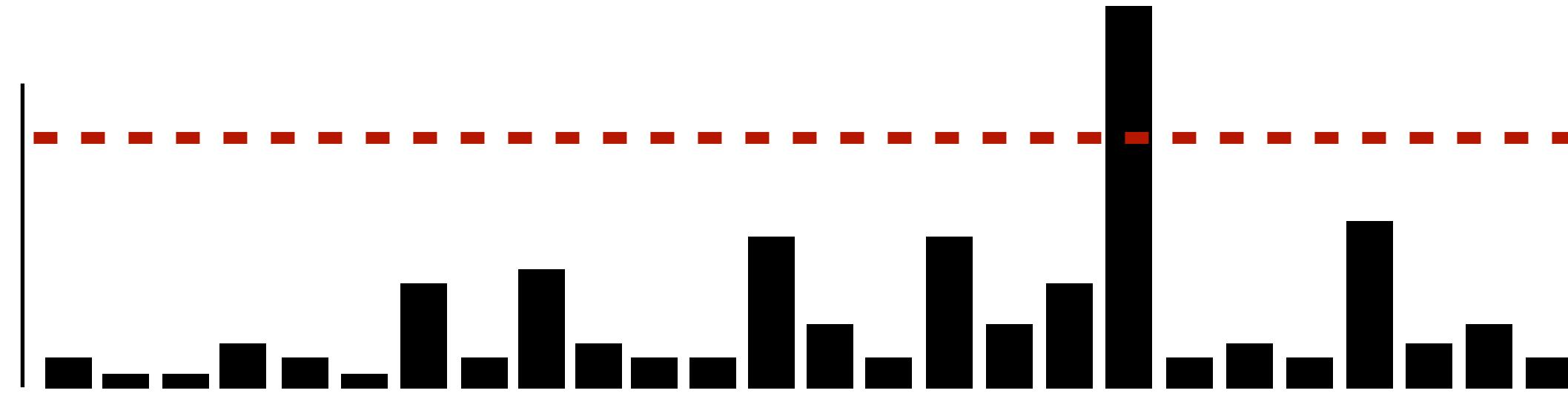
Когда известны $f_l(i), b_l(i)$, можно посчитать $P(\pi_i = k, x) = \frac{f_k(i)b_k(i)}{P(x)}$

- Оценка сложности по времени
 $O(ns^2)$
- По памяти
 $O(ns)$

Просмотр вперёд и назад. Замечания.

- Мы получили распределение для состояний на каждой позиции, а значит можем для любой последовательности состояний найти ее вероятность.
- Взяв распределение и некоторый порог, мы можем найти ответ и на первый вопрос

CTTCATGTGAAAGCAGACGTAAGTCA



Резюмируем.

- Для задачи генерации размеченных последовательностей хорошо подходят марковские модели со скрытым состоянием.
- Наиболее вероятная последовательность состояний для заданной НММ восстанавливается алгоритмом Витерби.
- Алгоритм просмотра вперед и назад позволяет для заданной НММ определить распределение вероятности по состояниям на каждом шаге генерации.

**Разбор задачи про лучшее
расстояние Хэминга.**

Задача

Даны 2 строки, a, b длины n, m соответственно, над алфавитом $ATGC$.
Пусть $n \geq m$.

Нужно находить подстроку в строке a , которая ближе всего по
расстоянию Хэмминга к b . Сложность алгоритма $O(n \log(n))$

Решение. Шаг 1, представление строк

Представим строку a следующим образом $a \Rightarrow a_A, a_T, a_G, a_C$

$$a_A = \{ \text{if } a_i == A \text{ then } 1 \text{ else } 0 \}_{i=1}^n$$

То же самое для строки b

$b \Rightarrow b_A, b_T, b_G, b_C$, но добавим в конец $n - m$ нулей

Решение. Шаг 2, решение за $O(nm)$

Сколько символов из b совпадает с символами из $a[t \dots t + m]$?

Решение. Шаг 2, решение за $O(nm)$

Сколько символов из b совпадает с символами из $a[t \dots t + m]$?

$$\sum_{j=0}^m a_A[j+t] * b_A[j] + \sum_{j=0}^m a_T[j+t] * b_T[j] + \sum_{j=0}^m a_G[j+t] * b_G[j] + \sum_{j=0}^m a_C[j+t] * b_C[j]$$

Посчитав массив c_t

$$c_t = \sum_{s \in \{A, T, G, C\}} \sum_{j=0}^n a_s[t+j] * b_s[j]$$

Найдем максимум

Решение. Шаг 2, решение за $O(nm)$

Пример $a = GATTACA$, $b = TTTC$

$$a_A = 0100101, \quad b_A = 0000000$$

$$a_T = 0011000, \quad b_T = 1110000$$

$$a_G = 1000000, \quad b_G = 0000000$$

$$a_C = 0000010, \quad b_C = 0001000$$

$$c_A = 0000000$$

$$c_T = 1221000$$

$$c_G = 0000000$$

$$c_C = 0010000$$



Решение. Шаг 3, решение за $O(n \log(n))$

$$c_t = \sum_{s \in \{A, T, G, C\}} \sum_{j=0}^n a_s[t+j] * b_s[j]$$

c - ничто иное как сумма четырех сверток.

По теореме свертки

$F(conv(x, y)) = F(x) * F(y)$, где F - это преобразование Фурье, а $*$ - это скалярное произведение.

Нам нужно сделать прямое преобразование Фурье, перемножить скалярно вектора и сделать обратное преобразование

$$conv(x, y) = F^{-1}(F(x) * F(y))$$