

7.91 / 7.36 / BE.490

第三讲

2004-03-02

# DNA 模体建模与发现

Chris Burge

# DNA 序列比较与比对回顾

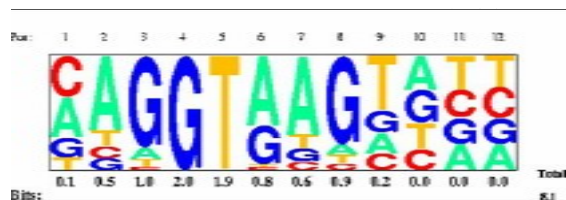
- 目标序列和错配罚分
- 真核基因结构
- 比较基因组学应用：
  - **Pipmaker** （两序列比对）
  - 系统发育投影（多序列）
- **DNA 序列模体介绍**

# 主题结构

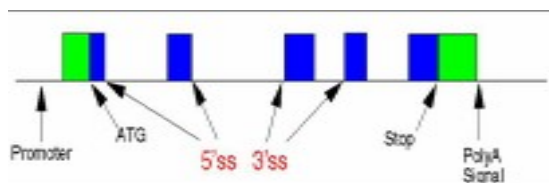
章节  
模型

对象  
关性

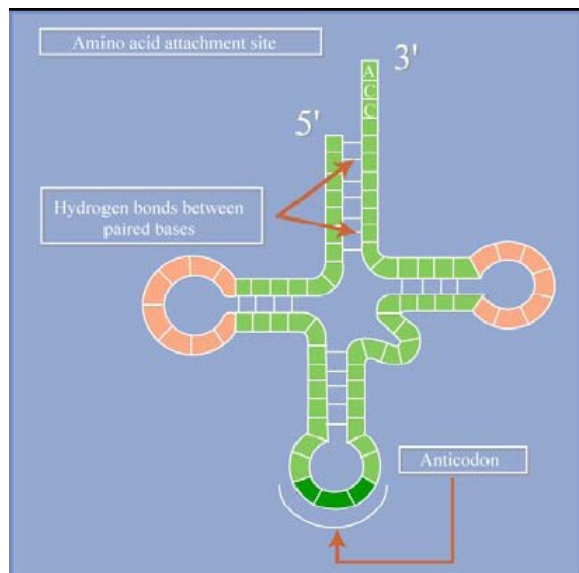
3/2



3/4



3/9



权重矩阵  
模型

无相关性

隐马尔可  
夫模型

局部相关性

能量模型  
共体模型

无局部相关性

# DNA 模体的发现与建模

- 回顾——剪切位点的权矩阵模型
- 模体的信息量
- 模体的发现与搜索的问题
- Gibbs 抽样

## Gibbs 抽样算法多媒体体验

- 模体模型——权矩阵之上

见 Mount 的第四章

# 剪切位点 I

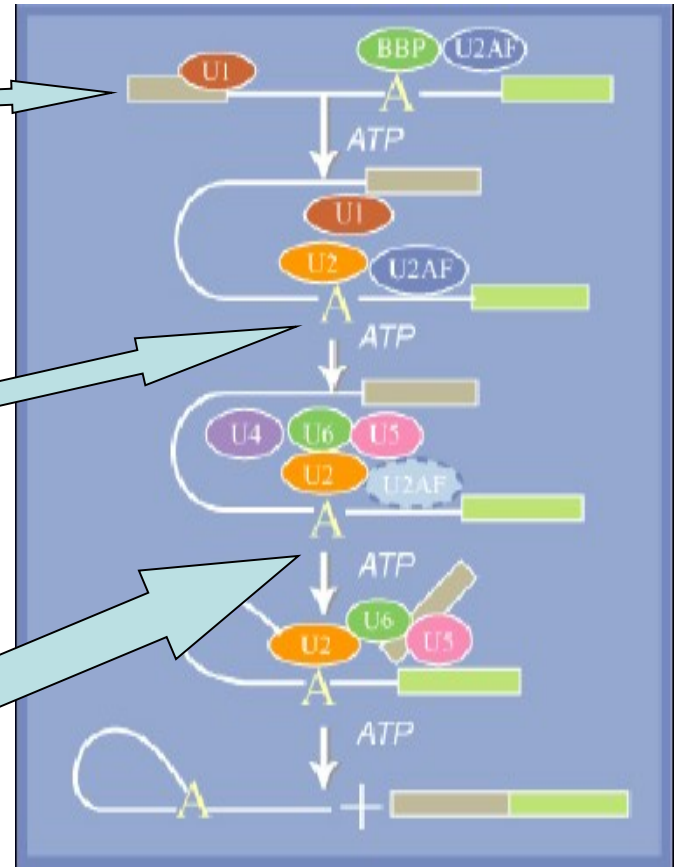
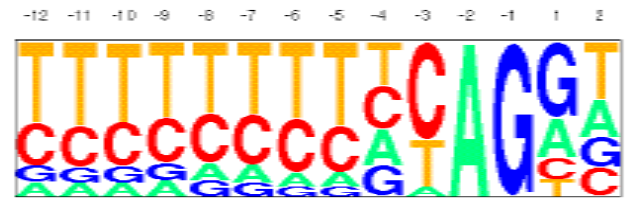
5' 剪切位点



分支点




3' 剪切位点



# 权矩阵模型 II

5' 端剪接信号



Con: C A G ... G T

可能性	-3	-2	-1	...	+5	+6
A	0.3	0.6	0.1	...	0.1	0.1
C	0.4	0.1	0.0	...	0.1	0.2
G	0.2	0.2	0.8	...	0.8	0.2
T	0.1	0.1	0.1	...	0.0	0.5

背景

可能性	普通的
A	0.25
C	0.25
G	0.25
T	0.25

$$S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9$$

概率系数

$$R = \frac{P(S|+) = P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \cdots P_5(S_8)P_6(S_9)}{P(S|-) = P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3) \cdots P_{bg}(S_8)P_{bg}(S_9)}$$

背景同源模型，假设为独立

# 权矩阵模型 III

$$S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9$$

概率系数

$$R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \cdots P_5(S_8)P_6(S_9)}{P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3) \cdots P_{bg}(S_8)P_{bg}(S_9)}$$
$$= \prod_{k=1}^{k=9} P_{-4+k}(S_k) / P_{bg}(S_k)$$

得分

$$s = \log_2 R = \sum_{k=1}^{k=9} \log_2 (P_{-4+k}(S_k) / P_{bg}(S_k))$$

Neyman-Pearson 定理:

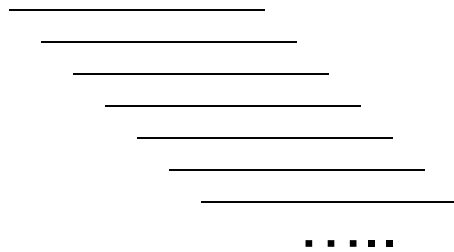
最优化判别规则的形式:  $R > C$

因为  $\log$  是单调函数,  $\log_2(R) > C'$

# 权矩阵模型 IV

- 沿序列滑动 WMM

ttgacctagatgagatgtcgttcacttttactgagctacagaaaa



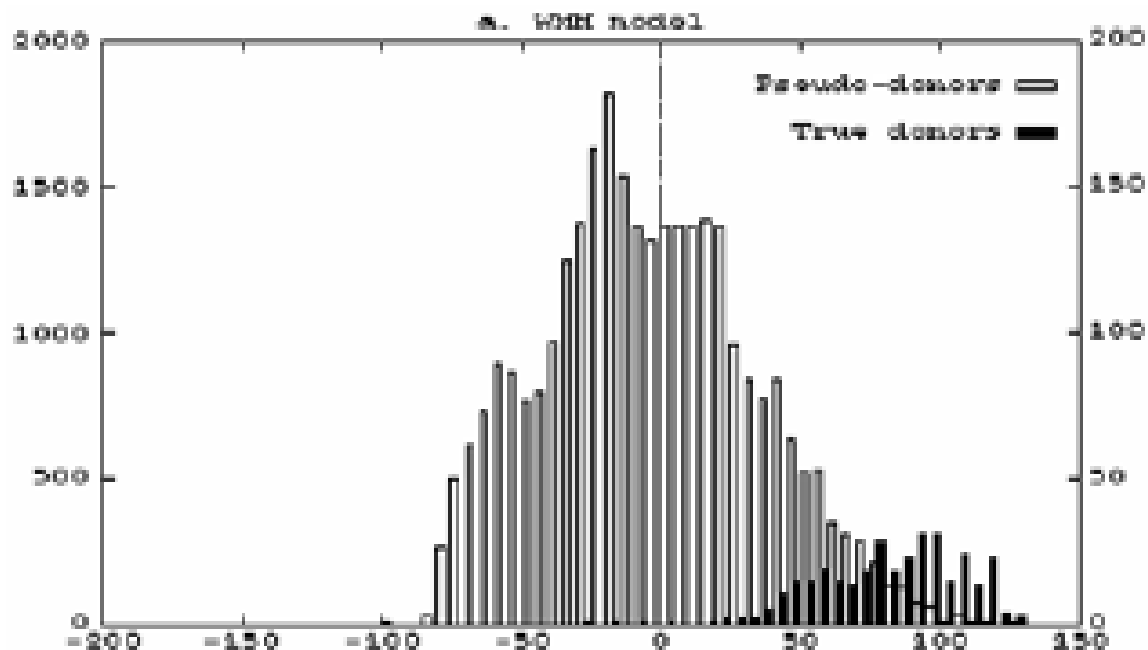
为每个 9 碱基窗口赋分值

用临界值预测 5' 端潜在剪接位点



# 5' 剪接位点柱状图

“Decoy”  
5' 剪接  
位点



真实  
5' 剪接  
位点

得分 ( 1/10 比特单位)

测量精度:

灵敏度: 真实位点 w/ score 百分数 > 临界值

特异性: 位点 w/ score 百分数 > 临界值  
为真实位点

Sn	20%	50%	90%
Sp	50%	32%	7%

# 这些结果告诉我们什么？

**A** ) 剪切机器也使用除 **5'** 剪接位点模体以外的信息来确定剪接位点

或

**B** ) 权矩阵模型不能准确捕捉到 **5'** 剪接位点用以识别的某些特征

(或都)

这是生物学中常常出现的情况

# 什么是 DNA、RNA 模体

一组拥有共同性质的 DNA 或 RNA 序列常常出现的模式，如调节蛋白结合位点。

常见模体形容词：

精密、精确 与 退化

强 与 弱（好 与 差）

高信息量 与 低信息量

# 信息论

- 我们以 **Shannon's** 著名的公式结束

$$H = - \sum_{i=1}^{20} P_i (\log_2 P_i)$$

其中 **H** = 比对中每个位点包含的比特 “信息熵”

这表示什么？

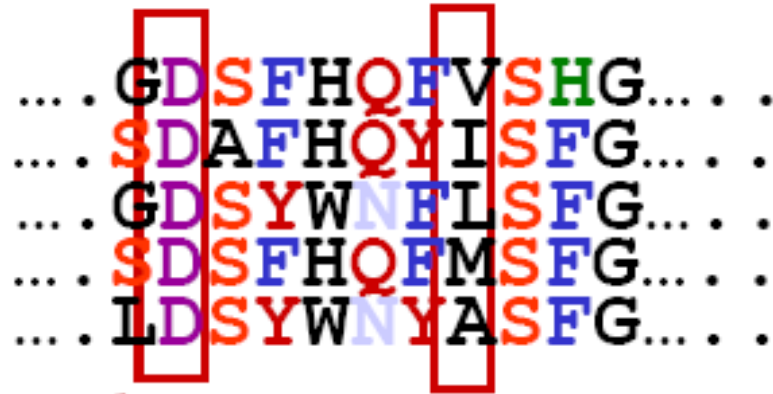
H 是熵或随机性或无序性的度量

…… 它告诉我们在模体某一位置有不同氨基酸丰度的不确定度

This slide courtesy of M. Yaffe

# 信息论

Courtesy of M. Yaffe



假设 20 种氨基酸有同样的可能性:  $H_{\text{before}} = 4.32$ ,  $H_{\text{after}} = 0$

因此, 该位点编码信息量为  $4.32 - 0 = 4.32$ ,

模体中另一个位置包含 20 种氨基酸:  $H_{\text{before}} = 4.32$ ,  $H_{\text{after}} = 4.32$

因此, 该位点编码信息量为  $4.32 - 4.32 = 0$

# DNA 模体的信息含量

- 位点  $j$  所包含的信息:  $I_j = H_{\text{before}} - H_{\text{after}}$
- 模体概率:  $p_k$  ( $k = A, C, G, T$ )
- 背景概率:  $q_k = 1/4$  ( $k = A, C, G, T$ )

$$I_j = -\sum_{k=1}^4 q_k \log_2 q_k - -\sum_{k=1}^4 p_k \log_2 p_k = 2 - H_j$$

$$I_{\text{motif}} = \sum_{j=1}^w I_j = 2w - H_{\text{motif}} \text{ (motif of width } w \text{ bases)}$$

Log base 2 gives entropy/information in  
'bits'

# 模体的平均比特得分

- 比特分数:  $\log_2 (p_k/q_k)$
- 平均比特分数: (模体宽度  $w$ ,  $n = 4^w$ ,  $q_k = 1/4^w$ )

$$\sum_{k=1}^n p_k \log_2 \left( \frac{p_k}{q_k} \right) = 2w - H_{\text{motif}} = I_{\text{motif}}$$

经验规律\*：每  $2^m$  随机序列有  $w/m$  比特的模体信息

\* 在常规表达中大约符合，在其他模体中符合

# 模体搜索的问题

未比对

Cgggcactagcccatgtgagagggcaaggaccagcggaa  
gtaattcagggccaggatgtatctttctcttaaaaataacatatcct  
acagatgatgaatgcaaatcagcgtcacgagcttggcgggc  
aagggtgcttaaaagataatatcgaccctagcgattcgggtacc  
gttcataaaagtacgggaatttcgggtaggttatgttaggcgag  
ggcaaaagtcataacttttaggtcaagagggcaatgcctcctc  
tgccgattcggcgagtgatcggtgggaaaatatgagacca  
ggggagggccacactgcagctgccgggctaacagacaca  
cgtctagggctgtgaaatctgtaggcgccgaggccaacgctg  
agtgtcgatgttgagaacattagtcgggtccaagagggcaac  
ttgtatgcaccgccgcggcccagtgcgcaacgcacagggc  
aaggttactgcggccacatgcgagggcaacctccctgtgtg  
ggcggttctgagcaattgtaaaacgacggcaatgttcgggtcgc  
ctaccttgataaagaggggggtaggaggtcaactcttccgt  
attaataggagtagagtagtggttaaactacgaatgcttataac  
atgcgagggcaatcgggatctgaaccttctttatgcgaagactc  
caggaggaggtcaacgactctgcatgtctgacaacttggtcat  
agaattccatccgccacgcggggtaatttgacgtgtgccaac  
ttgtgccgggggggctagcagcttcccgtaaacgcggttgag  
tgcaaacatacacagcccgggaatatagaaagatacagattc  
gatttcaagagttcaaaacgtgacggggacgaaacgagggc  
gatcaatgcccgataggactaataagtagtacaacccgctc  
acccgaaaggagggcaaatacctatatacagccaggggag  
acctataactcagcaagggtcagcgtatgtactaattgtggaga  
gcaaatcattgtccacgtg

---

已比对

Gcggagaggggcactagcccatgtgagagggcaaggacca  
atctttctcttaaaaataacataattcagggccaggatgtgtcacg  
agctttatcctacagatgatgaatgcaaatcagctaaaagataat  
atcgaccctagcgtggcgggcaagggtgctgtagattcgggtac  
cgttcataaaagtacgggaatttcgggtatacttttaggtcgttatgtt  
aggcgagggcaaaagtcactctgccgattcggcgagtgatcg  
aagagggcaatgcctcaggatggggaaaatatgagaccagg  
ggagggccacactgcacacgtctagggctgtgaaatctctgcc  
gggctaacagacgtgtcgatgttgagaacgtaggcgccgagg  
ccaacgctgaatgcaccgccattagtcgggtccaagagggc  
aactttgtctgcgggcggcccagtgcgcaacgcacagggcaa  
ggtttatgtgttgggcgggtctgaccacatgcgagggcaacctcc  
cgtcgcctaccctggcaattgtaaaacgacggcaatgttcgcgt  
attaatgataaagaggggggtaggaggtcaactcttcaatgctta  
taacataggagtagagtagtggttaaactacgtctgaaccttcttt  
atgcgaagacgcgagggcaatcgggatgcatgtctgacaactt  
gtccaggaggaggtcaacgactccgtgtcatagaattccatcc  
gccacgcggggtaatttgatcccgtcaaagtgccaacttgtgc  
cgggggggctagcagctacagcccgggaatatagacgcgttg  
gagtgc aaacatacacgggaagatacagattcgaattcaagag  
ttcaaaacgtgcccgataggactaataaggacgaaacgaggg  
cgatcaatgttagtacaacccgctcacccgaaaggagggca  
aatacctagcaaggttcagatatacagccaggggagacctata  
actcgtccacgtgcgtatgtactaattgtggagagcaaatcatt

---



# 模体搜索例子： Gibbs 采样

Gibbs 采样是一种蒙特卡罗方法，可以从输入序列数据中搜索最大似然率函数。  
在 A 位置有模体的序列 s 的似然率函数

权矩阵                      背景频率矢量

$$P(S, A | \Theta, \theta_B) =$$

$$\theta_{B,a} \times \dots \times \theta_{B,a} \times \Theta_{1,t} \times \Theta_{2,a} \times \dots \times \Theta_{8,c} \times \theta_{B,t} \times \dots \times \theta_{B,t}$$

s=

“actactg**tatcgt**actgactgattaggccatgactgcat”

模体位点 A

准备好

# Gibbs抽样多媒体体验

Gibbs 采样算法的描述:

- 图片
- 文字
- 视频

# Gibbs 采样算法 I

## 1. 在每条序列中选择随机位点

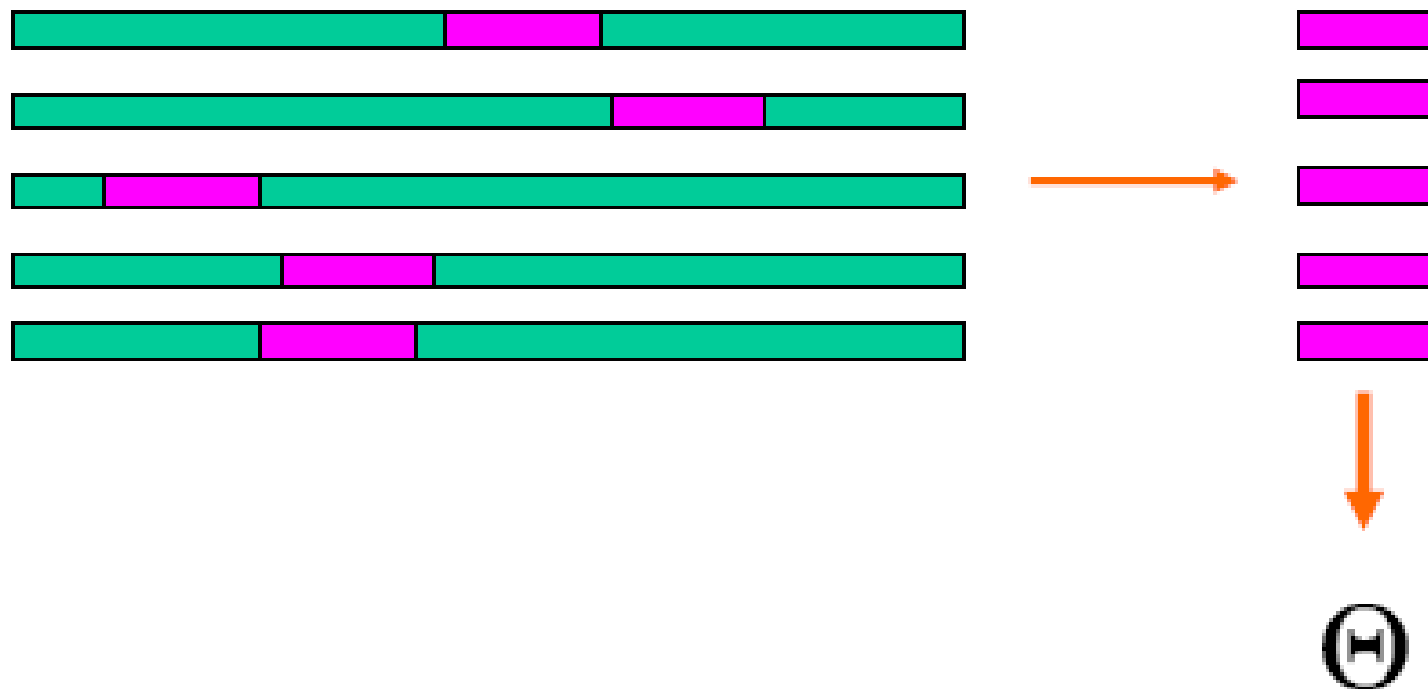
序列组

模体例子



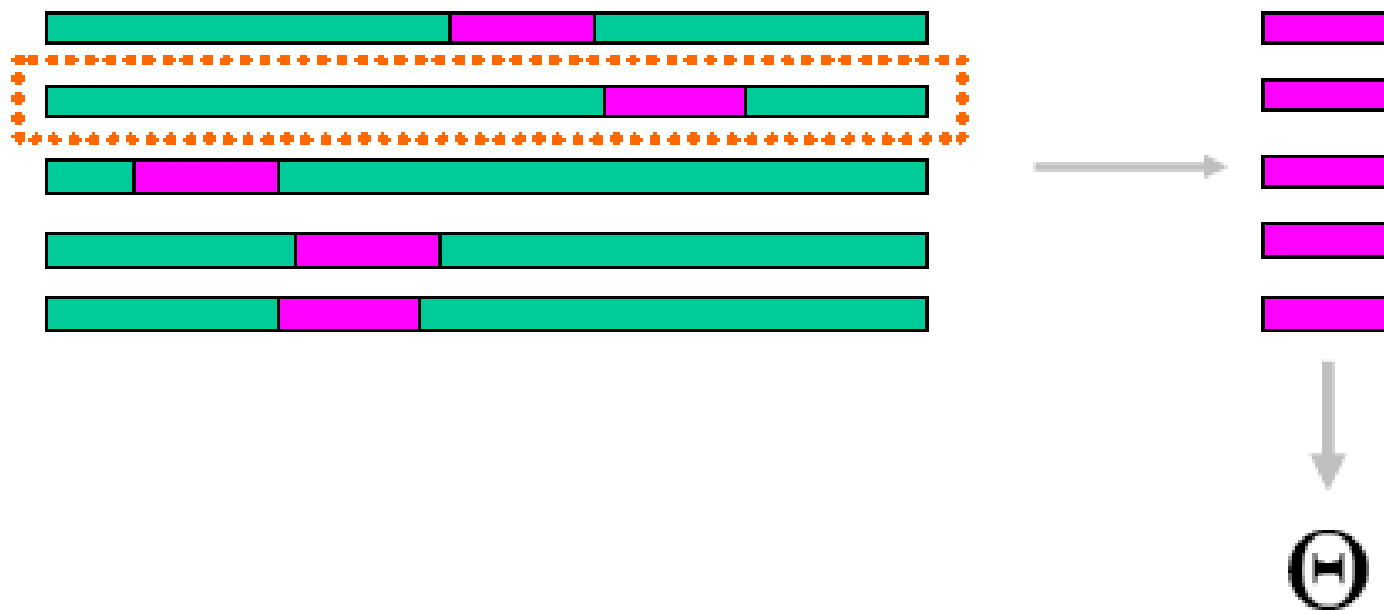
# Gibbs 采样算法 II

## 2. 建立权矩阵



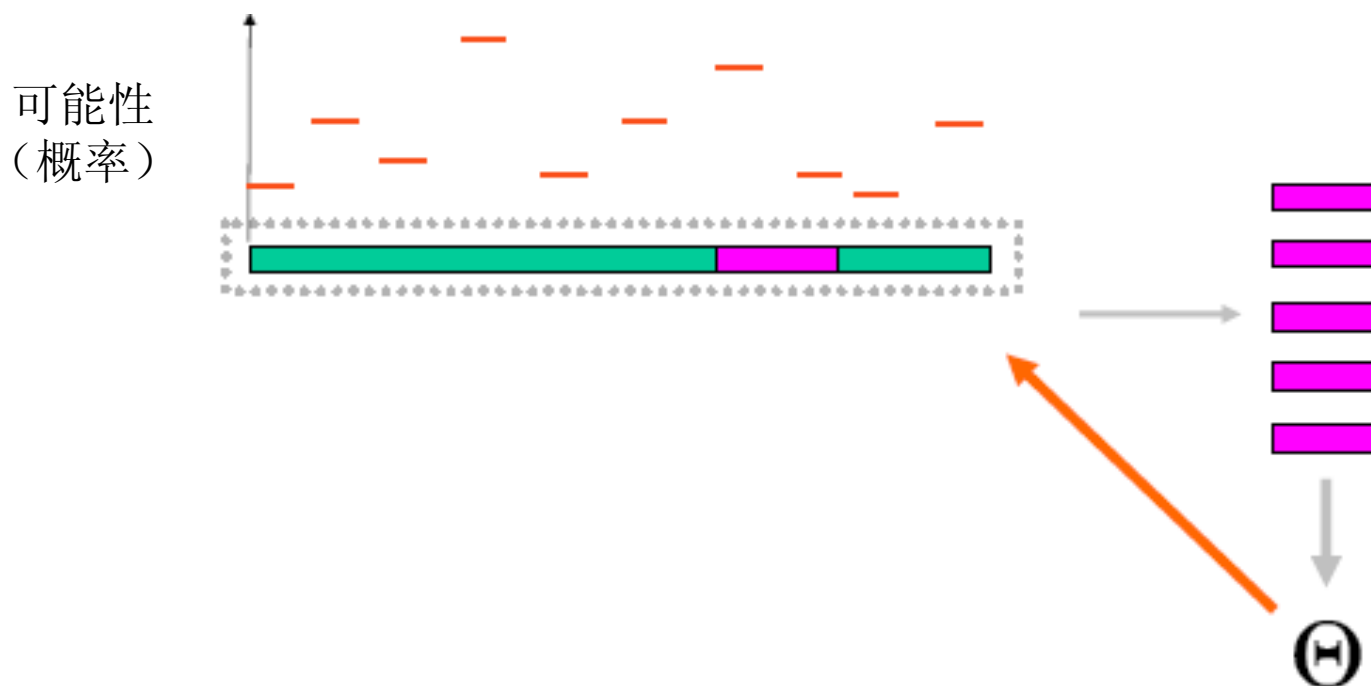
# Gibbs 采样算法 III

## 3. 随机选择一条序列



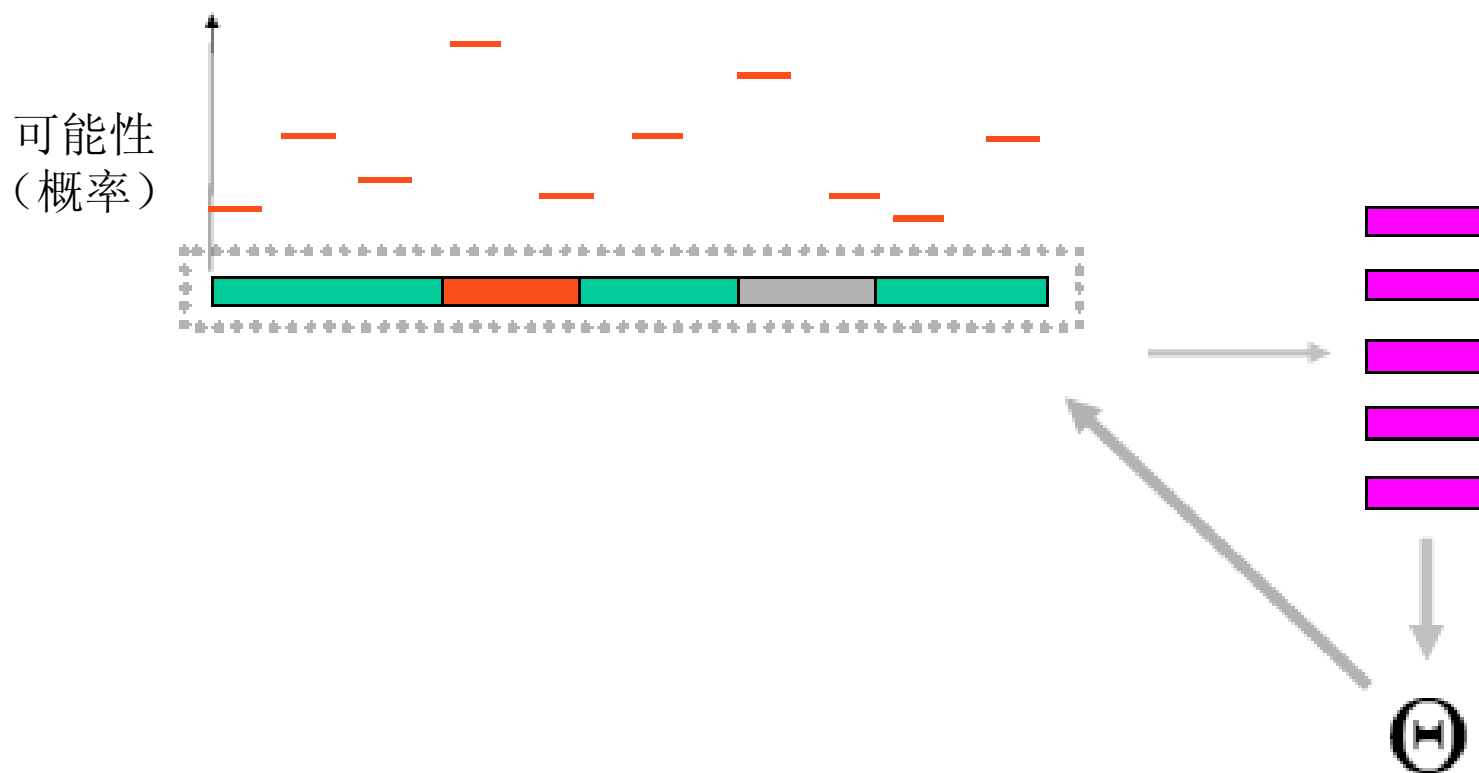
# Gibbs 采样算法 IV

## 4. 用权矩阵给序列中可能的位点打分



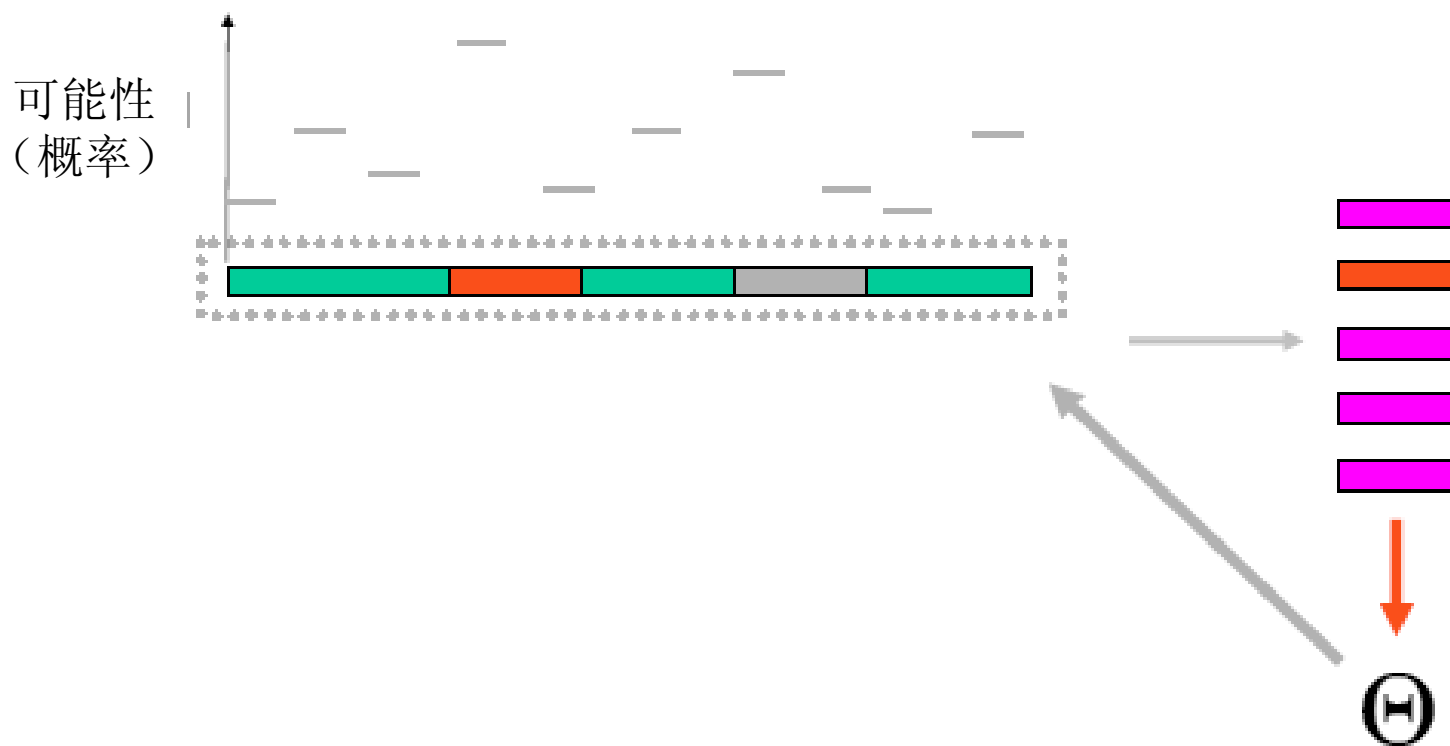
# Gibbs 采样算法 V

## 5. 抽样一具有相似可能性的位点



# Gibbs 采样算法 VI

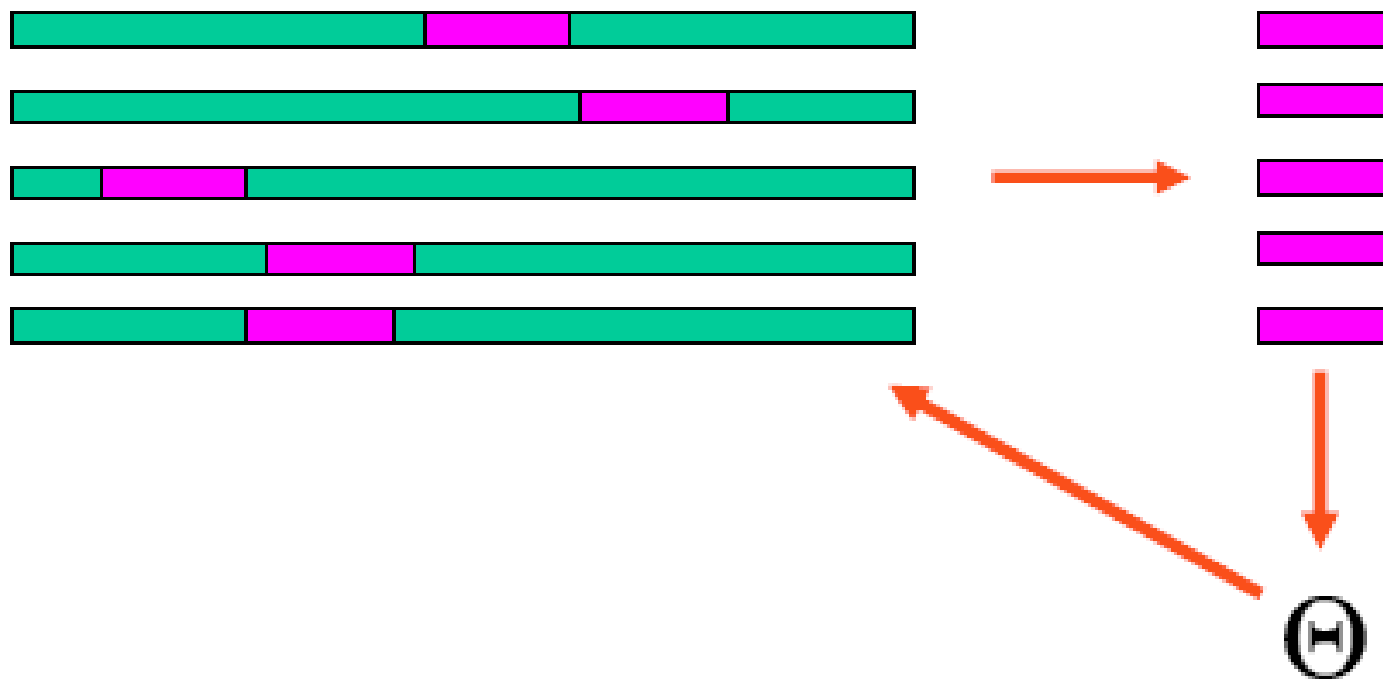
## 6. 升级权矩阵





# Gibbs 采样算法 VII

## 7. 迭代至收敛 (位点 $i|\Theta$ 不改变)



# Gibbs 采样算法文字描述 I

假设有宽度  $W$  期望模体，长度为  $L$  的序列  $N$ ：

步骤 1 ) 在每条序列中选择随机位点：序列 1  $a_1$ ，序列 2  $a_2$ ， $\dots$ ，序列  $n$   $a_n$ 。

步骤 2 ) 在序列组中随机选择序列（比如，序列 1）。

步骤 3 ) 为所有序列中宽度  $W$  的位点建立权矩阵，第 2 步中选中的序列除外。

步骤 4 ) 用第三步中建立的权矩阵为序列 1 中每个位点设置概率： $p = \{ p_1, p_2, p_3, \dots, p_{L-W+1} \}$

# Gibbs 采样算法文字描述 II

假设有宽度  $W$  期望模体，长度为  $L$  的序列  $N$

步骤 5 ) 根据该概率分布在序列 1 中抽样起始点，设该新位点为  $a_1$ 。

步骤 6 ) 从序列组中随机选取序列（比如说，序列 2）。

步骤 7 ) 为所有序列个位点建立宽度  $W$  的权矩阵模型，第 6 步中选中序列除外。

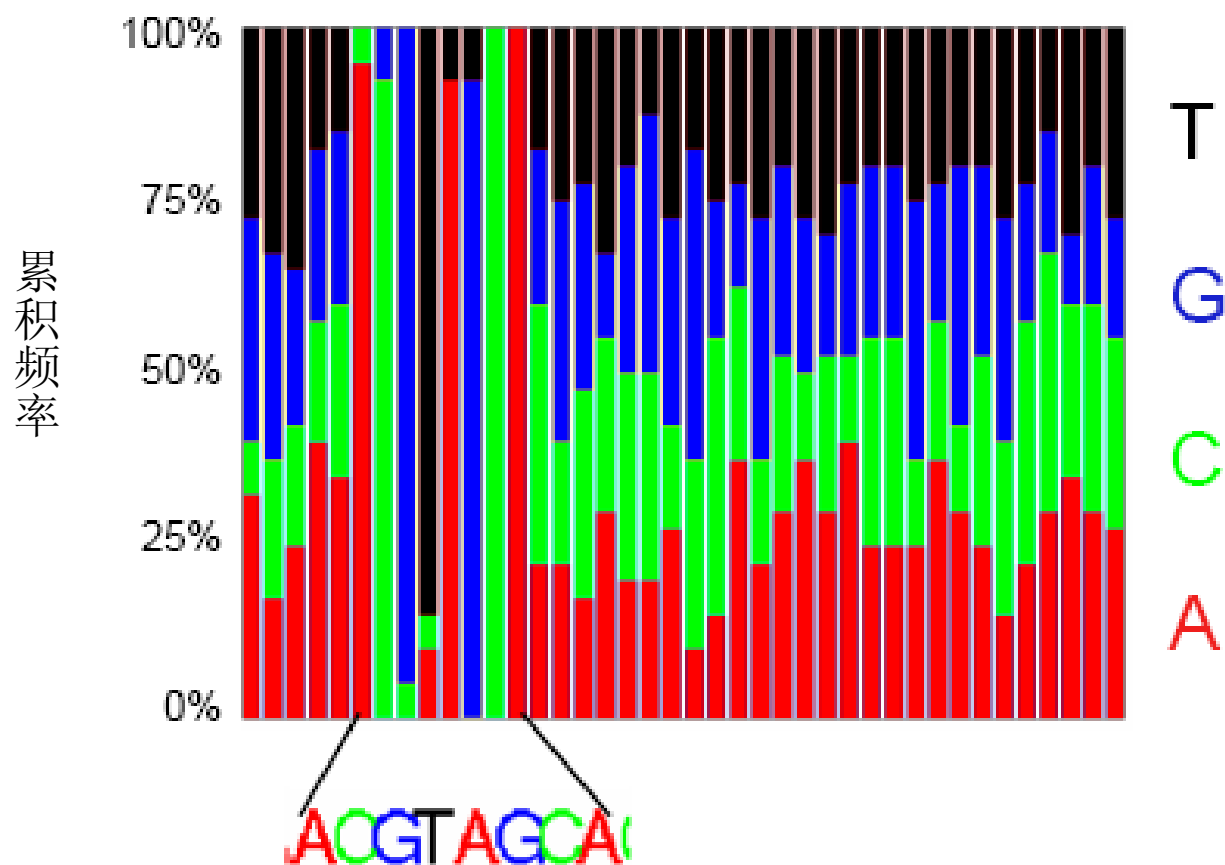
步骤 8 ) 用第七步建立的权矩阵，为序列 2 中每个位点赋予概率

步骤 9 ) 按照该 **dist**, 为序列 2sample 起始点

步骤 10 ) 重复直至收敛

如果有，这个算法实现了什么（除了让我们的电脑忙于运算外）？

# 强模体序列输入



# 输入序列（弱模体）

gcggaagagggcactagcccatgtgagagggcaaggacca  
atctttctcttaaaaataacataattcagggccaggatgtgtcac  
gagctttatcctacagatgatgaatgcaaatacagctaaaagat  
aatatcgaccctagcgtggcgggcaaggtgctgtagattcgggt  
accgttcataaaaagtacgggaatttcggtatacttttaggtcggtat  
gttaggcgagggcaaaaagtcactctgccgattcggcgagtgat  
cgaagagggcaatgcctcaggatggggaaaatatgagacca  
ggggagggccacactgcacacgtctagggctgtgaaatctctg  
ccgggctaacagacgtgtcgatgttgagaacgtaggcgccga  
ggccaacgctgaatgcaccgccattagtcgggtccaagagg  
gcaactttgtctgcgggcggcccagtgcgcaacgcacagggc  
aaggtttatgtgttgggcggttctgaccacatgcgagggcaacct  
cccgtcgcctaccctggcaattgtaaaacgacggcaatgttcg  
cgtattaatgataaagaggggggtaggaggtcaactcttcaatg  
cttataacataggagtagagtagtgggtaaactacgtctgaacc  
ttctttatgcgaagacgcgagggcaatcgggatgcatgtctgac  
aacttgtccaggaggaggtcaacgactccgtgtcatagaattc  
catccgccacgcggggtaatttgatcccgtcaaagtgccaac  
ttgtgccgggggggtagcagctacagcccgggaatatagacg  
cgtttgagtgcaaacatacacgggaagatacagagttcgatttc  
aagagttcaaaacgtgcccgataggactaataaggacgaaa  
cgagggcgatcaatgttagtacaacccgctcacccgaaagg  
agggcaaatacctagcaaggttcagatatacagccagggga  
gacctataactcgtccacgtgcgtatgtactaattgtggagagc  
aatcatt

# Gibbs 采样概要

- 模体发现的随机（蒙特卡罗）算法
- 在少量模体实例上进行反复计算，偏离权重矩阵，抽样更多的模体实例，进一步偏离权重矩阵，... 直至收敛
- 对结果进行比较，多次运算中，并不能保证每次都收敛于统一模体
- 用于 **DNA**， **RNA** 和蛋白质模体

# 模体识别算法（MEME）— Multiple EM for Motif Elicitation

- 是另一种流行的模体搜索算法—用 EM(expectation maximization) 算法优化近似的似然函数
- 不同于 Gibbs 抽样，MEME 具有确定性。

Bailey & Elkan, Proc. ISMB, 1994



# 权矩阵模型 II

5' 端剪接  
信号



Con:

C A G ... G T

可能性	-3	-2	-1	...	+5	+6
A	0.3	0.6	0.1	...	0.1	0.1
C	0.4	0.1	0.0	...	0.1	0.2
G	0.2	0.2	0.8	...	0.8	0.2
T	0.1	0.1	0.1	...	0.0	0.5

背景

可能性	普通
A	0.25
C	0.25
G	0.25
T	0.25

$$S = S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8 S_9$$

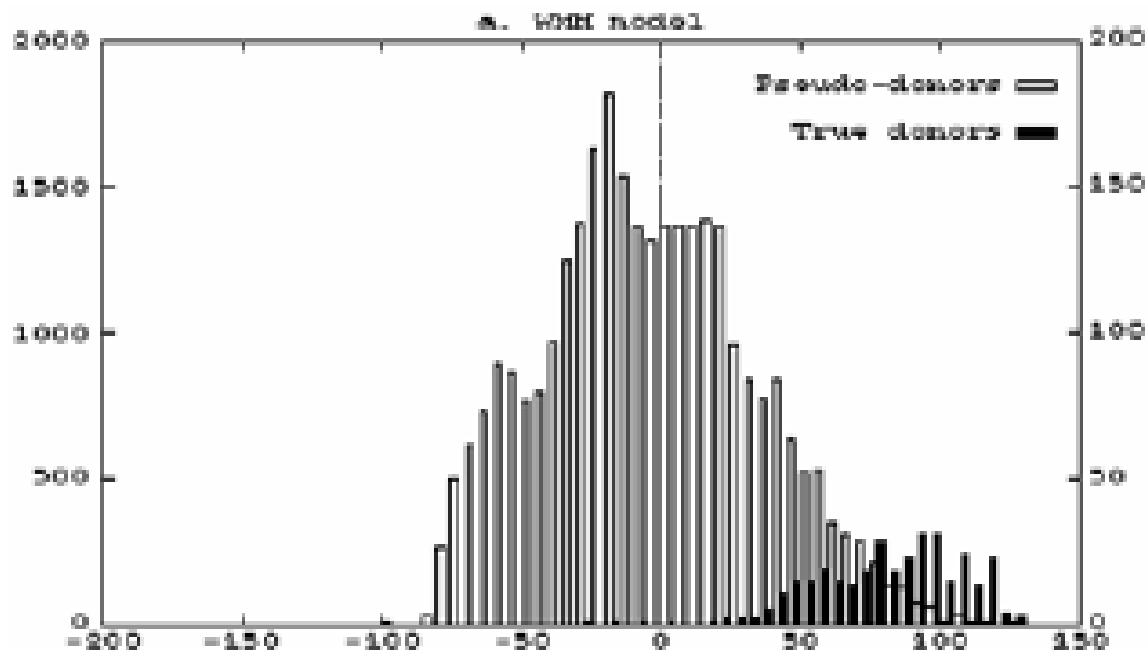
概率系数：

$$R = \frac{P(S|+) = P_{-3}(S_1)P_{-2}(S_2)P_{-1}(S_3) \cdots P_5(S_8)P_6(S_9)}{P(S|-) = P_{bg}(S_1)P_{bg}(S_2)P_{bg}(S_3) \cdots P_{bg}(S_8)P_{bg}(S_9)}$$

背景模型同源物，假设具独立性

# 5' 剪接位点柱状图

“Decoy”  
5' 剪接  
位点



真实  
5' 剪接  
位点

得分 ( 1/10 比特单位)

测量精度:

灵敏度: 真实位点 w/ score 百分数 > 临界值

特异性: 位点 w/ score 百分数 > 临界值  
为真实位点

Sn	20%	50%	90%
Sp	50%	32%	7%

# 这个结果说明什么？

A ) 剪接体也使用除 5' 剪接位点模体以外的信息来确定剪接位点

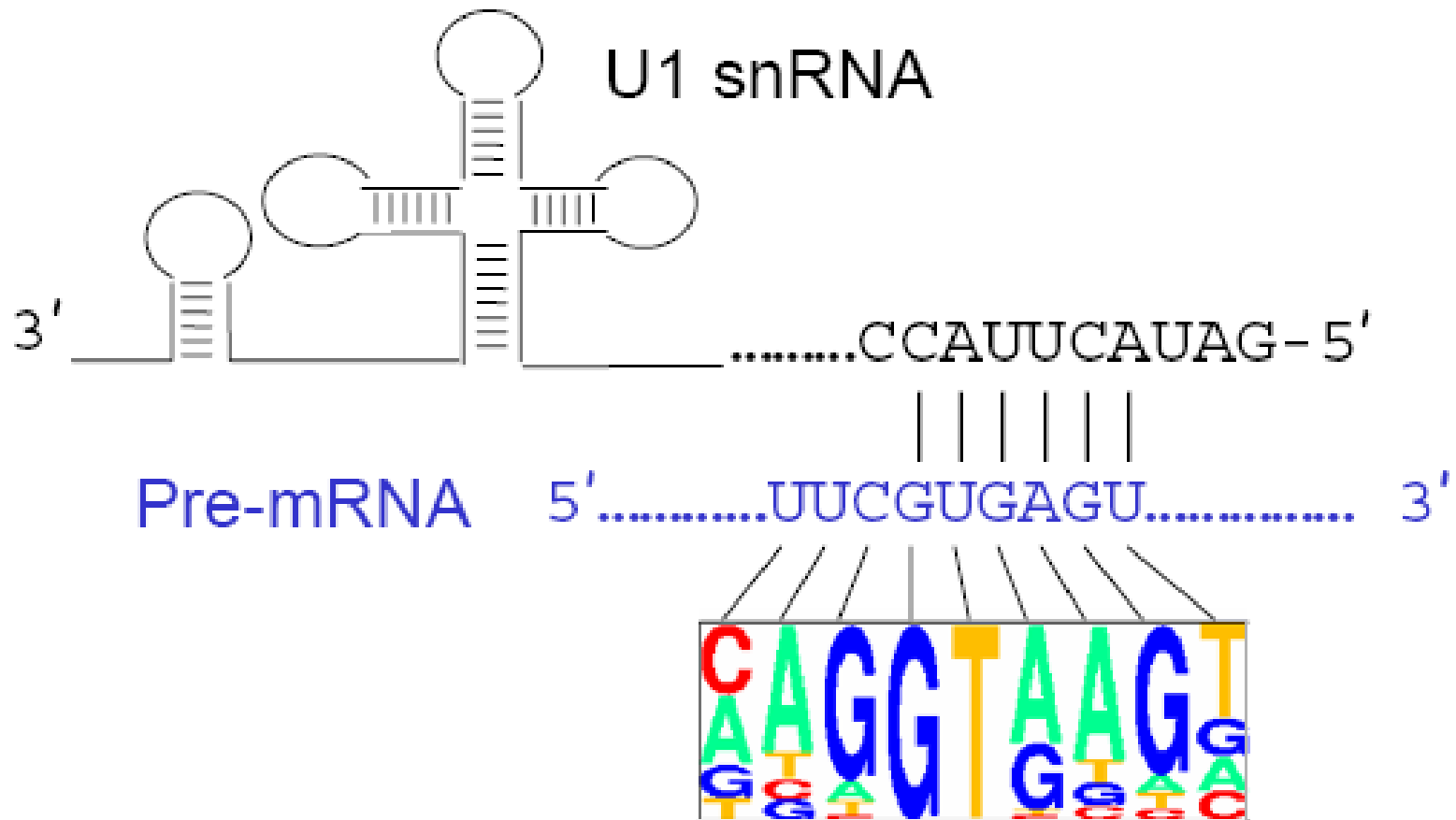
或者

B ) 权矩阵模型不能精确的捕捉到在识别中气作用的 5' 剪接位点的某些信息

(或二者都有)

这是生物学中常见的问题

# 5' 剪接位点是如何识别的

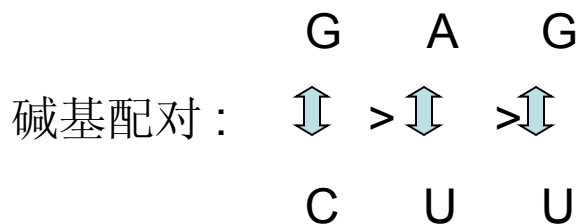


# RNA 热力学

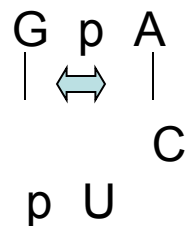
I

螺旋结构自由能

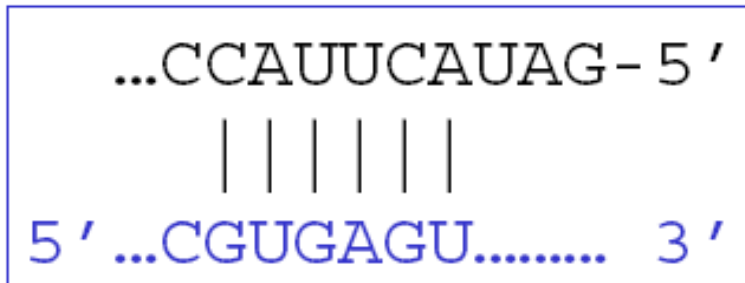
推导:



碱基堆积:

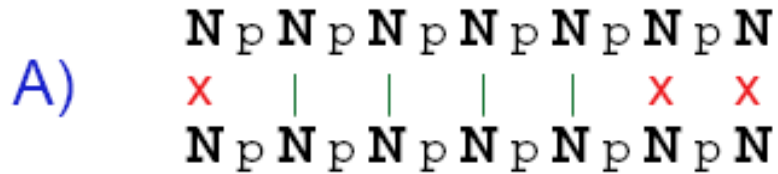


Doug Turner 能量规则:

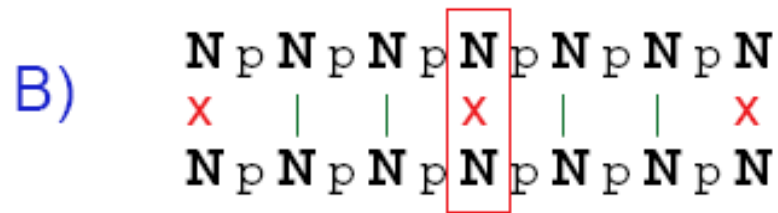


		5' --> 3'		
		UX		
		AY		
		3' <-- 5'		
			X	
<u>Y</u>	A	C	G	U
A	.	.	.	-1.30
C	.	.	-2.40	.
G	.	-2.10	.	-1.00
T	-0.90	.	-1.30	.

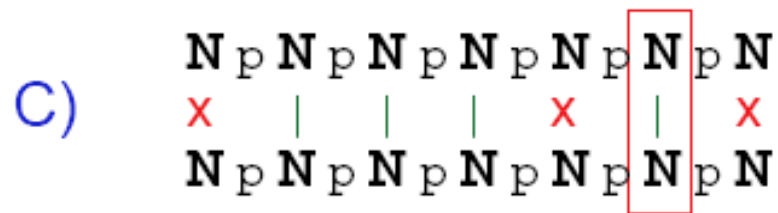
# RNA 热力学 II



多个连续碱基—好



内部环—不好



末端碱基配对不  
稳定—不好

通常，A 比 B 和 C 稳定

# 5' 剪接位点的条件频率



5' 剪接位点 +5 位有 G

可能性	-1		+3	+4	+6
A	9		44	75	14
C	4		3	4	18
G	78		51	13	19
T	9		3	9	49

5' 剪接位点 +5 位无 G

可能性	-1		+3	+4	+6
A	2		81	51	22
C	1		3	28	20
G	97		15	9	30
T	0		2	12	28

数据源自: Burge, 1998 "Computational Methods in Molecular Biology"

哪种模型可以使位点间相互作用得以结合？