

# **НММс. Алгоритм Витерби для поиска СрG островков.**

**Алгоритмы в биоинформатике**

**Мелешко Дмитрий**

**meleshko.dmitrii@gmail.com**

# Что было на прошлой лекции?

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

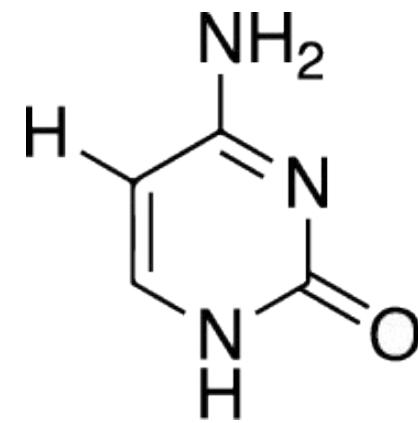
# Что будет на этой лекции?

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

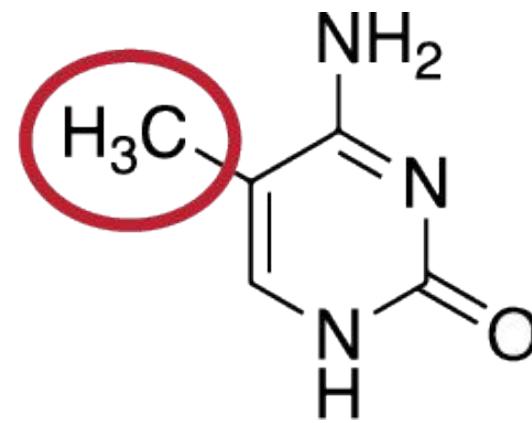
# CpG островки.



# CpG островки.



Цитозин



Метилированный  
цитозин



Тимин

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

# CpG островки.

CpG sites	GpC sites
<p>CCCCGGTC CGGGCG GGGAAAGAGC CG CCTCAA CG GCAGGGCCCATC CGCGA GAGGCCAG CGCCCG CG CGTCCAGCCCAGGCC CG CG CCTCC CG CCTCG GGCTGCTCCCTCG CGGCCCTGCAC CG CCTCCCTGCTACTTGGAC CG CTTC CTCA CGCCCTTCTCCACCC CGCGCG CCAGCCTCC CGCGCG CAGCGTGGGG ATCT CGGCCAATAAAGGAGAAAGGG CGCG GCC CGTA CGCGCG CAGGGTGC CGTGGCG AGACCAAGCTCA CGCCCCCTCCCTCCAGCGCG CCAAGGCCCG CGGCC ACAGCTGCTGGCTGCAGTCAGAAGCG TAGGCC CGAGACAAGGAAGGG CGC CTTGACT CGCACTTTGTC CG GTT CGAA CG TTCTGCTCAGTGGTG CGTGG AATG CGAGCGCGCTTAAAT CGATGGCGCCTAGGAGTCCATGAAATA CG GTACAGGCTTC CGCGCG CGATGCCCG CGCCCTCACCCA CGCTCC CGCCCT CGGGGATGCCCGACCCCG CGTGGCGTCCCG CGTCCC CGCGCAGGCG CGCT CGGGCTGCC CGTGGCTCTT CGCA CGCGCG CGATGCC CGACTCC CGAGC TGCAGCTGGTTGAGCAG CGGATCC CGAGCTTCCC CGACTTCCC ACCTCG GGCGTGGTATTCAAGGTGCA CGCACAGGC CGCCCT CGTGGCGCC CGACCT GCGGGCTAC CGATGGGA CGCGCG CGTCCC CGGGCGGG CGGGG CGGAAACCCCT CGCTTT CGCCCG CGGGCCCTGCCCTCC CGGCCCG CG CGTACCCAGGCCCTGCTTGGGTCCAGGGACATCT CGCCCG CGTCTGAAGG ACCC CGCTCCCTCG CGCGCG CGCAT CGCCCTCTGG CGCGACACCTGAAG GCGACCC CGGGGG CGCAT CGACTACAT CGCAGG CGAGTGCCCGTGGC CGCATCTAAGG CGCTTCC CGCTCTG CGCGCG CGAGGGCAGCA CGTGGC TCTG CGCG TCTGCTTGGGGAGGGGCTT TGGGTGCTTCAGGGGG CGCG GGACCGG CGCG CGTCTGGTGGT CGCC CGGAAAGGGTGTGAGATTGAGCCC CGEAGGG CGCG CGCAT CGTGCAGGCC CGCTTCC CGCAGGTT CGGGTCCC AGCCCAGGACAGG CGTACCG CGAGTGC CGGGTCAGTTGGTCTCCCTGGAG TGCCCAAGCTGAATCCACAGGGCCAGCTGCCCTGCTTCTGTTCTTCT CGAGCTGGTATTGAG CGCCCTGCCA CGAGCCAGGGCTTCCCTGGTGAAGA TCAC CGGAATGCCACCCAGGGAGGGCTGGAGGCC CGGGAGAGC CCAAGAGGTGGCCCAGGGAGAACAGAGTGTCCCTGGCG TCTTGCCTCTC CTAGGGTGTGACAGCCCACCTCCCTGGACACTGCCCTGAGGAAAG CGCGAG CTCTTGCTGGAGCCACAAACTGCCAGCTCCCTCACCTCTGCCAG GAAGCCCTCCCTGACCTCTGCCAGGC CGGGCAGGGTTCCCTGAGCG CCCCAACCATCACAGCTCAGGCCACCT CGAGAGACTCCCTTTAGACA GAAGCCCTGGTGCAGAGCTGCCCTTGAGAGTAAGCTGAGGCCCTGTGAGGT TTCTACAGGCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGGTGG TGACTCCCTAGGAACACACAGCTAAGAGTGGTCCCTTAAAGACAGAC CCAGGTCTGCACCTGACCTGGAGCAGCTCG CGGGTAGGTGATGGTAAC ATTCCCTAAATGGTGCATGCACTGGCTTCACTGGGAGGCCAACCGAG TACCCCTGCCAC CGGCCAACCTGGCCCTGGGATTCCTCATGCTGC CG AGTCACCTCTGTCACTTACCCCTGACAGGCCTAGACTCC CGAGGTTCTC TTTGGCCCTCCCTGGCCAGGAGCTGGACTGGCTG CGTGCATCG AAAG CGGGGAAGCTGCCAGGCCACTCTGTTGGCTCTTATCCCTGG AGTA CGGGGAAGGTAAGAGGGCTGGGTGGCCAGAGGAAGGGCAGGGCCAG GCCAC CGTGGCAACTCTCCCCAGTTCTAAAAGGCCCTCCAGG CGTGTG AAGTGGAGCTGCTGTGGTACAGTGGCTTGGAGCTCAGAGAGGGTGA ACATAGGCTGGGCTCACACAGCCAGGTACAGCAAGGTGGGTTGGAGTC AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTG CG GGAGGCTGAGGACCACACACCAACTCCCACTCCAGGCTGAGCTGGAGATT CAGAAAGA CGCCCTGGAGCCAGGACAGAGGGTGGT CGTGTGATGATCT GCTGGCCACTGGTGGTAAGGGTCTCC CGAGCCAACTGCTGTGGCTCCA AGGGCCTGGTGGAGTGGGACAGGACCT CGCTGTGACATGGGATGCAG CTTACTGTTGTCAGAGGGTGCCTGGTGGCAGGC CGACACCTCCCTCTC CCCAGGCCCTCCCCCAACCCAGGGCTGGGAGCAGCTGCTCT CTGCAGGCCAGGCCACTGGGACCTCACCCCTCCATCCCCAGGAACCAT GAA CGCTGCCCTGTGAGCTGCTGGCG CGCTGCCAGGCTGAGGTCTGGAGT CGCTGAGGCTGGTGGAGCTGACCT CGCTTAAAGGGCAGGGAGAAGCTGGCA CCTGTACCCCTCTCTCTGCCAGTATGAGTGA CGCACAGGGCTCCC AGCCCCACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGCAACT GCAGTGAACCAAGGGCAC CGCTGCCACAGGGAAACACATTCCCTTGCTGG GGTTCA CGCCTCTCCCTGGGCTGGAAAGTGCCAAAGCCTGGGCAAAGCT GTGTTTCA CGGCCACTGAACCAATTACACACAG CGGGAGAA CGCAGTAA ACAGCTTCCCAC</p>	<p>CCCCGGTCCGG CGGGGAAGA CGCGCTCAACG CGAGGCCCATCC CGGA GAGGCCAG CGCC CGCCCGGCCCGGCGTCCA CGCCAGGCC CGCCCGTCCGCCCTG GCT CGTCCCTCCGG CGCCCT CGCACC CGCCCT CGCTACTTGGACCG CGTTC CTCAC CGCCCTTCTCCACCC CGCCCGCCA CGCTCC CGCCCGA CGTGGGG ATCT CGGCCAATAAAGGAGAAAGGGCGCCCG CGCCCGTAC CGCCCGCAGGT CG GTGG CGGAGACCA CGCTCAC CGCCCTCCCTCCA CGCCGCCAAGGCCCG CGCC ACA CGTCCCTG CGT CAGTCAGA CGCGTA CGCCGAGACAGGAAGGCC CTTGACT CGCACTTTGTC CG GTT CGAA CGTCTGCT CGT CAGTGGT CGTGG AATG CGAGCGCGCTTAAAT CGATGGCGCCTAGGAGTCCATGAAATA CG GTACAGCGCGTCC CGCC CGACGGAT CGCCCGCAGTCCCGAGCG CGCCGGGAT CGCCCGCCCGTGTGCCCGTCCCGCCCGCGCGAGCG CGCTCGGGCGT CGCCCGTGTCTCGCACCGCGCCAT CGCCCGACTCCCGAGCG TCA CGTGGTGA CGA CGGGATCC CGA CGTCTCCCGACTTCCCACCC GCCGTGGTATTCAAGGT CGACCGACAGCGCCCGCGCGACCGT CGCGCGCTACCGGATGGGA CGCGCGTGC CGCCCGCGACCTCCCGGGCGGG CGGGAACCCCTCGTCTT CGCCCGCGCGCGCGCGCGCGCGCG CGTCACCGAGCGCGTGTCTTGGGTCCAGGGACATCT CGCCCGTCTCG ACCC CGCTCCCTCC CGCCCGCCAT CGCCCGTCTCG CGCCCGACACCTGAAG CGCACCC CGGGGGCGCGAT CGCAGCTACAT CGCAGCGAGT CGCCCGAGT CGCAGCGCGTACCGGATGGGA CGCGCGTGC CGCCCGCGACCTCCCGGGCGGG CGGGGAACCCCTCGTCTT CGCCCGCGCGCGCGCGCGCGCGCG CGTCACCGAGCGCGTGTCTTGGGTCCAGGGACATCT CGCCCGTCTCG ACCC CGCTCCCTCC CGCCCGCCAT CGCCCGTCTCG CGCCCGACACCTGAAG CGCACCC CGGGGGCGCGAT CGCAGCTACAT CGCAGCGAGT CGCCCGAGT CGCAGCGCGTACCGGATGGGA CGCGCGTGC CGCCCGCGACCTCCCGGG TCT CGCCCGTCT CGTGGGGAGGGCGCTTGGGTGCTTCAGGGGGCG GGACGGCGCCCGCGTGTGGTCCCGGGAGGGGGTGTGAGATTGA CGCC CCGAGGCC CGCC CGCGCGTGT CGCAGCGCGTCCCTCC CGA CGTTC ACCCAGGA CGCGCGTACCGAGT CGCGGGTCAAGTTGGTCTCCCTGGAG CGCCCAA CGTGAATCCACAGGCCCGA CGTGCCTCTCGTCTGTTCT CGGA CGTGGTATTGA CGCCCGT CGCACCGA CGCCCGTCCCTGGTGAAGA TCACGGAA CGCCACCCAGGGAGGGCGCTGGAGGCCCGCTCCGGGA CCAAGAGGT CGCCCGAGGGAGAACAGAGTGTCCCTG CGCGTCTT CGCTCTC CTAGGGTGTGACAGCCCACCTCCCTGGACACTGCCCTGAGGAAAG CGCGAG CTCTTGCTGGAGCCACAAACTGCCAGCTCCCTCACCTCTGCCAG GAAGCCCTCCCTGACCTCTGCCAGGC CGGGCAGGGTTCCCTGAGCG CCCCAACCATCACAGCTCAGGCCACCT CGAGAGACTCCCTTTAGACA GAAGCCCTGGTGCAGAGCTGCCCTTGAGAGTAAGCTGAGGCCCTGTGAGGT TTCTACAGGCCAGTTACAGATGGGCTGCTCAGCTCAGAGAGAGGGGGTGG TGACTCCCTAGGAACACACAGCTAAGAGTGGTCCCTTAAAGACAGAC CCAGGTCTGCACCTGACCTGGAGCAGCTCG CGGGTAGGTGATGGTAAC ATTCCCTAAATGGT CGATGTCAGTGCCTTCA CGTGGGA CGCACCCAGG TACCCCTGCCAC CGGCCAACCTGGCCCTGGGATTCCTCATGCTGC CG AGTCACCTCTGTCACTTACCCCTGACAGGCCTAGACTCC CGAGGTTCTC TTTGGCCCTCCCTGGCCAGGAGCTGGACTGGCTG CGTGCATCG AAAG CGGGGAAGCTGCCAGGCCACTCTGTTGGCTCTTATCCCTGG AGTA CGGGGAAGGTAAGAGGGCTGGGTGGCCAGAGGAAGGGCAGGGCCAG GCCAC CGTGGCAACTCTCCCCAGTTCTAAAAGGCCCTCCAGG CGTGTG AAGTGGAGCTGCTGTGGTACAGTGGCTTGGAGCTCAGAGAGGGTGA ACATAGGCTGGGCTCACACAGCCAGGTACAGCAAGGTGGGTTGGAGTC AGGGTCTAGGGTGGCAGCTGCCAAGCTGTGCAACAAAGCTGTTTCTG CG GGAGGCTGAGGACCACACACCAACTCCCACTCCAGGCTGAGCTGGAGATT CAGAAAGA CGCCCTGGAGCCAGGACAGAGGGTGGT CGTGTGATGATCT GCTGGCCACTGGTGGTAAGGGTCTCC CGAGCCAACTGCTGTGGCTCCA AGGGCCTGGTGGAGTGGGACAGGACCT CGCTGTGACATGGGATGCAG CTTACTGTTGTCAGAGGGTGCCTGGTGGCAGGC CGACACCTCCCTCTC CCCAGGCCCTCCCCCAACCCAGGGCTGGGAGCAGCTGCTCT CTGCAGGCCAGGCCACTGGGACCTCACCCCTCCATCCCCAGGAACCAT GAA CGCTGCCCTGTGAGCTGCTGGCG CGCTGCCAGGCTGAGGTCTGGAGT CGCTGAGGCTGGTGGAGCTGACCT CGCTTAAAGGGCAGGGAGAAGCTGGCA CCTGTACCCCTCTCTCTGCCAGTATGAGTGA CGCACAGGGCTCCC AGCCCCACATCTCCAGCTGGATCCCAGGGAAATATCAGCCTTGGCAACT GCAGTGAACCAAGGGCAC CGCTGCCACAGGGAAACACATTCCCTTGCTGG GGTTCA CGCCTCTCCCTGGGCTGGAAAGTGCCAAAGCCTGGGCAAAGCT GTGTTTCA CGGCCACTGAACCAATTACACACAG CGGGAGAA CGCAGTAA ACAGCTTCCCAC</p>

# 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?

CTTCATGTGAAAGCAGACGTAAGTCA

-----+-----

# CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

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

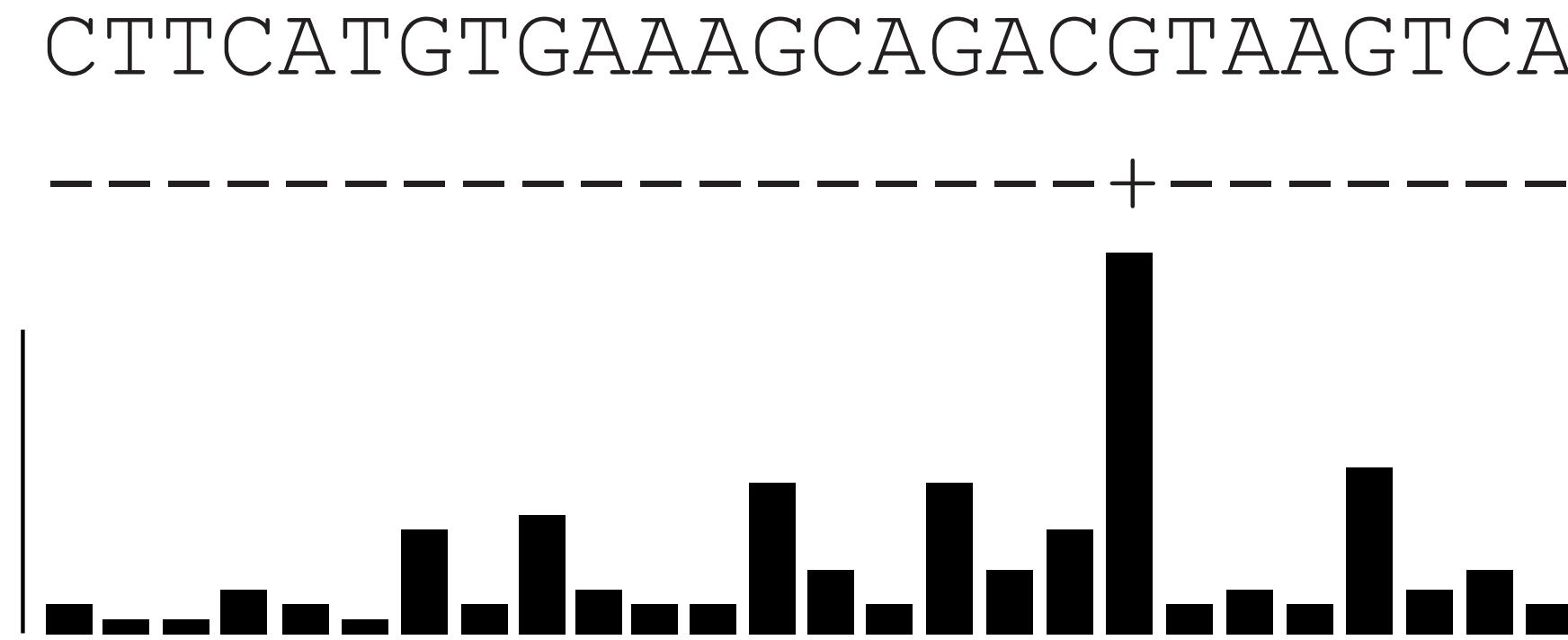
CTTCATGTGAAAGCAGACGTAAGTCA

-----+-----

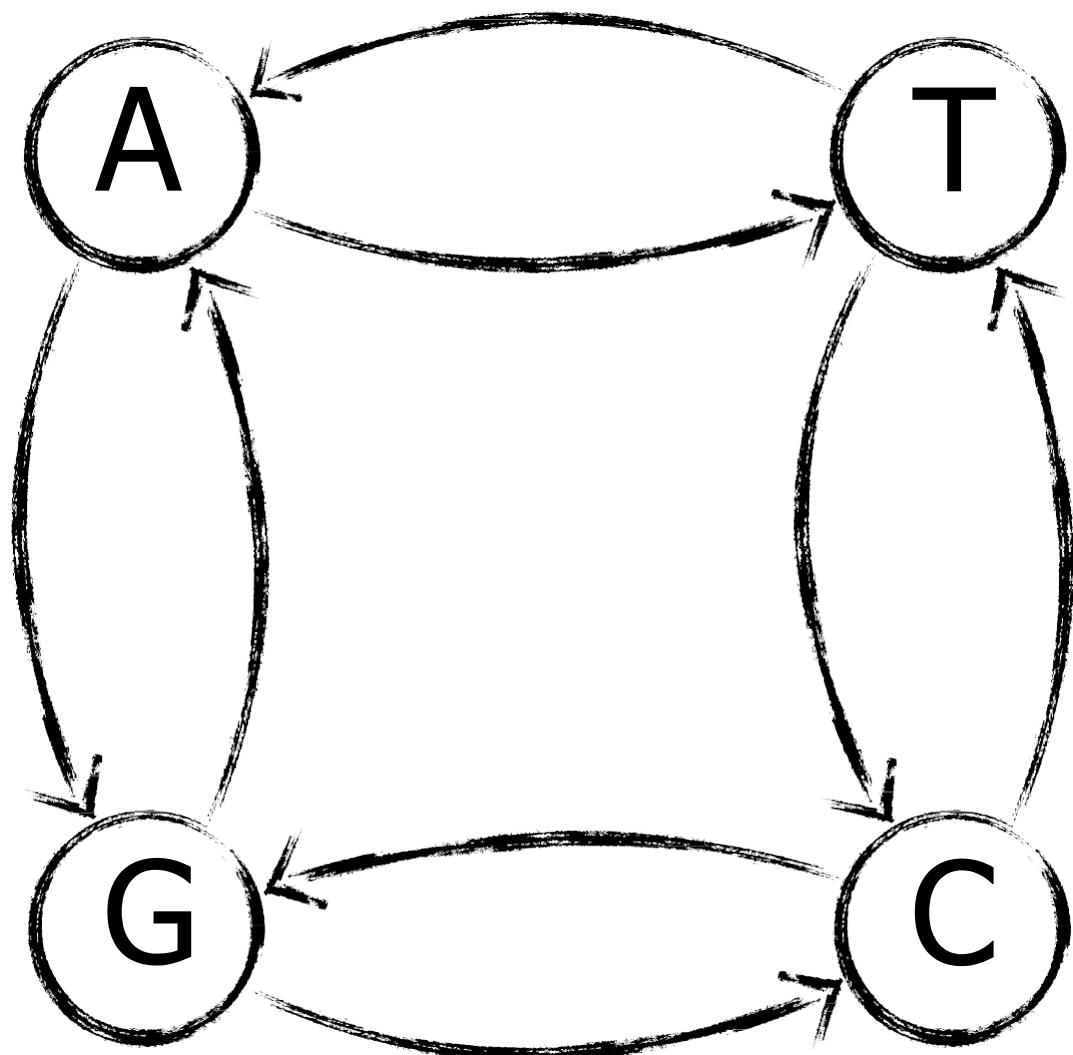
# CpG островки. Задачи.

Интересно научиться отвечать на следующие вопросы:

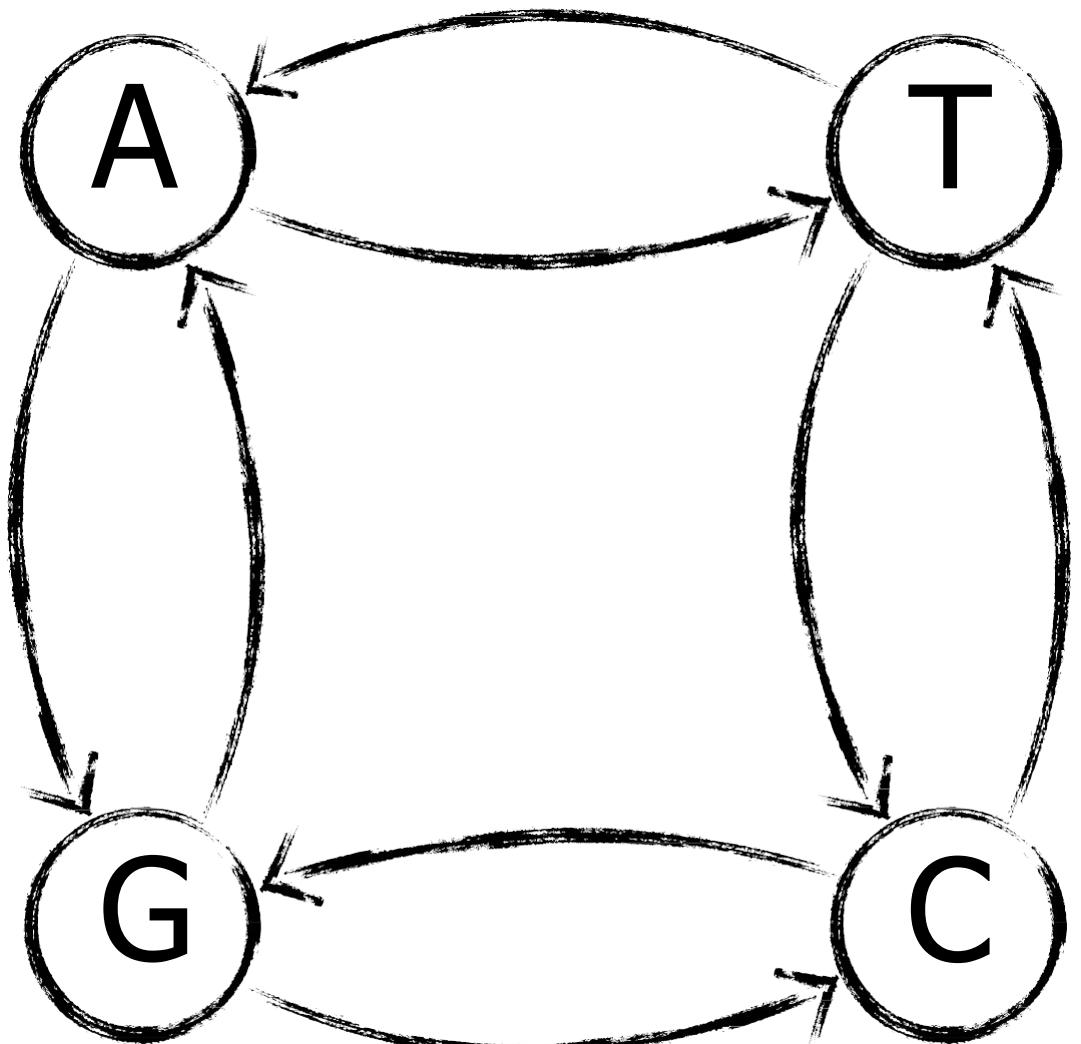
- Где CpG островки в последовательности?
- С какой вероятностью некоторая позиция принадлежит CpG островку?



# 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  
CTAAAAGCTATCCAATATCCACTTGCAGATTCTACAAAAAGAGTG  
TTTCAAAACTGCAGTATCAACAGAAAGGTTCAACTCTGTGAGCTGA  
GTACACACATCACAGAGAAGTTCTGGGAATGCTCTGTCTAGTT  
TTATGTGAAGATATTCCTTTTCAGCATAGGCCTCAATGGGTTCC  
AAATGTCCTTCCAGGTACTACAAAAAGAGTGTTCACAAACTGCT  
CTATGAAAGGGAATGTTCAACTCTGTGAGTTGAATGCAAACATCAT  
GAAGAAGTTCTGAGAATACTTCTGACTAGTTTATGTGAAGATA  
TTCCCATTCCAATGAAAGCCTCAAAGCTGCCAAATATTCCCTTG  
CAGATCCTACAAAGAGAGTGTTCACAAACTACTCTAAAAAGAAA  
TGTTCAACTCTGTGAGTTGAGTACACATATCACAAAGAAGTTCTT  
AGCATGTTCTGTCCTGTTTATTGTAGATCTTCCGGTTCCCG  
TGAAGGCCTCAAAGCTGTCCAA.....

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

# CpG островки. Веса модели.

## Участки из CpG

.....GAATTCTTCTGTCTAGTTTTATAGGAAGATGTTCCCTT  
TTTCAGCGTATGCATCAAAGAGCTCCAAGTTCCACTACAGAGTC  
TTCAAAAAGAATGTTCAAAACTGCTCTATGAAAAGGAATGTTCAC  
CTCTGTGAGTAGAATGCAAGCATCACAAAAAGTTCTGGGAATGC  
TTCTGTCTAGTTTATGTGAAGACATTCCCGTTCCAACGAAAGC  
CTAAAAGCTATCCAATATCCACTTGCAGATTCTACAAAAAGAGTG  
TTTCAAAACTGCAGTATCAACAGAAAGGTTCAACTCTGTGAGCTGA  
GTACACACATCACAGAGAAGTTCTGGGAATGCTCTGTCTAGTT  
TTATGTGAAGATATTCCTTTTCAGCATAGGCCTCAATGGGTTCC  
AAATGTCCTTTCCAGGTACTACAAAAAGAGTGTTCAAAAGTCT  
CTATGAAAGGGAATGTTCAACTCTGTGAGTTGAATGCAAACATCAT  
GAAGAAGTTCTGAGAATACTCTGACTAGTTTATGTGAAGATA  
TTCCCATTCCAATGAAAGCCTCAAAGCTGTCCAAATATTCCCTTG  
CAGATCCTACAAAGAGAGTGTTCAAAACTACTCTAAAAAGAAA  
TGTTCAACTCTGTGAGTTGAGTACACATATCACAAAGAAGTTCTT  
AGCATGTTCTGTCCTGTTTATTGTAGATCTTCCGGTTCCCG  
TGAAGGCCTCAAAGCTGTCCAA.....

	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.331	.375	.125
T	.071	.355	.384	.182

# СрG островки. Веса модели.

	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.331	.375	.125
T	.071	.355	.384	.182

СрG островки

	A	C	G	T
A	.300	.205	.285	.210
C	.322	.218	.078	0.302
G	.248	.246	.218	.208
T	.177	.231	.212	.212

Остальные участки генома

# СрG островки. Веса модели.

	A	C	G	T
A	.180	.274	.426	.120
C	.171	.368	0.274	.188
G	.161	.331	.375	.125
T	.071	.355	.384	.182

СрG островки

	A	C	G	T
A	.300	.205	.285	.210
C	.322	.218	0.078	0.302
G	.248	.246	.218	.208
T	.177	.231	.212	.212

Остальные участки генома

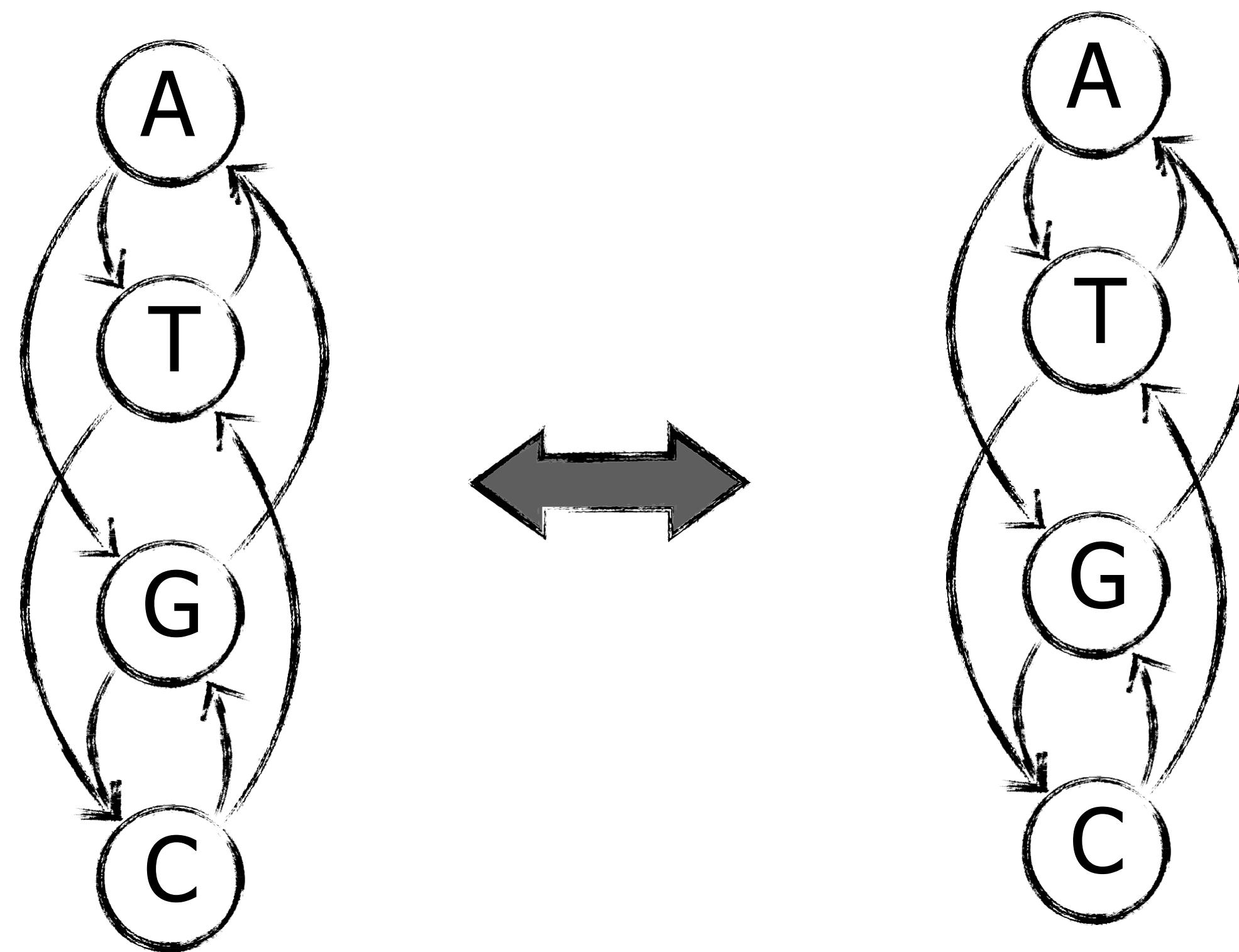
# СрG островки.

- Можем генерировать разные последовательности!
- Но как размечать?

# CpG островки. НММ.

CpG островки

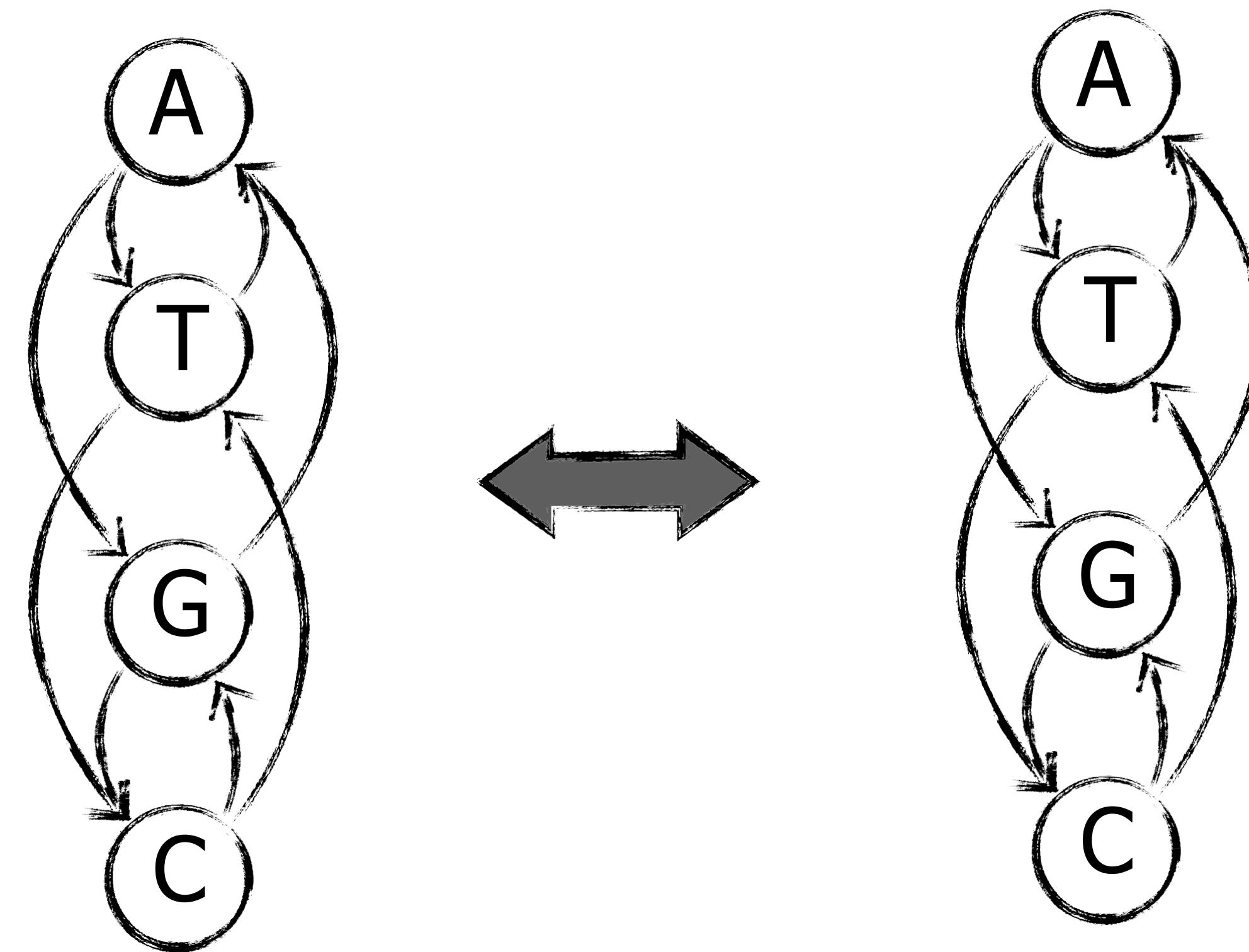
Остальные участки генома



# 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.

$\pi$ ,  $x$  – 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.

$\pi, x$  – 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  $\pi$ ,  $x$

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

# СрG островки. Вероятность.

Допустим мы знаем  $\pi, x$

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

# СрG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома  $x$

# СрG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома  $x$

Как определить последовательность состояний  $\pi$ ?

# СрG островки. Вероятность.

Допустим мы знаем параметры модели и подстроку генома  $x$

Как определить последовательность состояний  $\pi$ ?

Найдем наиболее вероятную последовательность  $\pi$ , 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 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  $v_k(i + 1)$ ?

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

# НММ. Алгоритм Витерби.

Динамика!

- Инициализация:

$$v_0(0) = 1, v_k(0) = 0$$

- Рекурсия:

$$v_l(i) = e_l(x_i) \max_k (v_k(i-1) a_{kl})$$

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

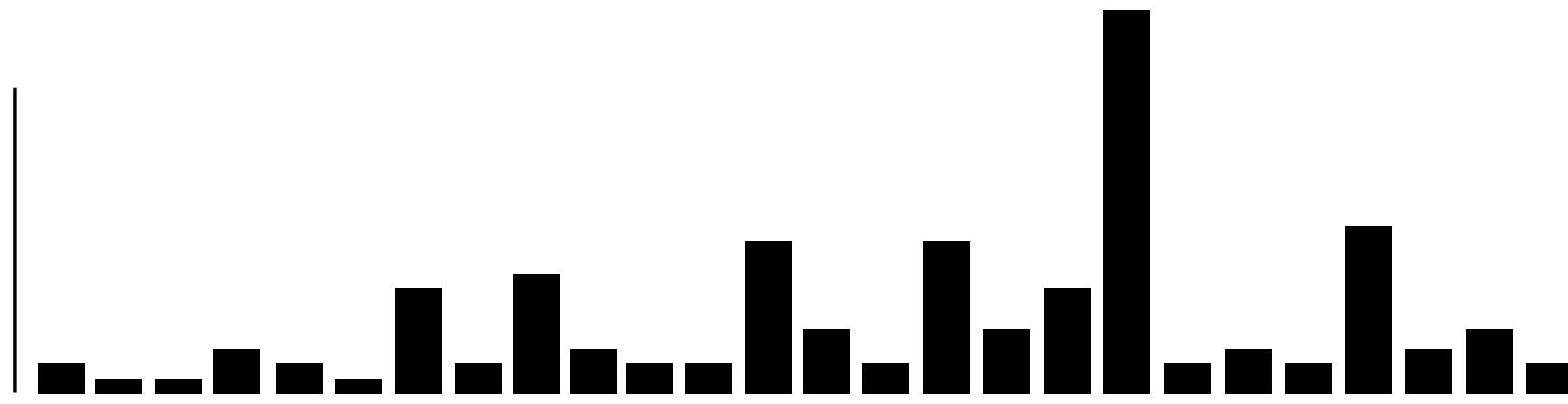
$$O(ns)$$

# 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  $\pi_i$

$b_k(i) = P(x_{i+1}, \dots, x_L, \pi_i = k)$  — вероятность на  $i$ -том наблюдении быть в состоянии  $\pi_i$ , 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  $\pi_i$

# 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  $\pi_i$

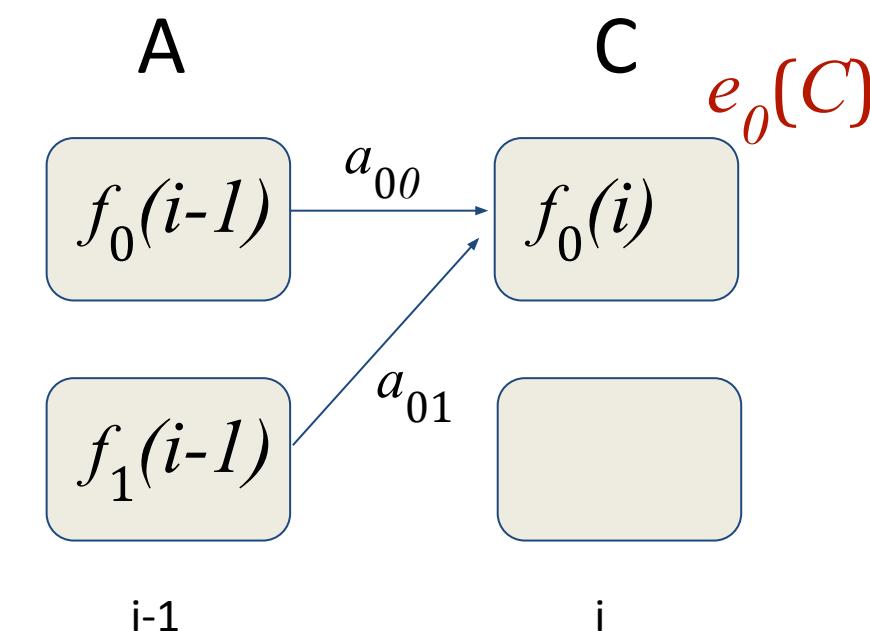
Нужно просто просуммировать вероятности с предыдущего шага!

- 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

$$O(ns^2)$$

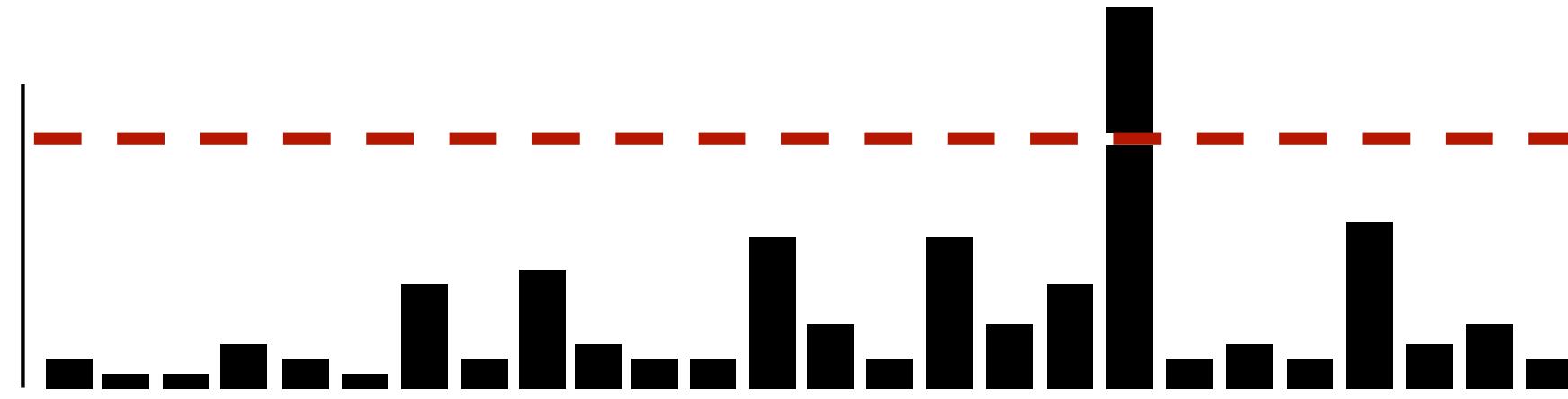
- По памяти

$$O(ns)$$

# 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

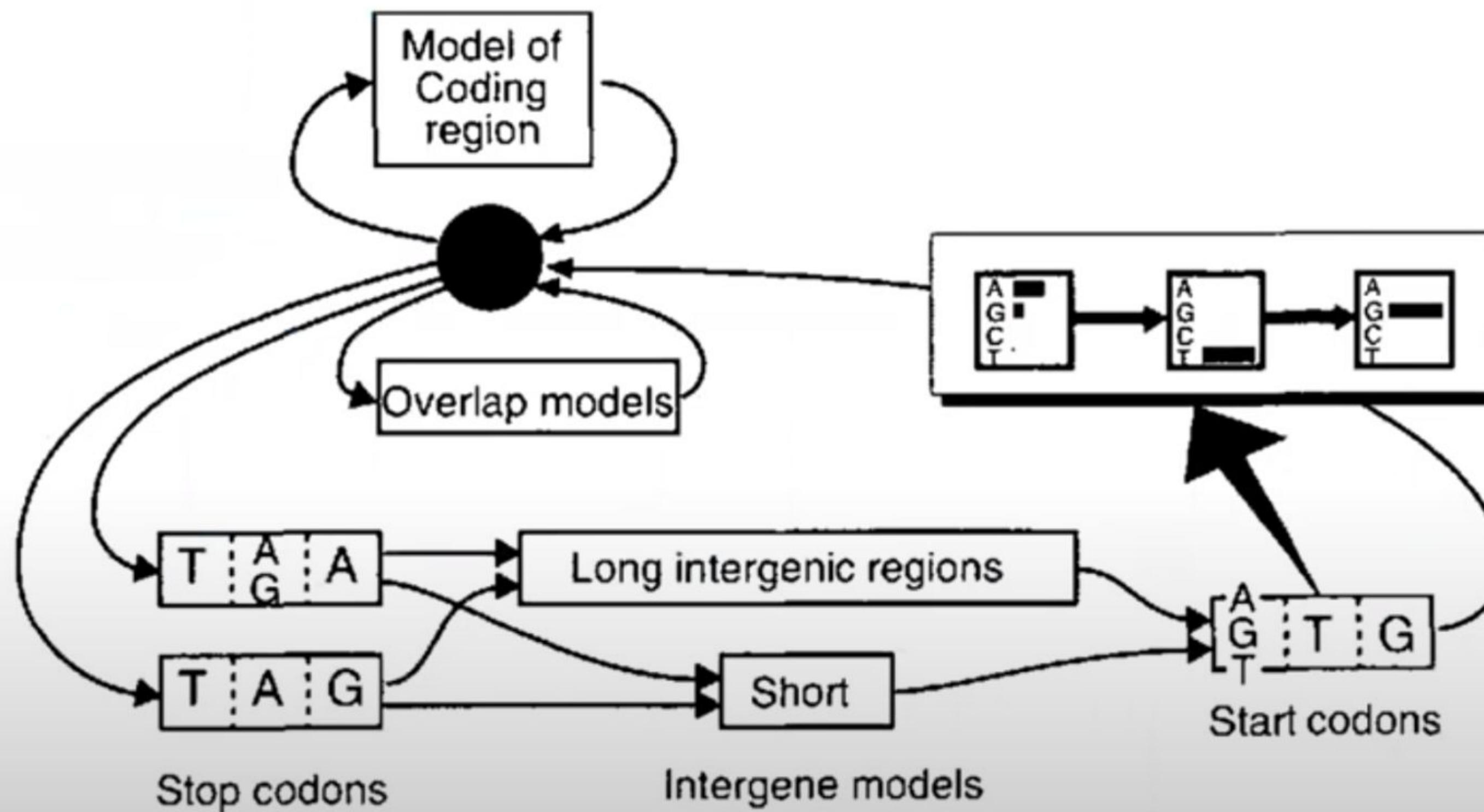


# Резюмируем.

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

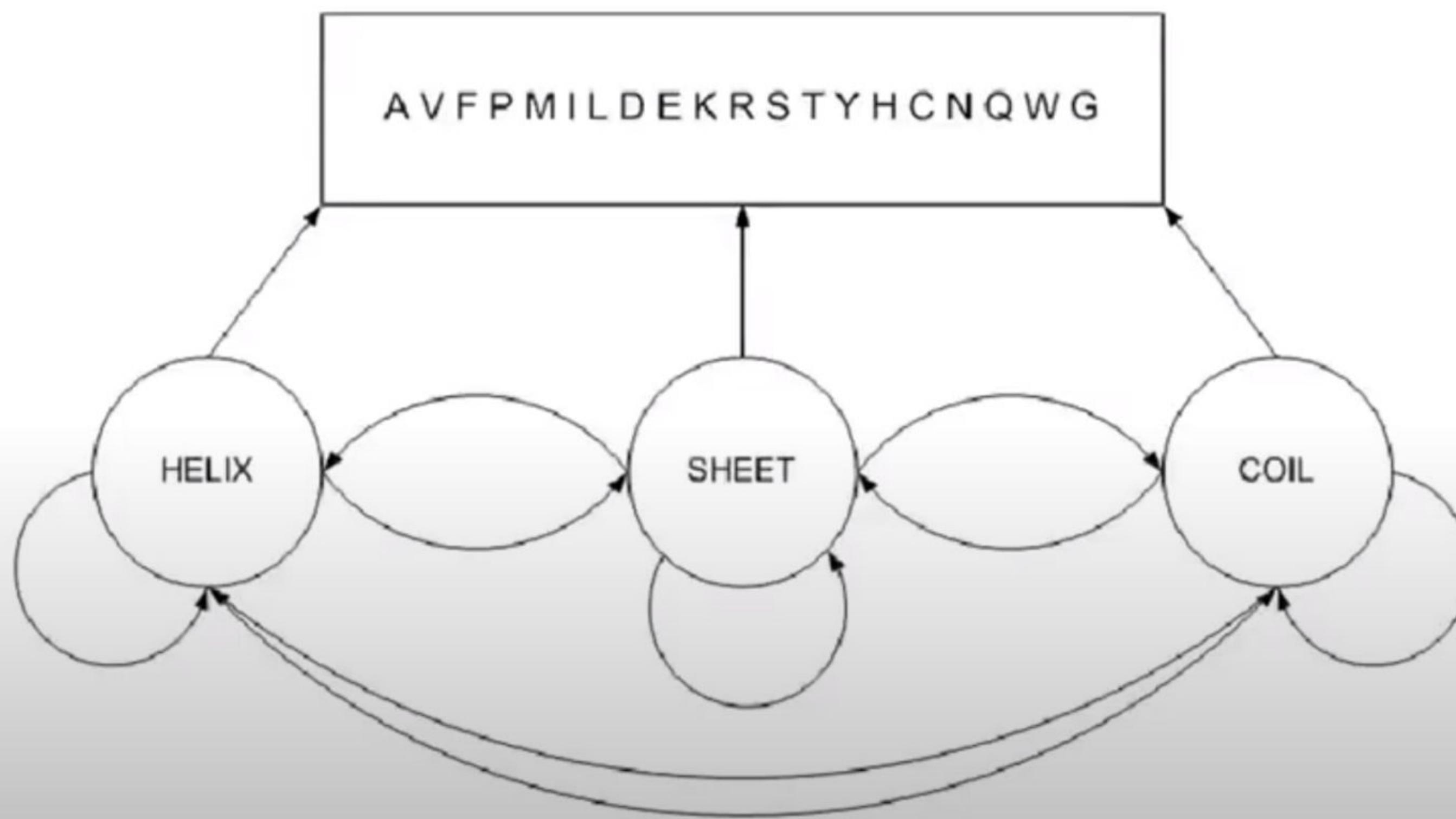
# Другие (древние) применения НММ в биоинформатике

- Gene prediction



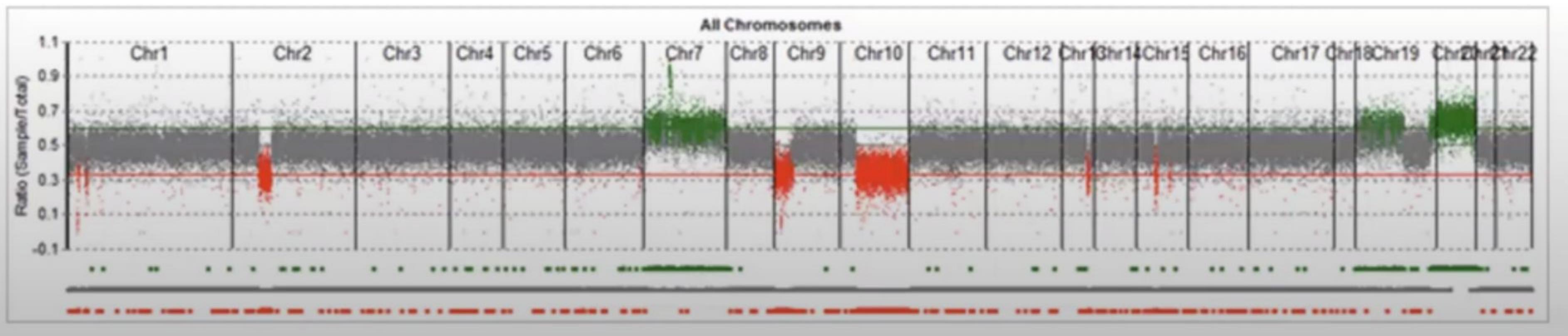
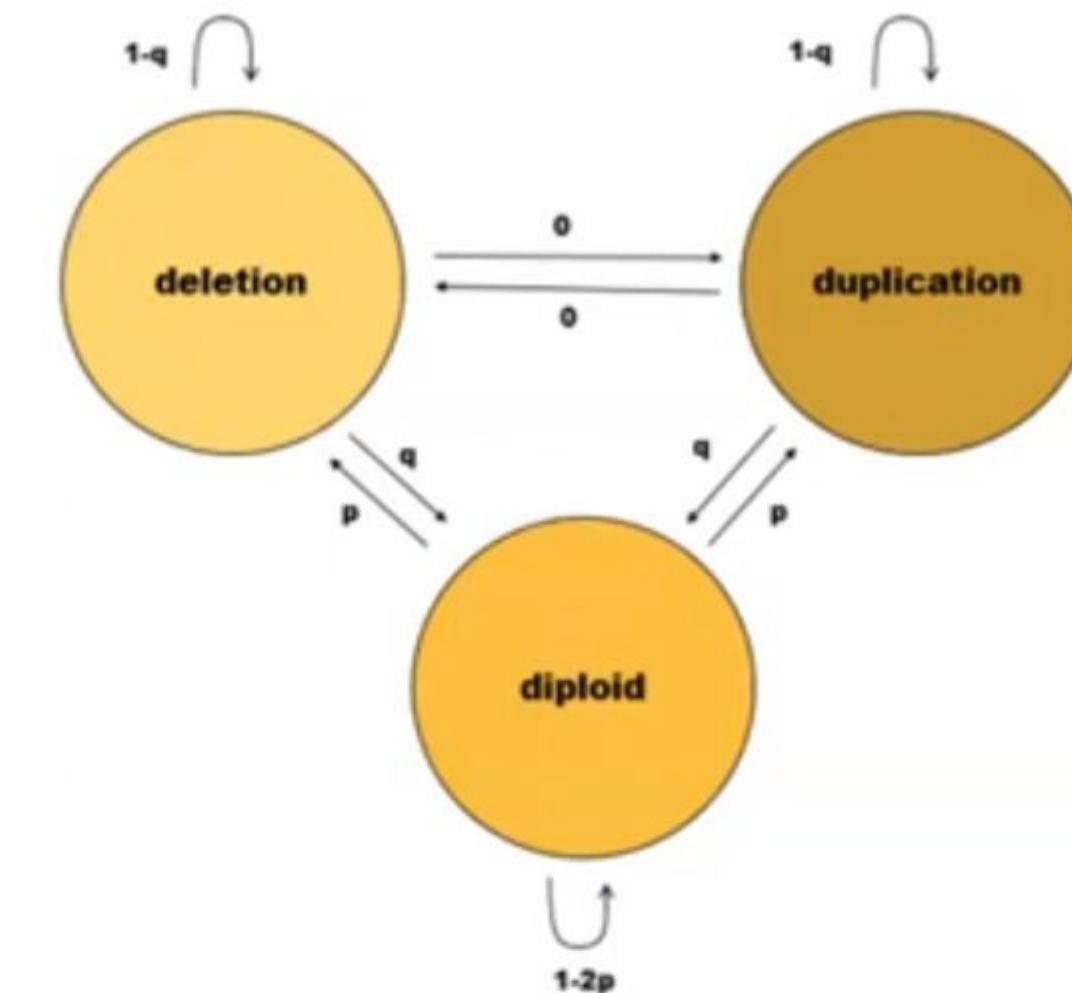
# Другие (древние) применения НММ в биоинформатике

- Protein structure prediction



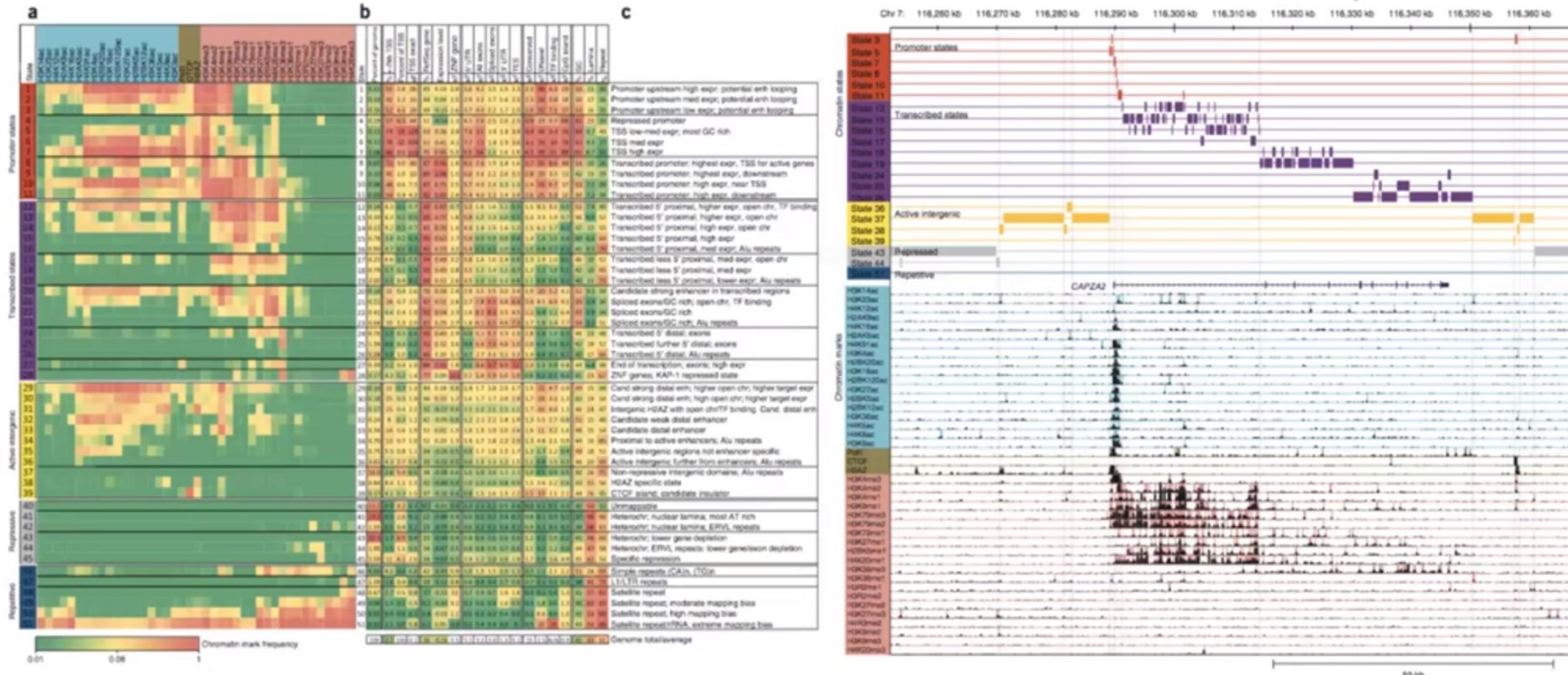
# Другие (древние) применения НММ в биоинформатике

- Copy number variation



# Другие (древние) применения НММ в биоинформатике

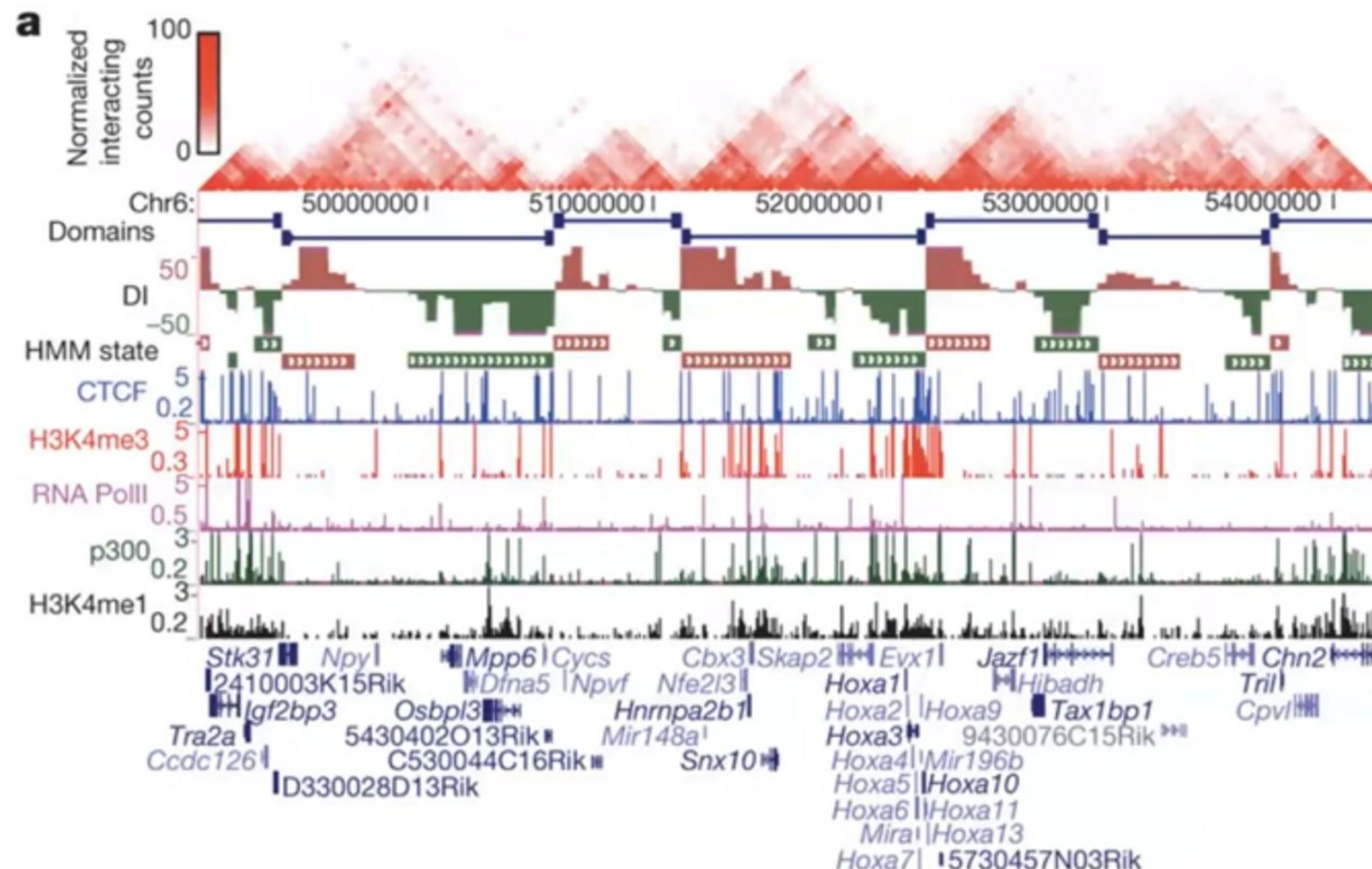
- Chromatin domains based on histone mark ChIP-seq



ChromHMM, Ernst et al, Nat Meth 2012

# Другие (древние) применения НММ в биоинформатике

- Chromatin topologically associating domains



Dixon et al, Nat 2012