生物信息学

第六章 序列模式识别（1）

生物信息学：预测

- 生物信息学最核心的问题：预测
- 生物信息学工具的作用：预测
- 生物信息学所有的分析：预测
- 基本假设（贝叶斯的哲学理念）：我们能够通过对已知世界的观察，总结经验，并以此来预测未知世界已经存在或者即将发生的事物/事件
- 在生物信息学中的应用：对现有的数据，使用合适的算法，进行训练，构建计算模型和计算工具，预测未知的现象

序列模式

- 功能结构域，functional domain
- 模体，motif
- 模块，BLOCK
- 模式，pattern/profile

功能结构域/Domain

- 具有完整的、独立的三级结构
- 具有特定的生物学功能
- 一般长度，几十到几百个氨基酸
- 允许插入/缺失，即允许存在gap

```
KDC2_DROME/1-5    KHLNWNDVYS----KKLKPPILPDVH--------HDGDTKNFD-DYPEKDWKP-------
CONSENSUS/80%     ttlsWptl.......ph.sPhhP.lp........s..Dhp.FD.thspt...........
CONSENSUS/65%     pslsWcpl.p....pplpPPahPplp........s.tDsssFD.casppts.........
CONSENSUS/50%     +sIDW-cLpp....+clpPPFpPplp........uspDsoNFD.-FTcpsss........


KDC2_DROME/1-5    -AKAVDQ----------------------RDLQYFNDF
CONSENSUS/80%     ..p............................a.shsh..
CONSENSUS/65%     .hssssp...t...................Ftuasaht
CONSENSUS/50%     .hosssssshhpshppp.............FtGFoass
```

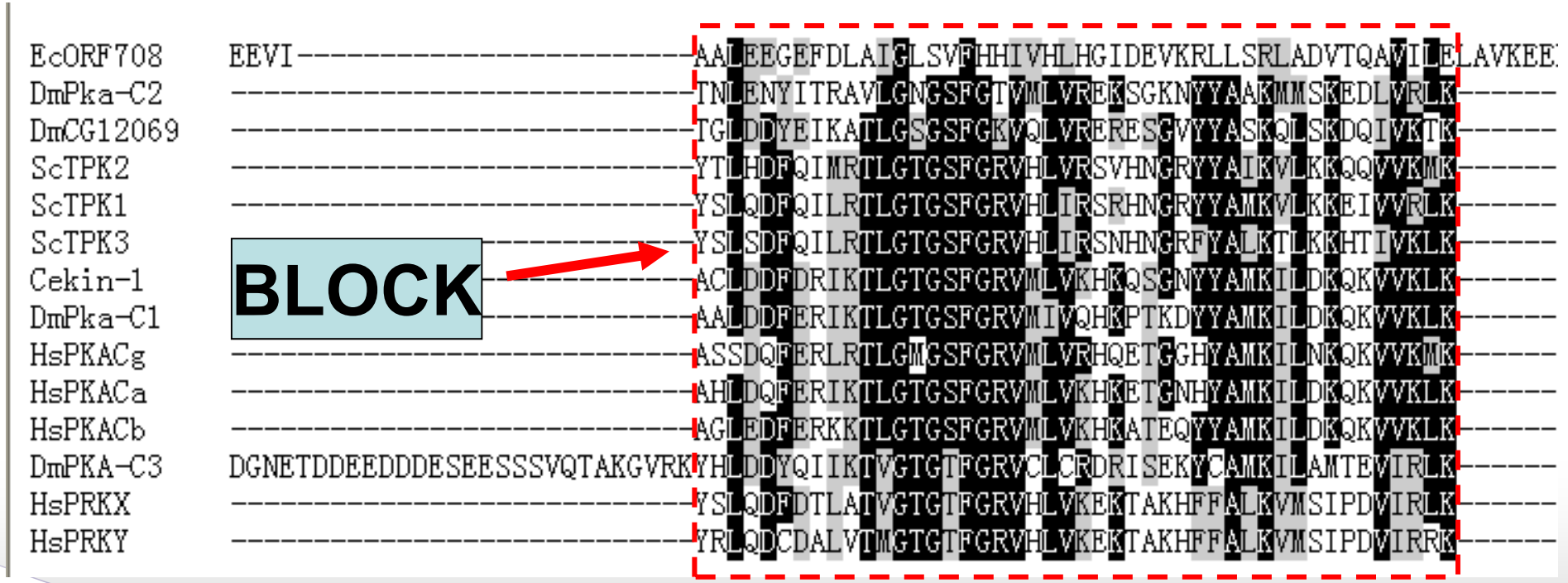
模体/Motif

- 不具有独立的三级结构
- 具有特定的生物学功能：结合，修饰，细胞亚定位，维持结构，等
- 长度一般几个到几十个氨基酸或者碱基；
- 例如，SUMO化的序列模体：Ψ-K-X-E (Ψ:A, I, L, V, M, F, P; X: 任意氨基酸)

Functional site class:	Sumoylation site
Functional site description:	Motif recognised for modification by SUMO-1
ELM(s):	MOD_SUMO
***MOD_SUMO* description:**	Motif recognised for modification by SUMO-1
Pattern:	[VILMAFP]K.E

模块/BLOCK

- 几个到几十个氨基酸
- 无gap，从全局多序列比对的结果直接处理得到
- 描述蛋白质家族或者一类蛋白质的序列保守性



模式/Pattern/Profile

- 在算法上用来描述一类功能结构域，模体或者模块的表示方式
- 根据序列数据，构建的预测模型
- 数据形式：概率表示
- 用来预测新的可能符合特定模式的序列
- 例如，直接将Ψ-K-X-E视为SUMO化位点的，普适的“模式”，则可以预测所有包含该模式的蛋白质序列

本章内容提要

- 预测性能检验和评估
- 位点特异性打分矩阵/权重矩阵模型
  - Position Specific Scoring Matrix (PSSM), Weight Matrix Model (WMM)
- 模体发现：Gibbs Sampler等
- 马尔科夫及隐马尔科夫模型
- 翻译后修饰位点预测
- 模式识别的其他算法简介

预测性能的计算和检验

- 样本/检验数据：阳性数据 (P)，阴性数据 (N)
  - 阳性数据 (P)：真实的，被实验所证实的数据
  - 阴性数据 (N)：被实验所证明为无功能的数据
- 对于预测结果的评测，定义：
  - 真阳性 (TP): 阳性数据中被预测为阳性的数据
  - 假阳性 (FP): 阴性数据中被预测为阳性的数据
  - 真阴性 (TN): 阴性数据中被预测为阴性的数据
  - 假阴性 (FN): 阳性数据中被预测为阴性的数据

常用的检验指标

- 灵敏度 (Sensitivity, *Sn*): 对于真实的数据，能够预测成“真”的比例是多少 - (Type II error)
- 特异性 (Specificity, *Sp*): 对于阴性的数据，能够预测成“假”的比例是多少 - (Type I error)
- 准确性 (Accuracy, *Ac*): 对于整个数据集(包括阳性和阴性数据)，预测总共的准确比例是多少
- 马修相关系数(Mathew correlation coefficient, *MCC*): 当阳性数据的数量与阴性数据的数量差别较大时，能够更为公平的反映预测能力，值域[-1,1]

常用的检验指标

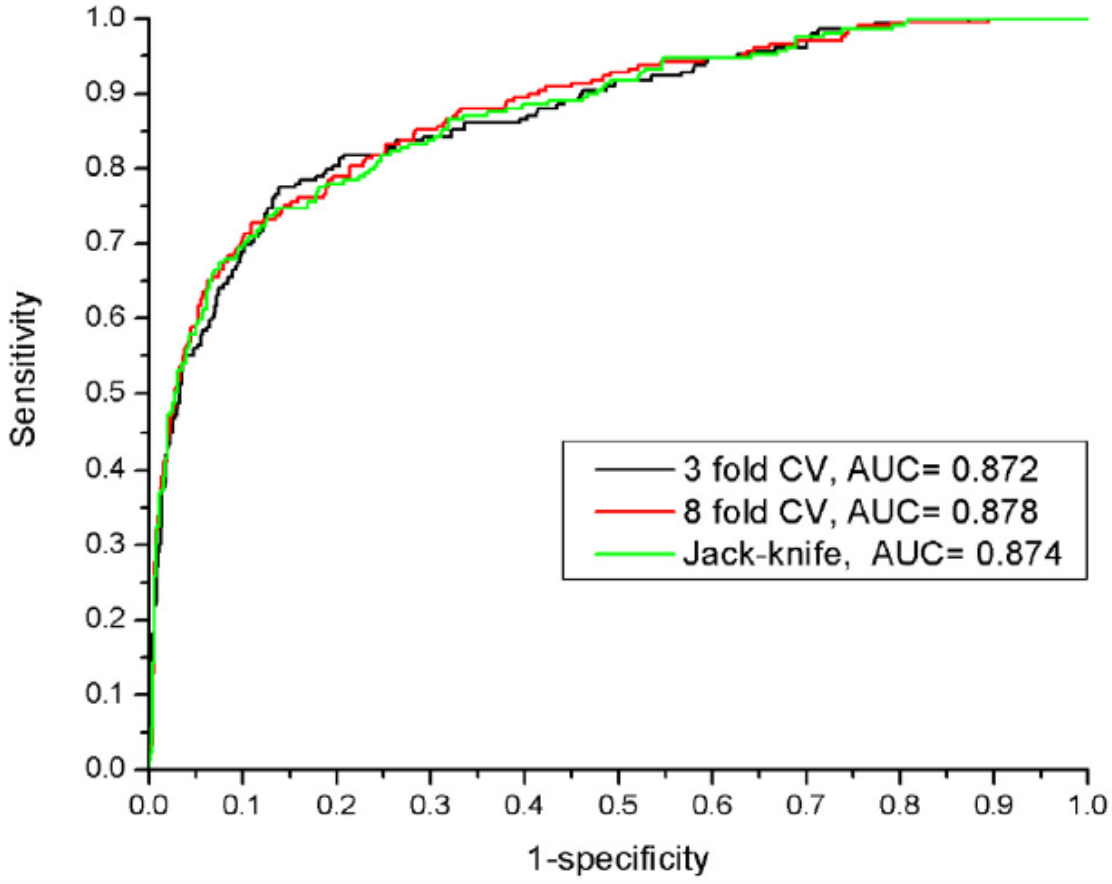
$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN},$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

ROC curve

- X轴：1-*Sp*
- Y轴：*Sn*
- ROC的面积越大，表明其预测能力越强



预测性能的计算

- 自适应法/自检法（Self-consistency validation）
  - 训练数据当成测试数据
  - 训练数据中所有的阳性数据为测试数据中的阳性数据
  - 训练数据中所有的阴性数据为测试数据中的阴性数据
- 反映当前预测工具对目前已知的数据的预测能力
- 假设：根据目前已知的数据所构建的计算模型能够反映未知的数据的模式
- 缺点：不能反映计算模型的稳定性

预测性能的检验 (1)

- 除一法/留一法（Leave-one-out validation）
  - 每次从数据集中去掉一个，包括阳性数据和阴性数据
  - 利用剩下的数据重新训练，并构建新的计算模型
  - 对去掉的这一个数据进行打分
  - 保证每个数据去掉一次，从而得到所有数据的分值
  - 计算各个阈值的*Ac*, *Sn*, *Sp*和*MCC*

预测性能的检验 (2)

- N折交叉法（*n*-fold cross-validation）
  - 将数据集分成*n*组，并保证阳性数据与阴性数据的比例与原数据相同
  - 随意将*n-1*组作为训练数据，重新训练并构建计算模型
  - 对剩下的1组进行打分，计算性能
  - 重复若干次 (一般20次或以上足够)
  - 计算平均值

预测性能及稳定性

- 自适应法/自检法: 反映预测性能
- 除一法/留一法 & N折交叉法: 反映预测系统的稳定性
- 预测性能 *vs*. 检验性能
  - 差距较小：系统稳定
  - 差距过大：系统不稳定，数据过训练

阈值的确定

- **Threshold 或 Cut-off：**
  - 人为设定，主要依据经验
  - 给定阈值以上或以下预测为阳性
  - 即利用阈值进行“一刀切”
- **确定阈值的一般方法**
  - 传统策略：平衡*Sn*和*Sp*，使两者大致相当
  - 实际应用：高*Sp*低*Sn*保证预测结果的可靠性
  - MCC最大值，保证综合预测性能最高
  - …

过训练（Overfitting/Overtraining）

- 根据已知数据构建的模型只能很好的适用于训练数据
- 不适合用来预测
- 对训练数据的微小改变对于预测性能影响过大
- 预测工具过训练：只能很好的符合训练数据，而对新数据则性能很差

如何评估算法的准确性？

- 例：某预测工具*X*使用400个阳性数据和1600个阴性数据训练计算模型。利用该数据集对软件进行除一法评测时，可预测出450个阳性结果，其中360个包含在已知阳性数据中。则该软件的预测性能是多少？（2017年期末考试题目）

- *Sn* = TP/(TP+FN) = 360/400= 90%
- *Sp* = TN/(TN+FP) = [1600-(450-360)]/1600

= 1510/1600= 94.4%

位点特异性打分矩阵

- Position Specific Scoring Matrix (PSSM)/ Weight Matrix Model (WMM)
- 对蛋白质家族进行多序列比对分析，发现结果中保守的BLOCK
- 根据BLOCK序列推导相应的PSSM
- 不考虑gap的影响
- BLOCK长度一般在几个~几十个残基/碱基

NCBI

NCBI Structures

NCBI Field Guide

This tool was created as part of an NCBI course and is still under development.

PSSM Viewer

Amino Acid Explorer

PSSM Viewer Help

CDD Help

Show Color Key

Course Main Page

Questions or comments

Scores

10 9 8 7 6 5 4 3 2 1 0 -1 -2 -3 -4 -5 -6 -7 -8

cd02249 : Zinc finger, ZZ type. Zinc finger present in dystrophin, ...

Change PSSM/Sequence | Stacked Bar View | Reset | Download Matrix to File | Tutorial

Draw table showing only those positions where the consensus is Any

Draw table showing only this feature: Zinc-binding sites

Hide Scores

Click on any score to compare the two residues.

Click on any column to sort the matrix by that column's scores.

P - consensus sequence position **C** - consensus sequence residue

Master:	1TOT_A	View Master in CD

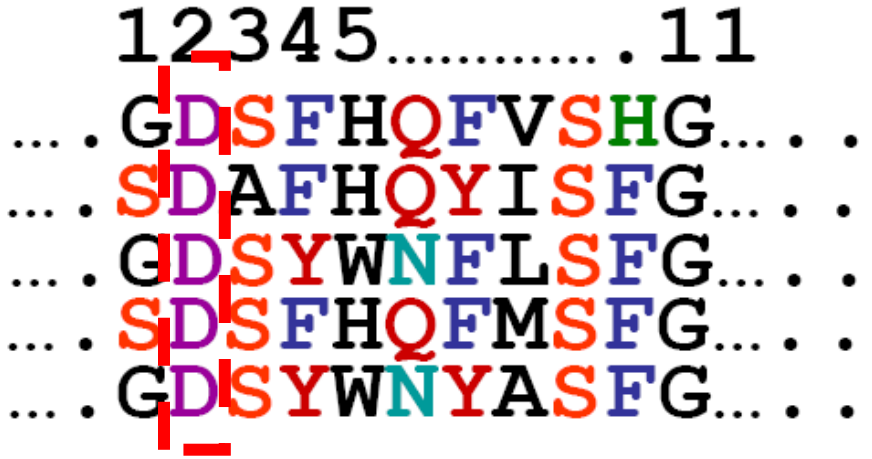
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
1	Y	7 - Y	-3	-5	2	-2	1	1	3	-2	0	-4	0	-3	6	0	-3	-2	-3	1	-4	-4
2	S	8 - T	0	-4	2	-1	0	-2	1	-4	0	1	3	3	-3	-3	-2	-3	0	1	-3	-1
3	C	9 - C	-2	-5	-3	-3	-3	-3	-4	-4	-5	11	-3	-3	-4	-5	-5	-5	-5	-6	-6	-6
4	D	10 - N	-1	-3	-5	-5	-5	-4	-5	-6	-4	-5	-2	-1	-4	4	-2	3	0	-1	6	-1
5	G	11 - E	0	4	1	-2	-1	-3	-4	-4	-4	4	-1	0	0	1	-3	3	-3	-4	-3	0
6	C	12 - C	-2	-5	-3	-3	-3	-3	-4	-4	-5	11	-3	-3	-4	-5	-5	-5	-5	-6	-6	-6
7	L	13 - K	-3	3	-1	2	-1	-2	-4	-4	-4	-4	0	-3	-4	1	2	-3	1	0	-1	-1
8	K	14 - H	-1	1	0	-1	-1	2	-3	-4	-4	-4	1	1	0	1	1	1	2	-2	-3	0
9	P	Gap	-1	0	-3	-1	-2	-3	-3	-4	5	-4	-2	0	1	1	-2	2	-2	-3	1	3
10	I	15 - H	-3	-5	6	1	1	-1	2	-4	-5	-3	-1	-3	-2	-1	-4	2	-4	-4	-5	-4
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E
11	V	16 - V	0	-4	2	-1	1	3	-2	5	1	-3	0	1	2	-1	0	-3	-3	-1	-1	-3
12	G	17 - E	-2	5	-5	-5	-4	-4	-5	-4	2	-4	-2	-3	-4	1	-2	-3	-2	-1	-2	3
13	V	18 - T	-1	-4	0	-1	2	0	0	-4	3	-4	0	0	0	0	-2	1	1	2	-3	1
14	R	19 - R	-3																0	7	-4	-2
15	Y	20 - W	-4																-4	-4	-6	-5
16	H	21 - H	-1																4	2	-2	1
17	C	22 - C	-2																-5	-5	-5	-5
18	L	23 - T	1																1	1	-1	0
19	V	24 - V	-2	-4	0	-3	4	1	-4	-5	-4	-4	0	2	-4	1	0	-3	-1	1	2	1
20	C	25 - C	0	-4	-3	-3	-3	-3	-4	-4	-5	11	-3	-3	-4	-4	-4	-5	-1	-5	-5	-5
P	C	Master	A	G	I	L	V	M	F	W	P	C	S	T	Y	N	Q	H	K	R	D	E

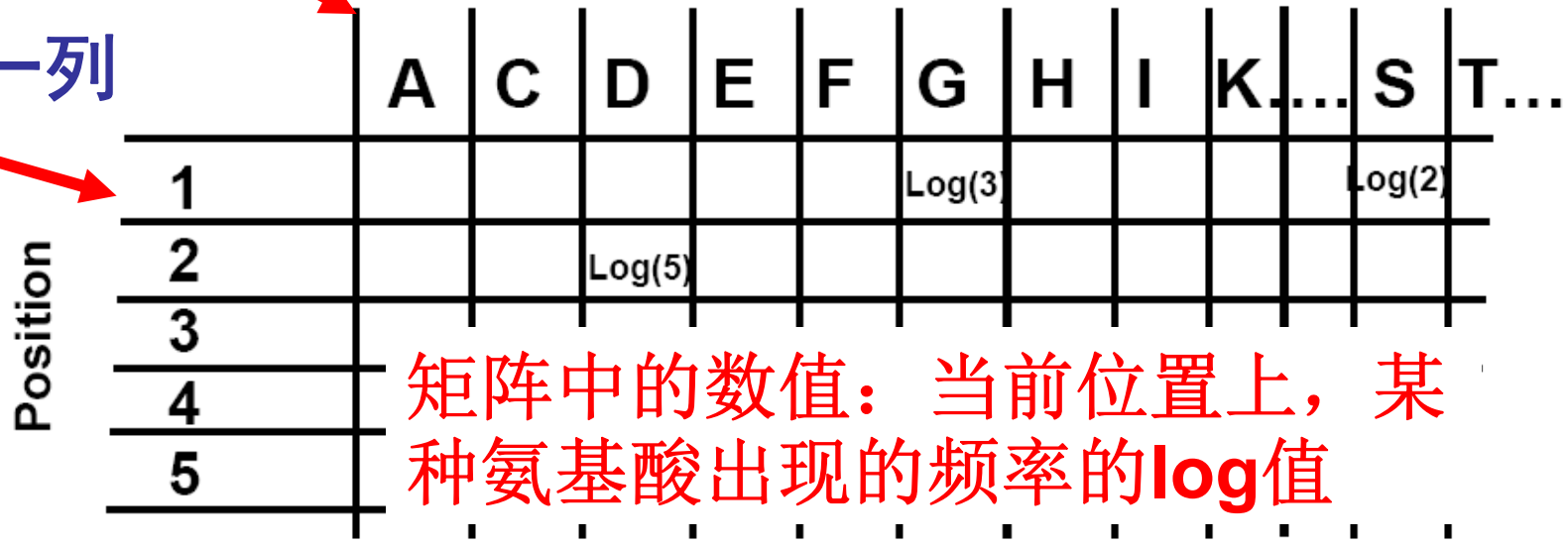
锌指功能结构域的PSSM

BLOCK -> PSSM

二十种
氨基酸

代表每一列





矩阵中的数值：当前位置上，某种氨基酸出现的频率的log值

第二种PSSM

- 每一个位置上显示每种氨基酸或者碱基出现的频率

四种碱基

碱基的位置

```
>ABF17121SCPD

A   0   0  11   2   3   7   3   5   5  14   0   1
T  14   0   2   1   6   2   7   3   0   0   0   0
G   0   0   0   0   2   3   1   2   3   0   1  13
C   0  14   1  11   3   2   3   4   6   0  13   0.
```

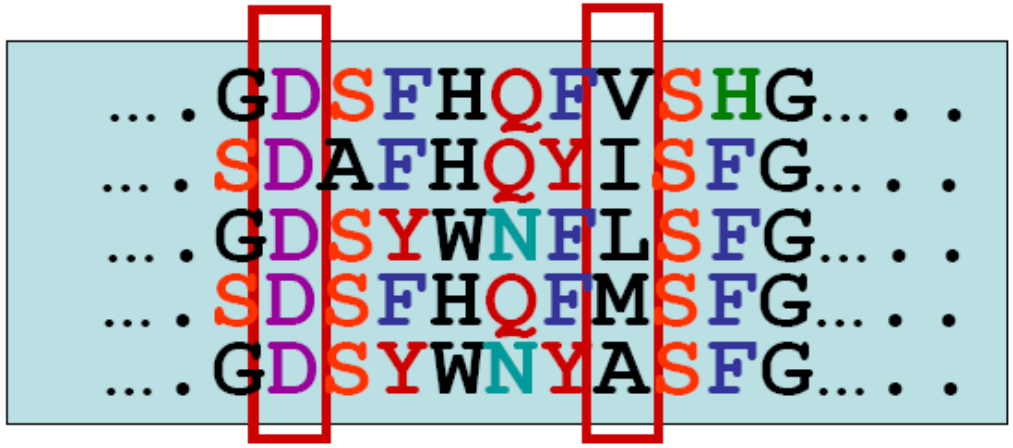
第三种PSSM

- 每一个位置显示氨基酸/碱基出现的概率

Pos	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**0.3**	**0.6**	0.1	0.0	0.0	**0.4**	**0.7**	0.1	0.1
C	**0.4**	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.2
G	0.2	0.2	**0.8**	**1.0**	0.0	**0.4**	0.1	**0.8**	0.2
T	0.1	0.1	0.1	0.0	**1.0**	0.1	0.1	0.0	**0.5**

PSSM：思考与应用

- 可以根据BLOCK推导得到的PSSM进行数据库的搜索，发现包含该模式的新的蛋白质，并预测功能
- 需要思考的问题：
  - 根据PSSM如何计算新的序列？
  - PSSM中究竟包含着何等信息？



问题一：PSSM->发现

- 计算log-odds ratio/Odds ratio
- Do not miss: 性能检验！！！
- 结果需要计算*Sn*, *Sp*, *Ac* & *Mcc*
- 需要计算Self-consistency, Leave-one-out validation & *n*-fold cross-validation

问题一：PSSM->发现

- 计算log-odds ratio/Odds ratio
- Do not miss: 性能检验！！！
- 结果需要计算*Sn*, *Sp*, *Ac* & *Mcc*
- 需要计算Self-consistency, Leave-one-out validation & *n*-fold cross-validation

计算log-odds ratio

Pos	-3	-2	-1	+1	+2	+3	+4	+5	+6
A	**0.3**	**0.6**	0.1	0.0	0.0	**0.4**	**0.7**	0.1	0.1
C	**0.4**	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.2
G	0.2	0.2	**0.8**	**1.0**	0.0	**0.4**	0.1	**0.8**	0.2
T	0.1	0.1	0.1	0.0	**1.0**	0.1	0.1	0.0	**0.5**

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

- **P(S|+)，根据阳性训练数据计算出来的概率**

Then, P(S|-)?

- 负样本/阴性数据的概率计算
- 计算方法：
  - A. DNA序列，四种碱基出现的频率
  - B. 蛋白质序列，20种氨基酸出现的频率

Odds Ratio

5′ splice signal

Con:	C	A	G	…	G	T
Pos	-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**

Background

Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

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

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

Log-odds Ratio

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

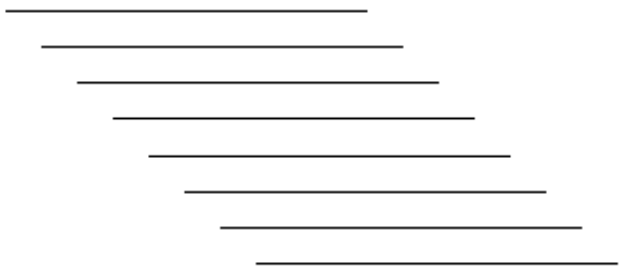
$$= \prod_{k=1}^{k=9} P_{-4+k}(S_k)/\, P_{bg}(S_k)$$

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

计算流程：滑动窗口

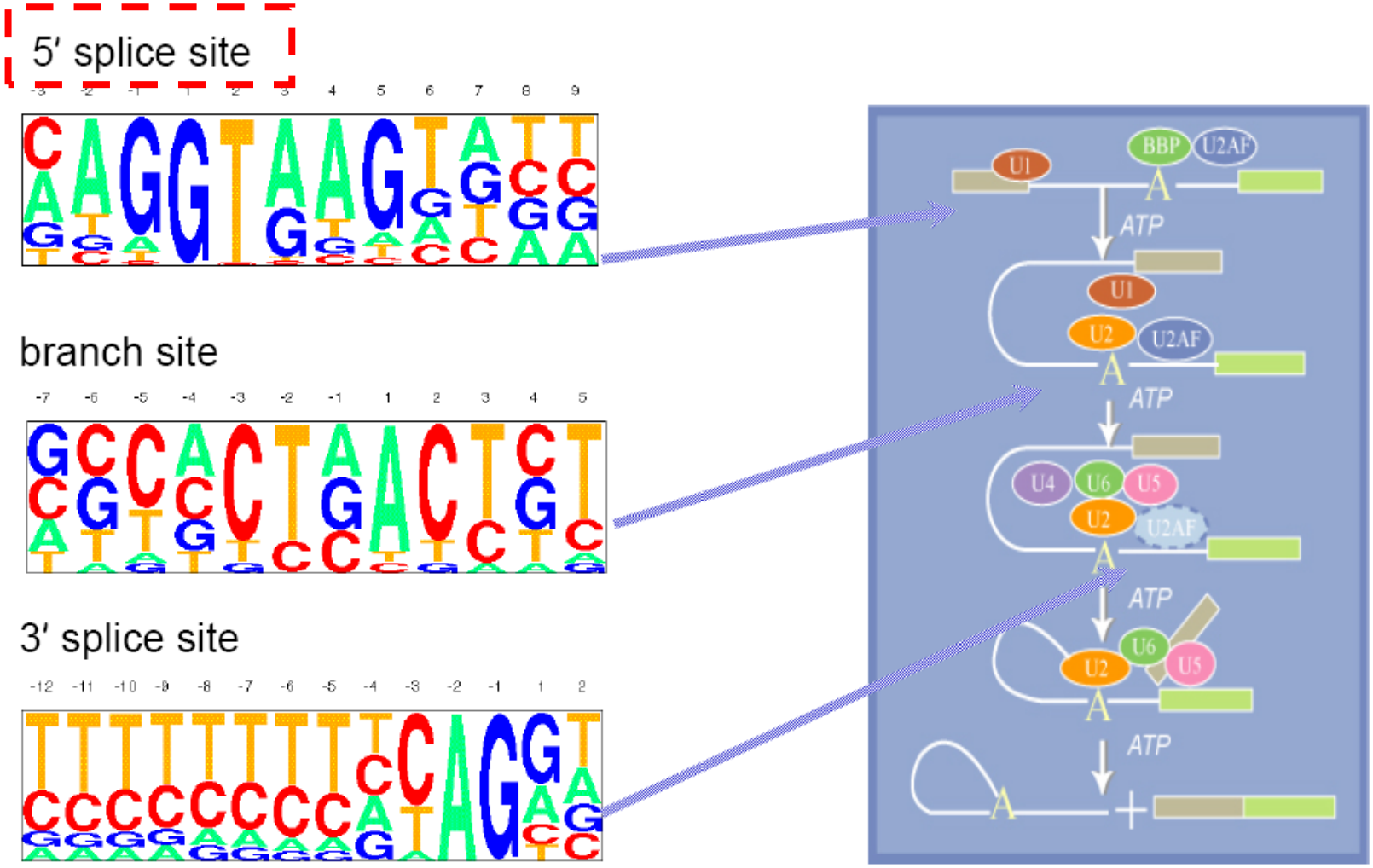
Slide WMM along sequence:

ttgacctagatgagatgtcgttcacttttactgagctacagaaaa



- 设定域值；窗口宽度12 bp；依次打分预测

例：剪接模型 (Splicing)



计算log-odds ratio

5′ splice signal

Con: C A G … G T

Pos	-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**

Background

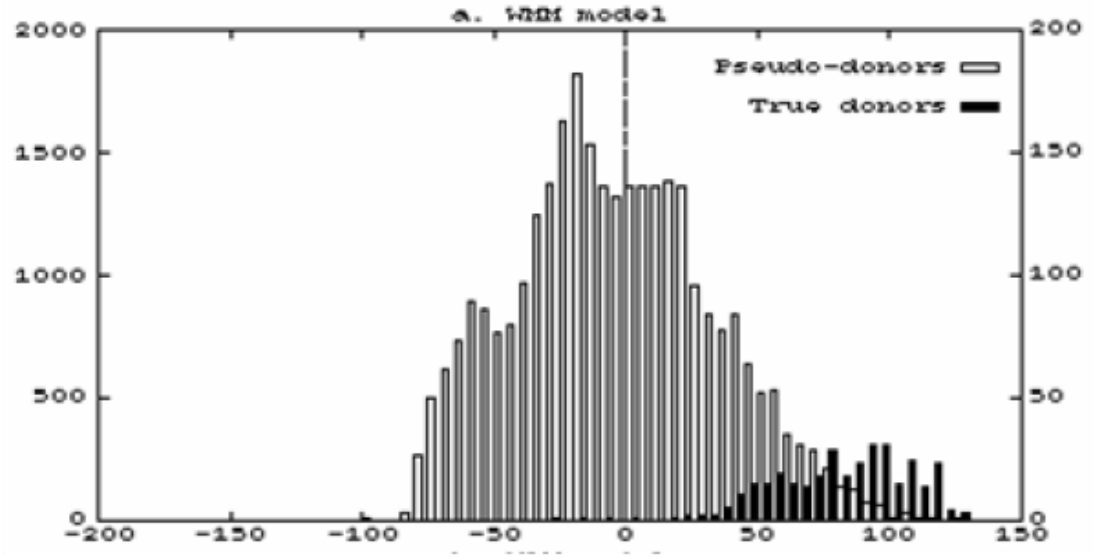
Pos	Generic
A	0.25
C	0.25
G	0.25
T	0.25

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

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

真实的打分情况：5’ SS

"Decoy" 5′ Splice Sites



True 5′ Splice Sites

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

结果解释

- 细胞中的剪接机器（Splicing machinery）可能识别其他的，不包括在训练数据中的模式
- PSSM模型不能很好的反映真实的5’ SS的识别情况
- 两种可能：either or both

Log-odds ratio *vs.* 贝叶斯

- X-box序列, 1000个4 bp的DNA序列有100个为真实的X-box。分布如下：

	第一位	第二位	第三位	第四位
A	70%	10%	1%	5%
T	10%	10%	97%	5%
C	10%	70%	1%	5%
G	10%	10%	1%	85%

P(X-box)=0.1

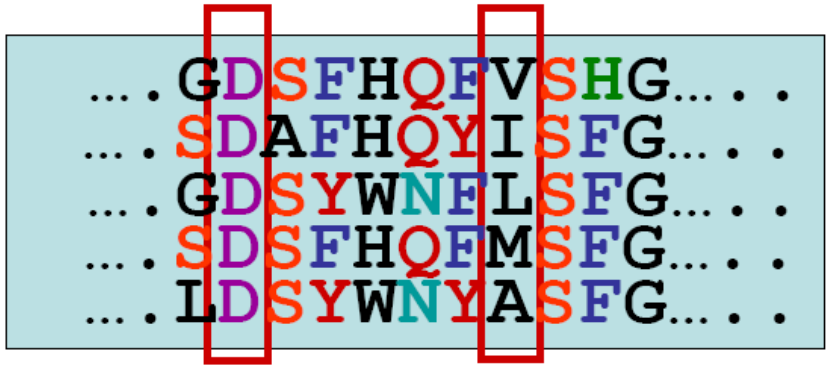
$$P(X-box \mid X_1X_2X_3X_4) = \frac{P(X-box)\prod_i q_{xi}^{x-box}}{P(X-box)\prod_i q_{xi}^{x-box} + P(nonX-box)\prod_i q_{xi}^{nonX-box}}$$

Log-odds ratio *vs*. 贝叶斯 (2)

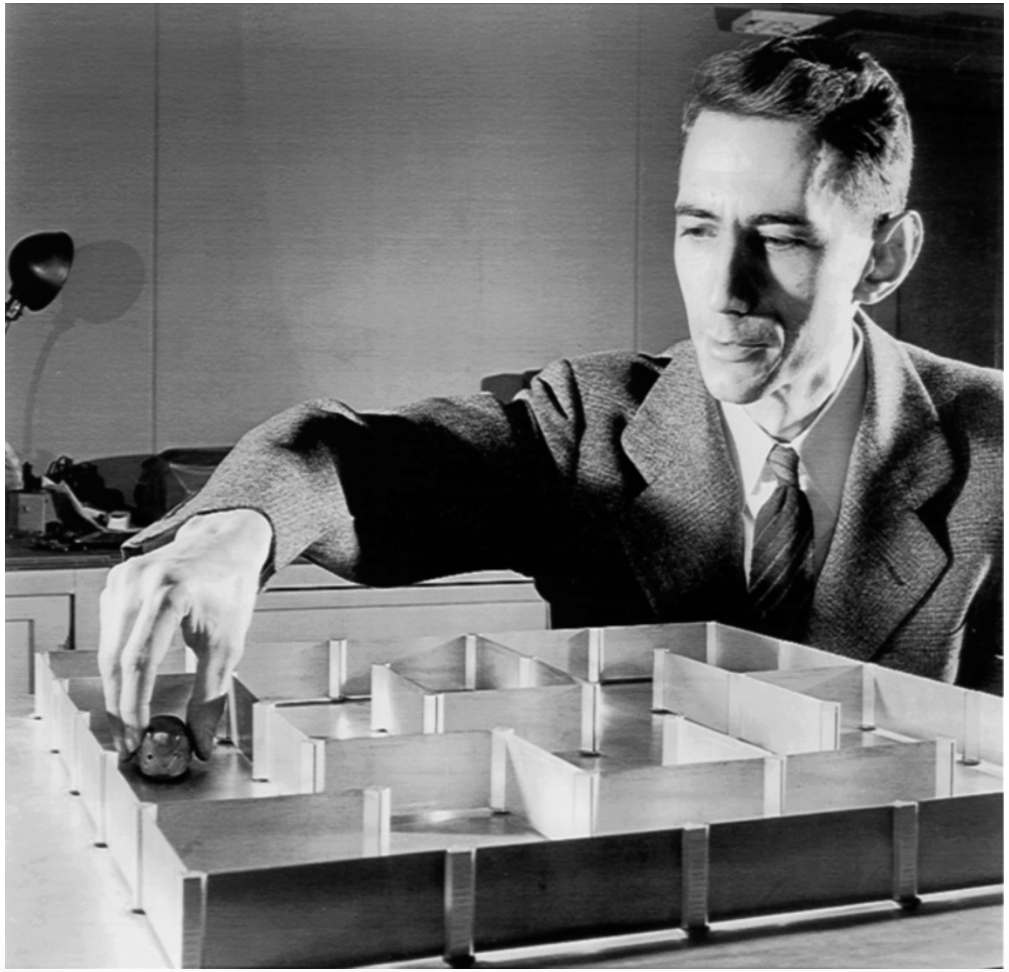
- 贝叶斯方法：必须估计真实的与错误的数量上的比例；另外，假设在未知数据中，(+)与(-)之间的比例不变
- **Log-odds ratio**: 不需要知道(+)与(-)之间的比例；观测到的数据，其为(+)与(-)的比值
- 域值的确定：都需要计算***Sn***, ***Sp***, ***Ac*** & ***MCC***

问题二：PSSM->信息？

- PSSM/motif/domain/BLOCK：每一个位置上究竟包含了什么样的信息？
- 对于同一个motif/PSSM：有些位点较其他位点提供更多的信息，why?
- 如何定量化“信息”？



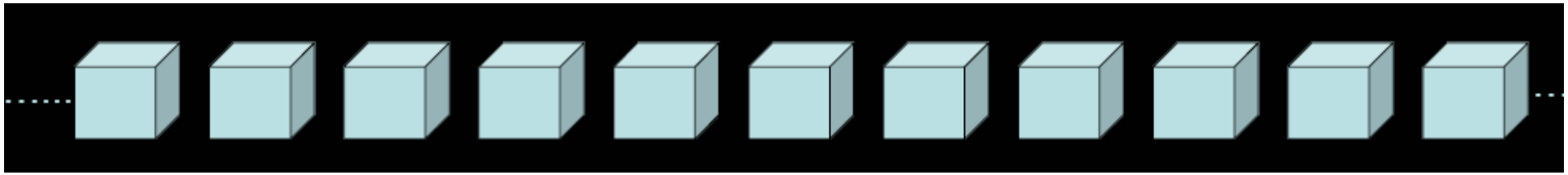
信息论：Claude Shannon



- 信息论的奠基人

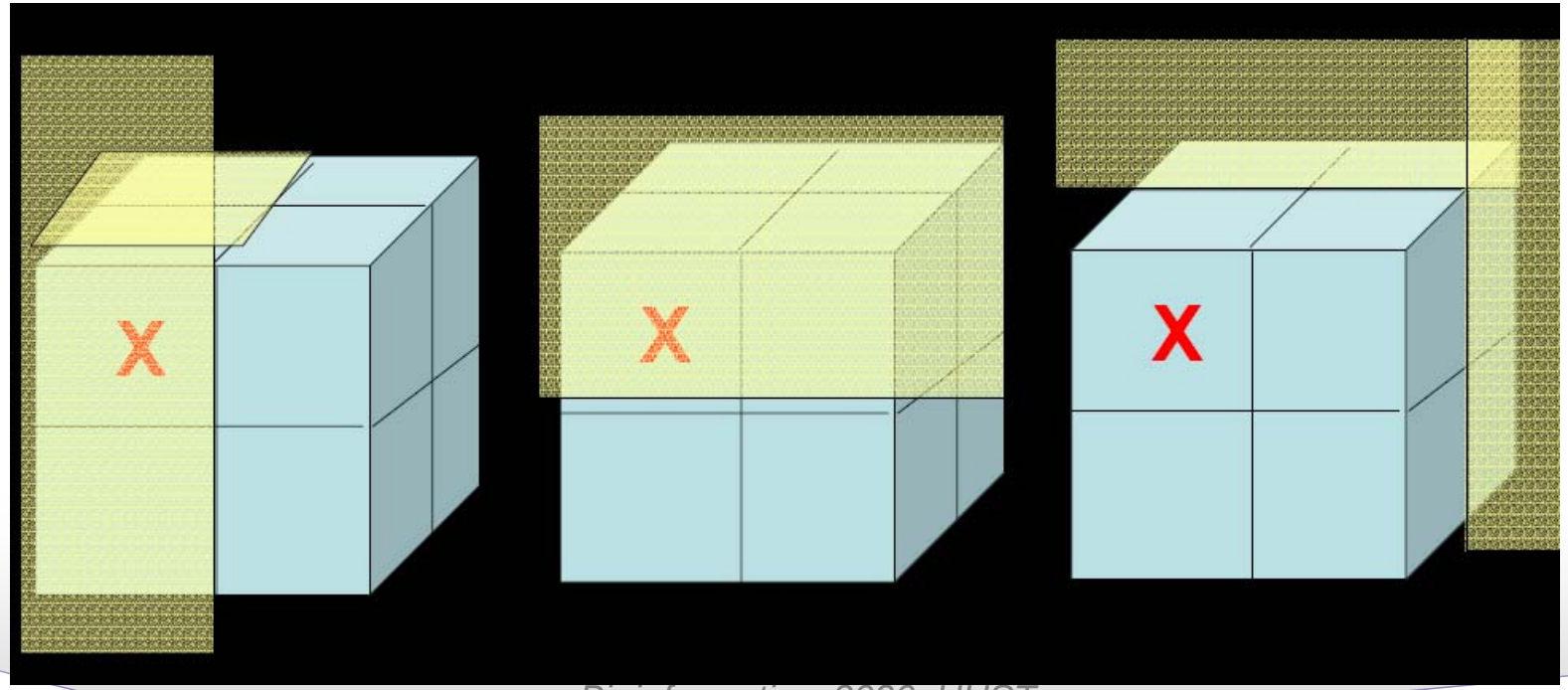
1,048,576个盒子：Yes/No?

- 随机将10000 RMB的支票放入1,048,576个盒子之一
- Play 20 questions: yes/no



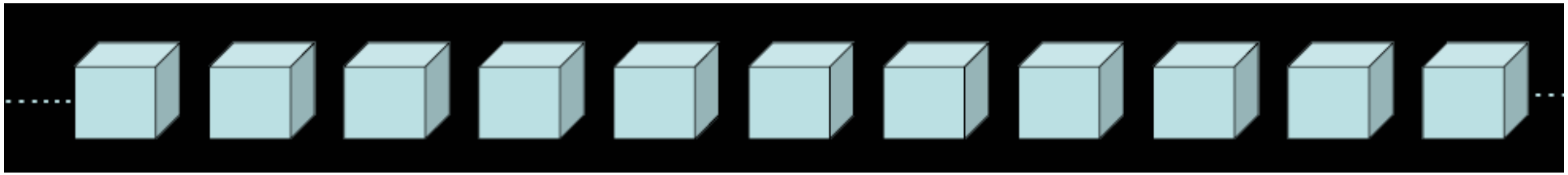
8个盒子

- 最少问多少个yes/no的问题能够定位支票？
- Answer: $\log_2 8 = 3$



1,048,576个盒子：Yes/No?

- 随机将10000 RMB的支票放入1,048,576个盒子之一
- Play 20 questions: yes/no



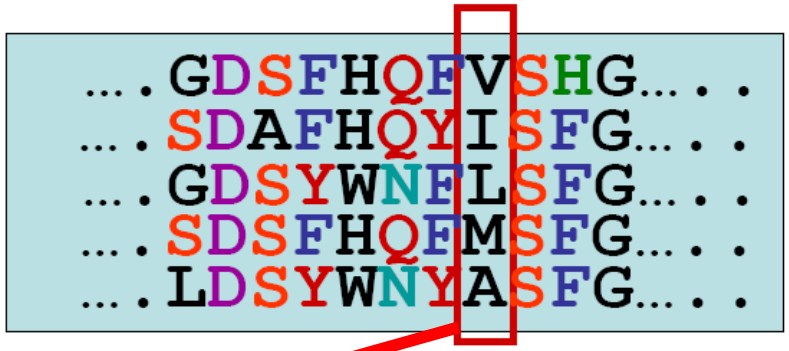
- $2^{20} = 1,048,576$

信息论

- $2^b = M$; b为bit (binary digit) 信息
- M：所有概率事件的总数；因此：
- $b = \log_2(M)$; => $b = -\log_2(1/M)$ => $b = -\log_2(P)$; 所有事件概率相同，则P=1/M
- 例：对于某一个motif的一个位置上，可能存在20种氨基酸，且概率相等，则P=1/20
- 香农熵：$b = -\log_2(1/20) = 4.32$ bits

信息论 (2)

- 若概率不等同，如何处理？
- 定义 $u_i = -\log_2(P_i)$



平均熵值 = $\dfrac{u_V+u_I+u_L+u_M+u_A}{N}$

N：全部序列的数目

香农熵的平均值 = $\dfrac{\sum_{i=1}^{20} N_i u_i}{N}$

N_i：在该位置上为氨基酸i的序列的数目

信息论 (3)

$$\frac{\sum_{i=1}^{20} N_i u_i}{N} \quad \Rightarrow \quad \sum_{i=1}^{20} \frac{N_i u_i}{N}$$

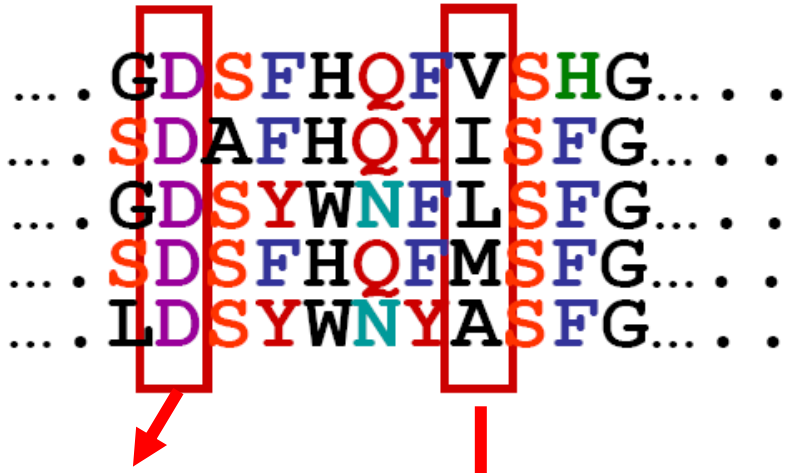
- 上式中，$N_i/N=P_i$;因此，上式可转化为：$\sum_{i=1}^{20} P_i u_i$
- 因此，香农熵公式为：

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

信息论：意义？

- 香农的信息熵公式：
$$H = -\sum_{i=1}^{20} Pi(\log_2 Pi)$$
- H为每个位置上的“香农熵”
- 香农熵：不确定性！
- 在每一个位置上，各种氨基酸出现的不确定性

信息论 (4)



- P(D)=1, 因此， H= -1*$\log_2$(1) = -1 * 0 =0

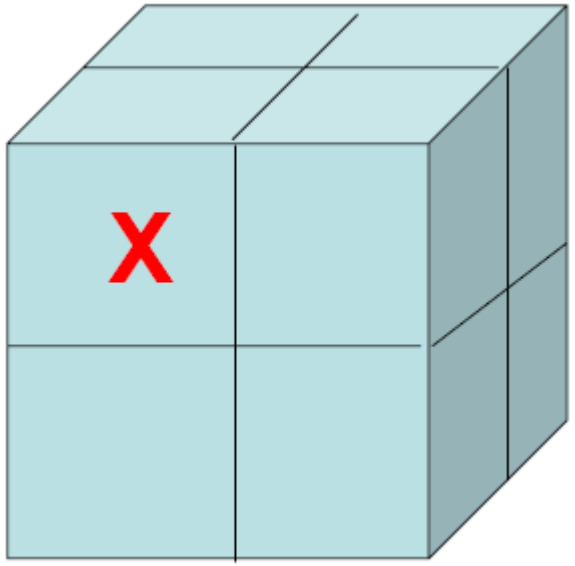
无不确定性

  - P(V) = P(I) = P(L) = P(M) = P(A) = 1/5;
  - H= -(1/5)*$\log_2$(1/5)- (1/5)*$\log_2$(1/5)-(1/5)*$\log_2$(1/5)-(1/5)*$\log_2$(1/5)-(1/5)*$\log_2$(1/5)=2.32 bit

高不确定性

不确定性 -> 信息

- 盒子模型；
- 假设只能回答两个问题；则
  - A. 回答问题之前，不确定性为3 bits
  - B. 回答问题之后，不确定性为1 bit
- 获得信息R：
- $R = H_{before} - H_{after} = 3-1 = 2$ bits



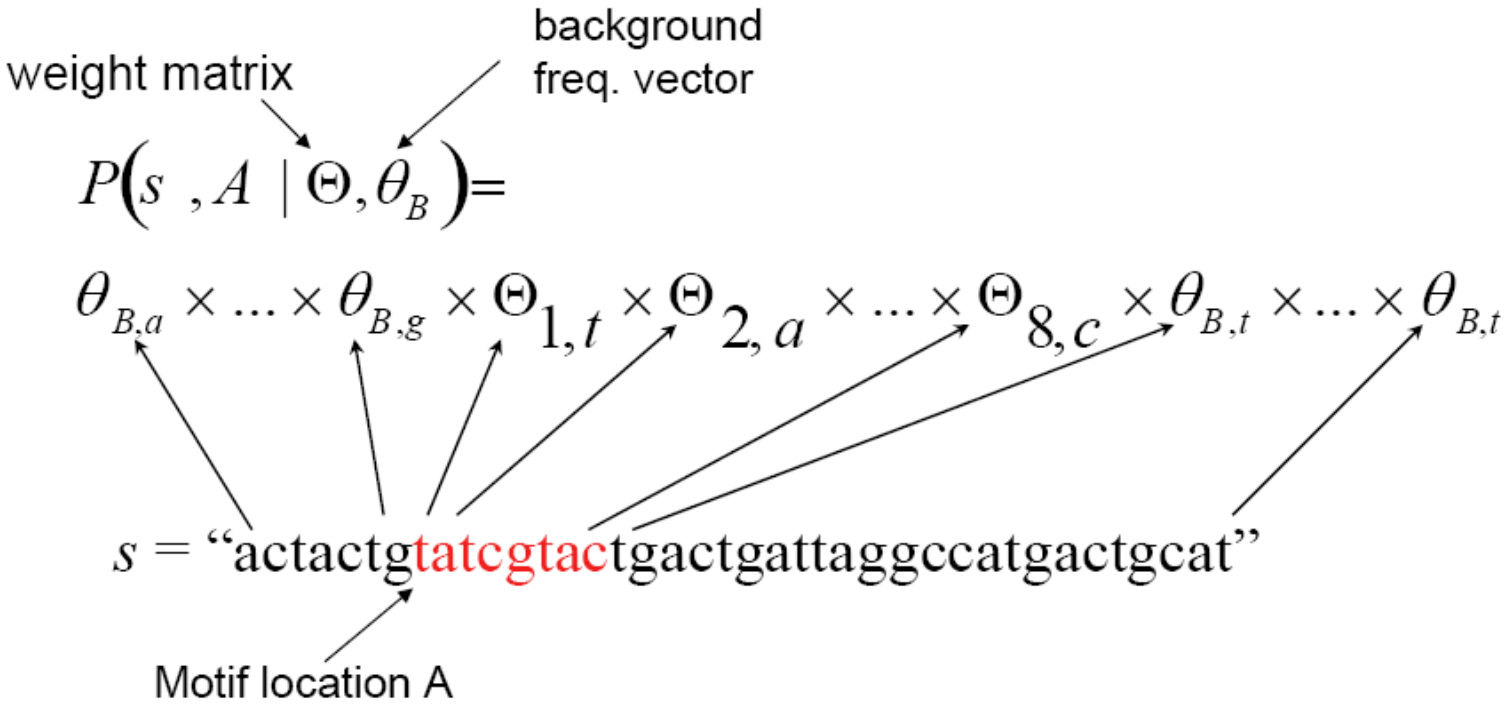
不确定性 -> 信息 (2)

- 假设，所有氨基酸出现的频率是相等的；则
- H_{before} = 4.32;
- H_{after} = 0;
- Motif在该位置的信息量为：4.32 bits

- H_{before} = 4.32;
- H_{after} = 2.32;
- Motif在该位置的信息量为：2 bits

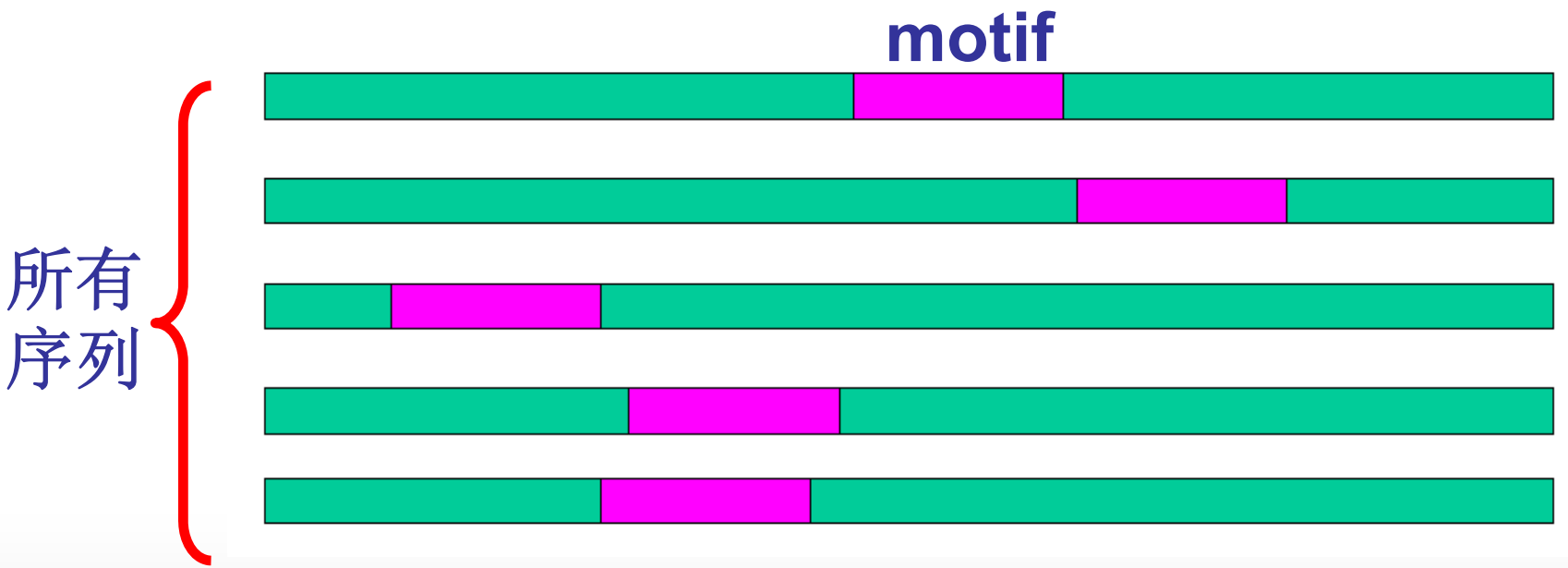
模体发现：Gibbs Sampler

- Gibbs Sampler是一种Monte-Carlo类的方法，对于输入序列，找到一个最大的似然函数
- 对于序列s，且在位置A有一个motif的似然函数，定义如下：



Gibbs Sampling 算法 (1)

- 从每条序列上随机的抽取一段序列，序列长度固定

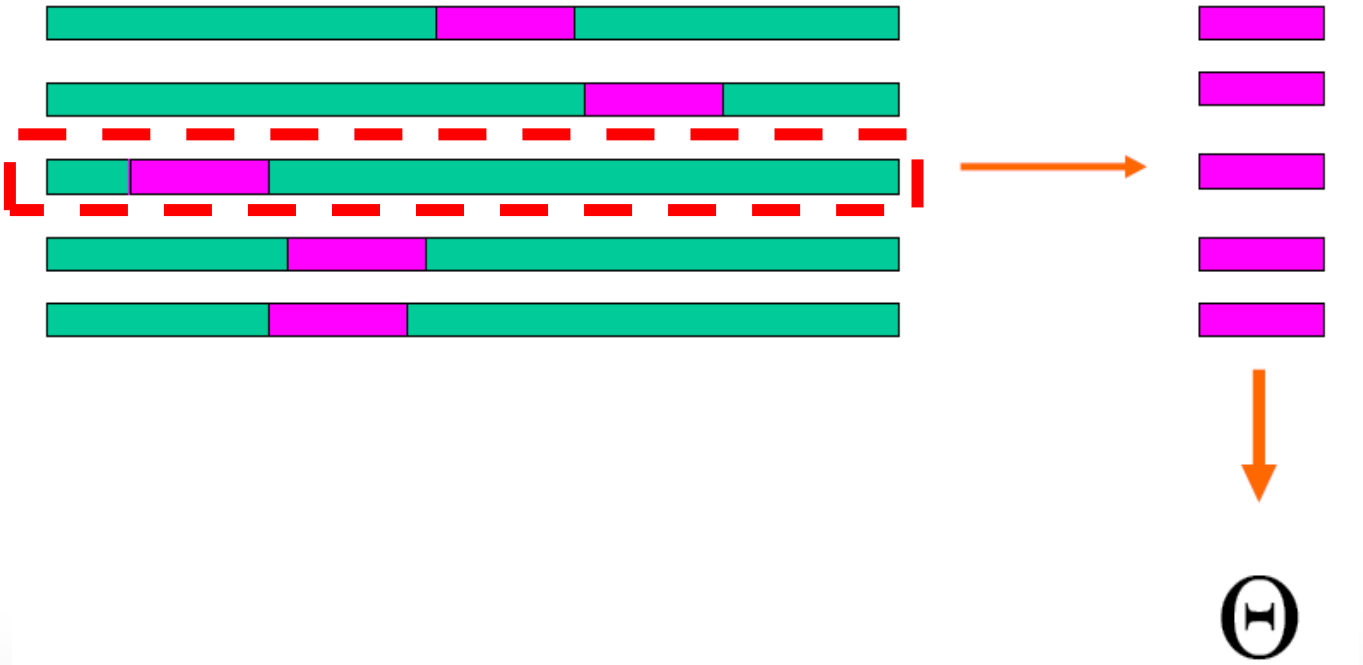


Gibbs Sampling 算法 (2)

- 构建PSSM/权重矩阵

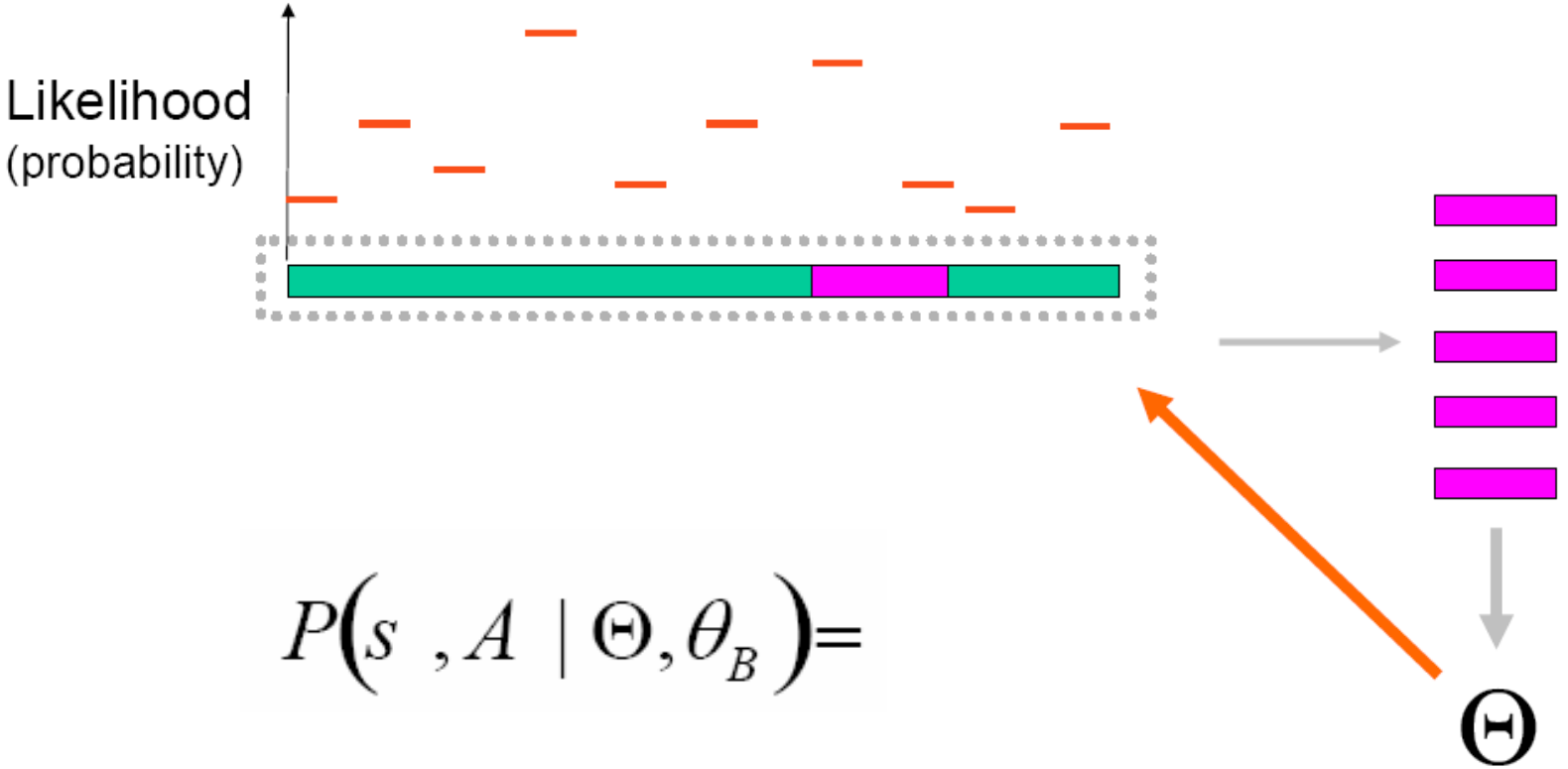
Gibbs Sampling 算法 (3)

- 随机挑选一条序列



Gibbs Sampling 算法 (4)

- 用构建好的PSSM对该序列上所有可能的motif进行打分 (窗口滑动，每次1个氨基酸或者碱基)



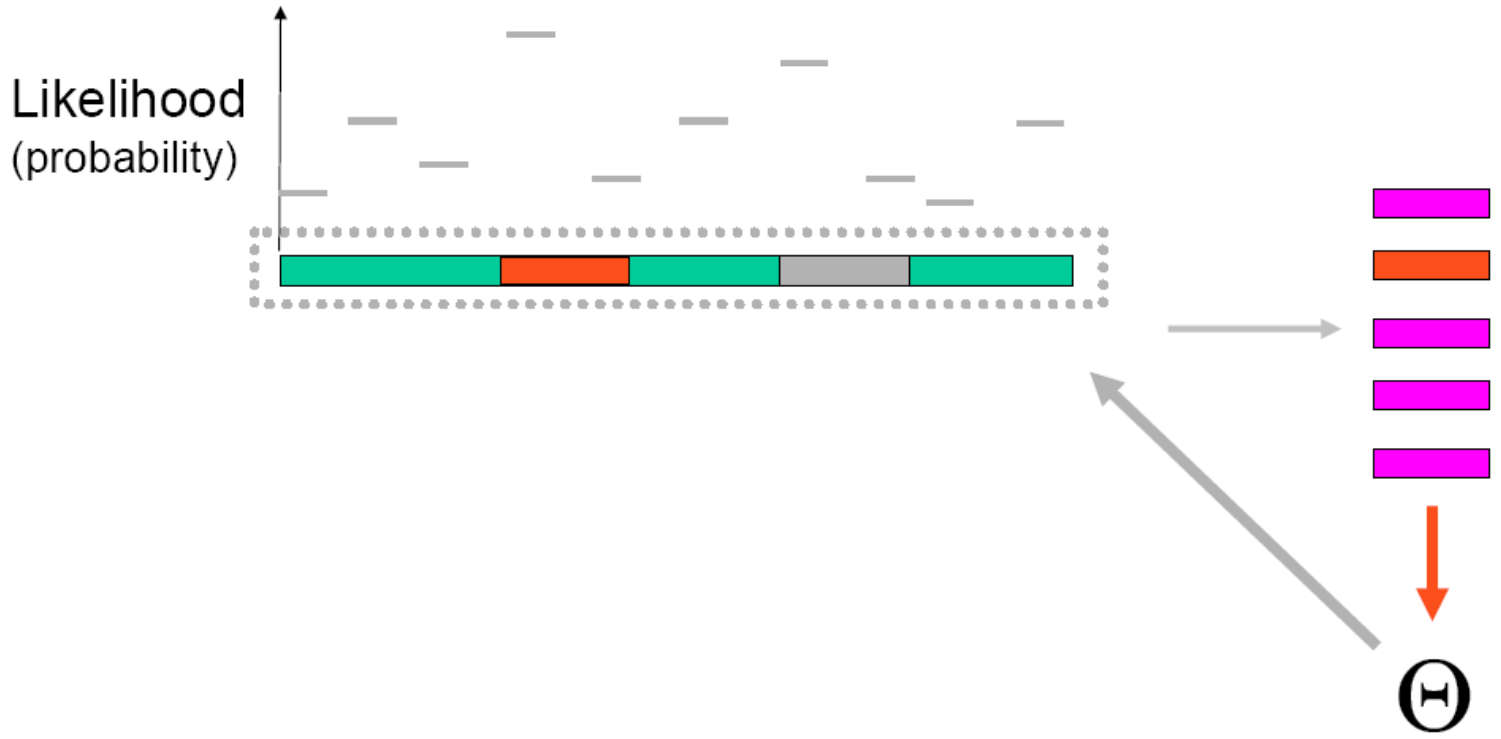
$$P(s, A \mid \Theta, \theta_B) =$$

Gibbs Sampling 算法 (5)

- 根据似然性的计算，得到似然值最大的模体，即新的motif

Gibbs Sampling 算法 (6)

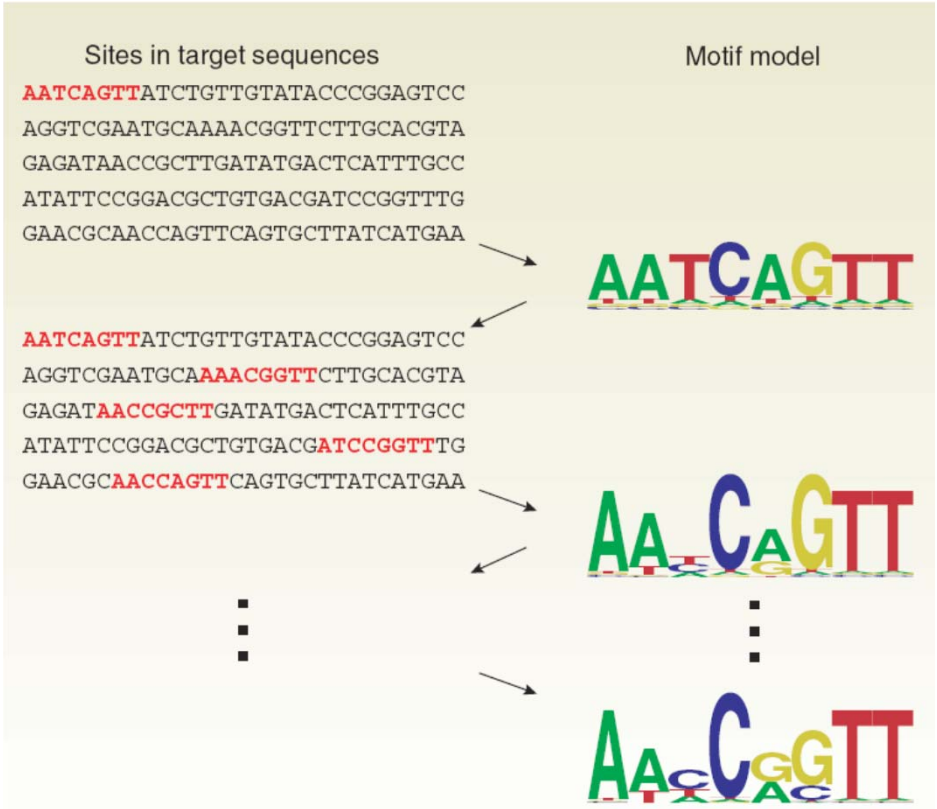
- 更新PSSM矩阵

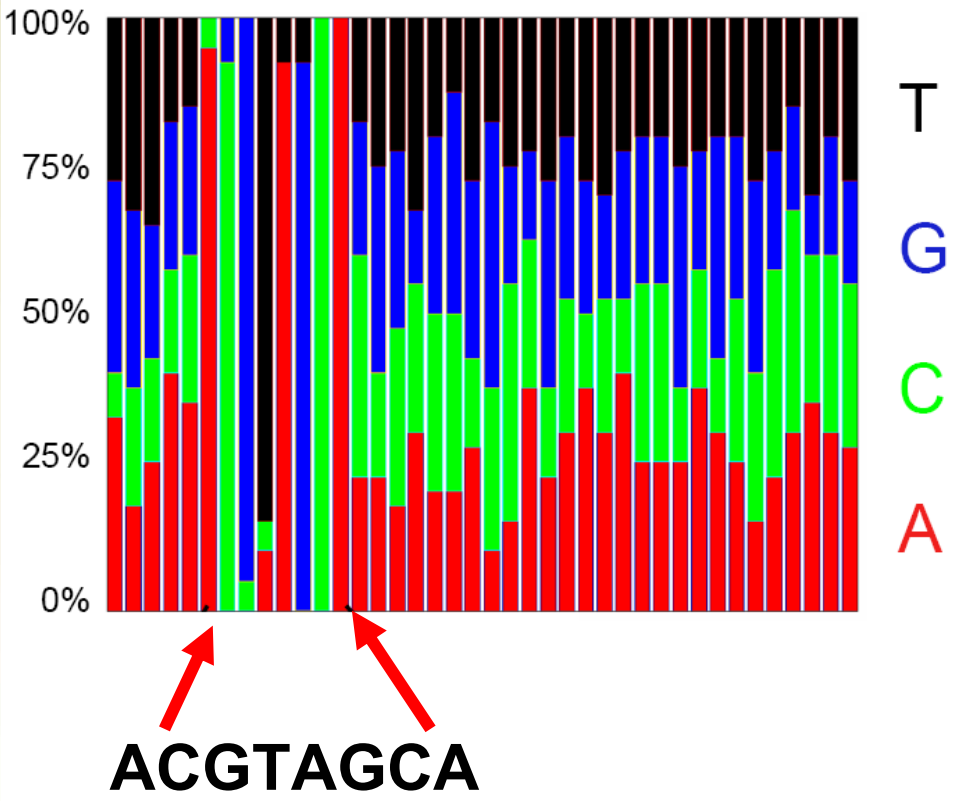


Gibbs Sampling 算法 (7)

- 反复迭代计算，直到似然性结果与PSSM不再发生变化

Strong Motif





Gibbs Sampler: 总结

- 模体发现的一种随机算法（Monte Carlo）
- 寻找次优解的算法
- 根据PSSM/WMM对随机抽取的序列进行打分来调整采样，直到结果收敛
- 不能够保证每次运算的结果一致：需要多运算几次，并进行比较
- 对蛋白质、DNA、RNA序列模体的发现有帮助