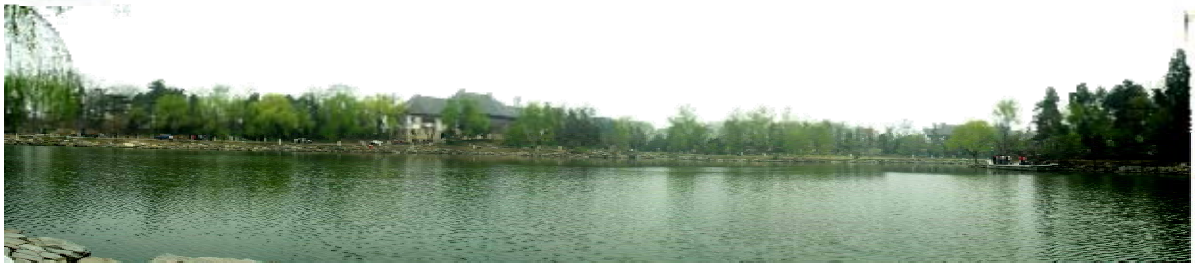



Introduction to Computational Biology

《计算生物学导论》

第七章

计算生物学的随机过程方法



Peking University

Introduction to Computational Biology

A mainstream topic in computational biology is the problem of sequence annotation: given a sequence of DNA/RNA or protein, …, *etc*., we want to identify "interesting" elements

Examples:

–DNA/RNA: genes, promoters, splicing signals, binding sites, *etc*

–Protein: coiled-coil domains, transmembrane domains, signal peptides, phosphorylation sites, *etc*

–Generally: homologs, *etc*

The sequence of many of these interesting elements can be characterized statistically, so we are interested in modeling them.

By modeling, we mean find statistical models then can:

– Accurately describe the observed elements of provided sequences;

– Accurately predict the presence of particular elements in new or un-annotated sequences;

– If possible, be readily interpretable and provide some insight into the actual biological process involved (*i.e.* not a black box).

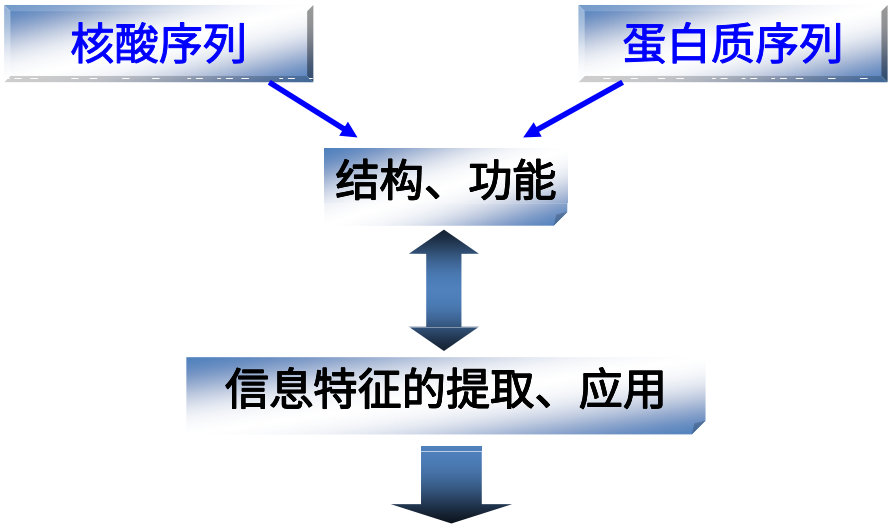
Peking University
Introduction to Computational Biology

随机过程 (Random process, or Stochastic process)

A collection of random variables, representing the evolution of some a system of random values over time, which is usually just a kind of mathematical expression.

- **A stochastic process *X* is a collection of random variables $\{X_t, t \in T\}$**
- **The variable *t* is often interpreted as time, and *X*(*t*) is the state of the process at time *t*.**
- **A realization of *X* is called a sample path.**
- **There are a number of special cases that are of high interest in computational biology, in particular Markov processes.**

§7.1 生物序列的Markov过程模型



Peking University
Introduction to Computational Biology

生物序列分析的两类问题

（1）判别问题

Does the sequence belong to a particular structure class (family)?

（2）信息结构的识别

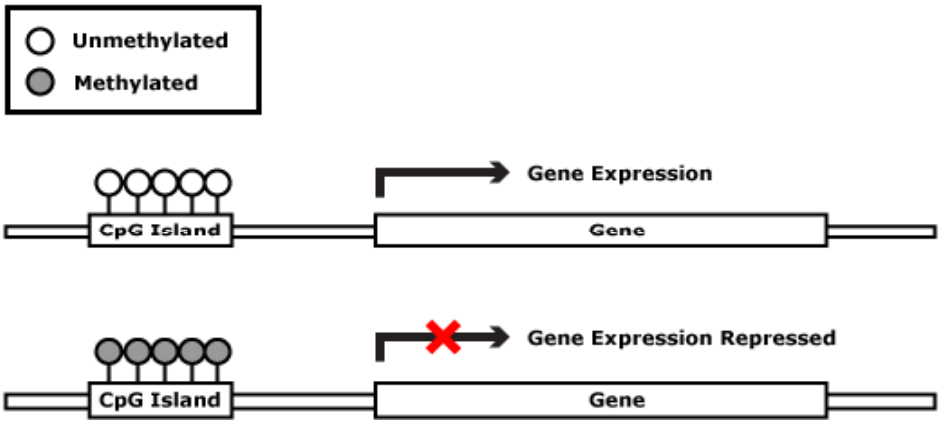
Assuming the sequence includes some structure classes (families), what can we say about its internal structure?

例子：人类基因组中与启动子相关的CpG岛信号

在人类基因组的许多基因的转录起始区域中，二核苷酸CG出现的频率往往要高于基因组其它的区域，因此，CG的高含量区（通常长达～10^3碱基）可能意味着转录启动子的存在，这种“ CG”高含量区被称为CpG岛（CpG islands）。对CpG岛的识别，有助于转录起始信号的识别。

【CpG岛（CpG islands）：指DNA上一个区域，此区域含有大量相连的胞嘧啶（C）、鸟嘌呤（G）以及使两者相连的磷酸酯键（p）。哺乳类基因中的启动子上，含有约40%的CpG岛（人类约70%）。一般CpG岛的长度约300到3000bp。常用的正式定义是指一个至少含有200bp的区域，其中GC所占比例超过50%，且CpG的观察值／预测值比例必须高于0.6。此部分的CpG岛与基因相连，可用来作为限制酶的识别位置。】

Peking University
Introduction to Computational Biology



两个问题：

（1）给定一段DNA序列片段，判别它是否CpG岛？
对应于Markov模型问题

（2）给定一段DNA序列，识别其中的CpG岛？
对应于隐Markov模型问题

Andrei A Markov was a graduate of Saint Petersburg University (1878), where he began a professor in 1886. His early work was mainly in number theory and analysis, continued fractions, limits of integrals, approximation theory and the convergence of series.

After 1900 *Markov* applied the method of continued fractions, pioneered by his teacher *Pafnuty Chebyshev*, to probability theory. He proved the central limit theorem under fairly general assumptions.

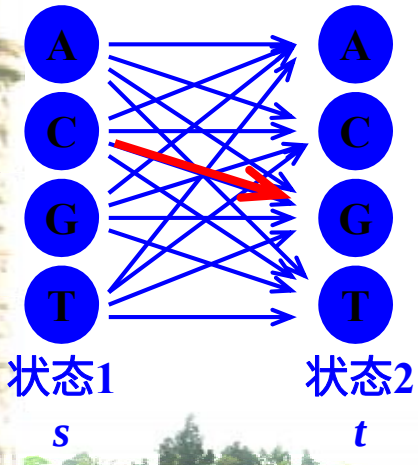
Markov is particularly remembered for his study of Markov chains, sequences of random variables in which the future variable is determined by the present variable but is independent of the way in which the present state arose from its predecessors. This work launched the theory of stochastic processes.

Andrei Andreyevich Markov (1856~1922)

1）符号序列的Markov模型

…ACGCCGATCCTCGCGACGACGGTGCGACGTCGTCCG…

例：CpG岛区域的双核苷酸



序列模型的假设：

一种（符号）状态转移到另一种（符号）状态时，后一状态仅取决于前一状态。

1阶Markov过程（Markov chain）

这种假设丢掉了许多信息，但是得到了复杂序列的简化的概率模型。

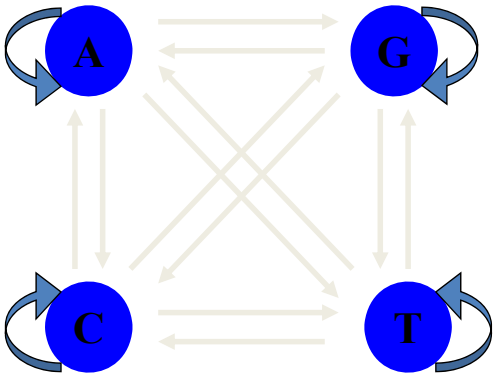
定　义

Markov过程：从一种状态转移到另一种状态时，过程仅取决于前面*n*种状态，是一种有序*n*模型。 *n*是影响下一个状态选择的状态数，由此定义的称之为*n*阶Markov过程。

最简单的Markov过程是一阶过程，状态的选择完全取决于前一状态，这种选择是依照概率来选择的。

注意：状态的选择是概率的，而非确定的。故Markov过程本质上是一种随机过程。

Peking University
Introduction to Computational Biology



例：某段DNA序列的1阶Markov过程模型 (s → t)

	t A	t C	t G	t T
s A	0.18	0.27	0.42	0.12
s C	0.17	0.37	0.27	0.19
s G	0.16	0.34	0.38	0.13
s T	0.08	0.36	0.38	0.18

定义：Markov过程
由下列参数构成：
（1）状态（state）：A、C、G、T
（2）状态转移概率 (state transition probability)矩阵：AA, AC，…的概率大小，记为a_{st}

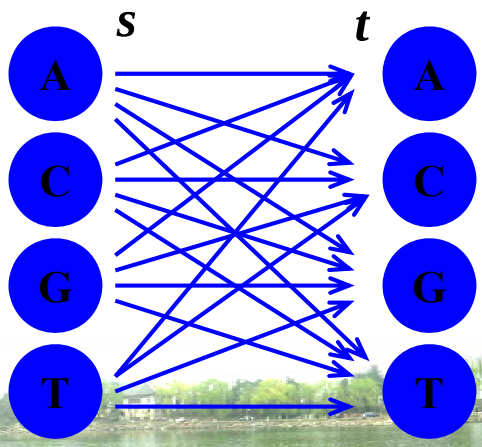
2）DNA序列的1阶Markov模型

假定DNA序列长度为L，序列x写成$x_1x_2\ldots x_L$。

状态转移概率为 a_{st}

$$a_{st} = P(x_i = t \mid x_{i-1} = s) \qquad i = 2,3,\ldots L$$

a_{st}表示碱基s后面紧邻碱基t的概率，或者状态s后紧邻状态t的概率。
（a_{st}标志状态s转移到状态t的箭头联系的强度。）



Peking University
Introduction to Computational Biology

运用乘法公式，可以写出序列$x_1x_2\ldots x_L$的一般意义下的概率模型表达式：

$$P(x) = P(x_L, x_{L-1},\ldots, x_1)$$
$$= P(x_L \mid x_{L-1},\ldots, x_1)P(x_{L-1} \mid x_{L-2},\ldots, x_1)\ldots P(x_1)$$

1阶Markov过程的特点是碱基x_i只取决于它前面x_{i-1}，而非整个序列x。

亦即：

$$P(x_i \mid x_{i-1},\ldots, x_1) = P(x_i \mid x_{i-1}) = a_{x_{i-1}x_i}$$

因此，

$$P(x) = P(x_L, x_{L-1}, \ldots, x_1)$$
$$= P(x_L \mid x_{L-1}, \ldots, x_1) P(x_{L-1} \mid x_{L-2}, \ldots, x_1) \ldots P(x_1)$$

可写成：

$$P(x) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) \ldots P(x_2 \mid x_1) P(x_1)$$
$$= P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

$$P(x) = P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$

Peking University
Introduction to Computational Biology

练习

证明：所有长度为L的序列的概率总和为1，即

$$\sum_{\{x\}} P(x) = \sum_{x_1} \sum_{x_2} \ldots \sum_{x_L} P(x_1) \prod_{i=2}^{L} a_{x_{i-1} x_i}$$
$$= 1$$

3）DNA序列的起始状态和结束状态

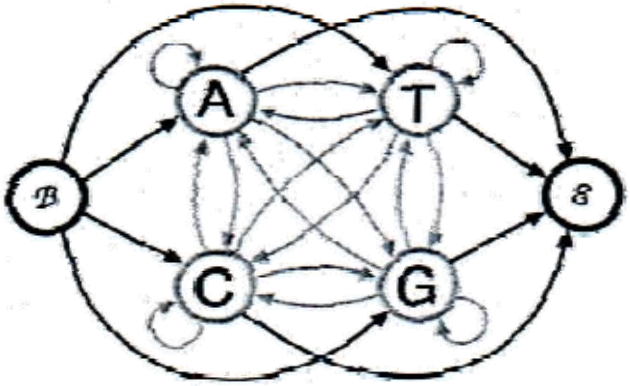
为描述序列起始和结束的状态转移，在A、C、G、T 4个状态之外，人为添加定义2个状态符号*B*（起始）、*E*（结束）。

对于序列的起始，定义$x_0=B$，对应x_0 → x_1的转移概率为：

$$a_{Bs} = P(x_1 = s)$$

对于序列的结束，对应x_L → *E*的转移概率为：

$$a_{tE} = P(E \mid x_L = t)$$



4）CpG islands判别的应用

$$P(x) = P(x_L \mid x_{L-1})P(x_{L-1} \mid x_{L-2})\ldots P(x_2 \mid x_1)P(x_1)$$
$$= P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i}$$

序列*x*属于某一Markov过程的似然率

衡量序列*x*是否满足某一Markov模型的度量 → 判别分析问题

已知48条人类DNA序列（全长60kb），每条都含有确认的CpG islands位置、长度注释信息。据此构造训练集：

正方训练集（Positive） ← CpG islands区的DNA序列

负方训练集（Negative） ← CpG islands区之外的DNA序列

分别从positive训练集和negative训练集中统计出双核苷酸*st*的频率，再计算概率转移矩阵：

$$a^{+}_{st} = \frac{c^{+}_{st}}{\sum_{t'} c^{+}_{st'}} \qquad a^{-}_{st} = \frac{c^{-}_{st}}{\sum_{t'} c^{-}_{st'}}$$

Peking University
Introduction to Computational Biology

a^{+}_{st}

+	t A	t C	t G	t T
s A	0.18	0.27	0.43	0.12
s C	0.17	0.37	0.27	0.19
s G	0.16	0.34	0.38	0.13
s T	0.08	0.36	0.38	0.18

a^{-}_{st}

−	t A	t C	t G	t T
s A	0.30	0.21	0.29	0.21
s C	0.32	0.30	0.08	0.30
s G	0.25	0.25	0.30	0.21
s T	0.18	0.24	0.29	0.29

对于任一段DNA序列片段，计算下列分值：

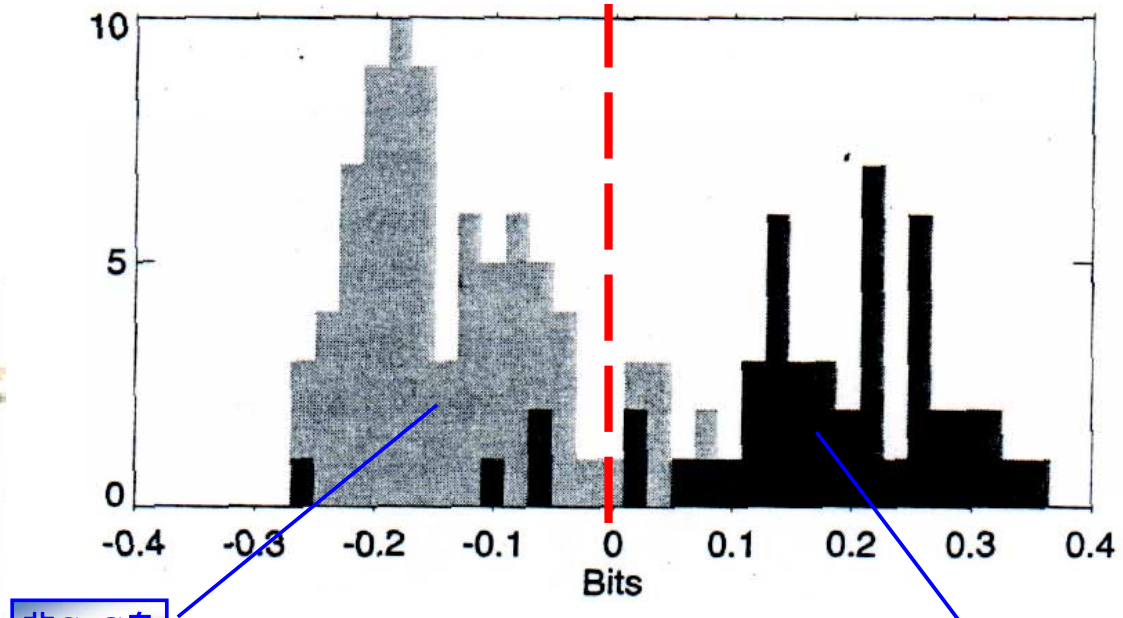
$$S(x) = \log \frac{P(x \mid \text{model}\ +)}{P(x \mid \text{model}\ -)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

β	t A	t C	t G	t T
s A	-0.74	0.42	0.58	-0.80
s C	-0.91	0.30	1.81	-0.69
s G	-0.62	0.46	0.33	-0.73
s T	-1.17	0.57	0.39	-0.68

作长度L的归一处理：

$$S(x) = \frac{1}{L} \sum_{i=1}^{L} \beta_{x_{i-1}x_i}$$

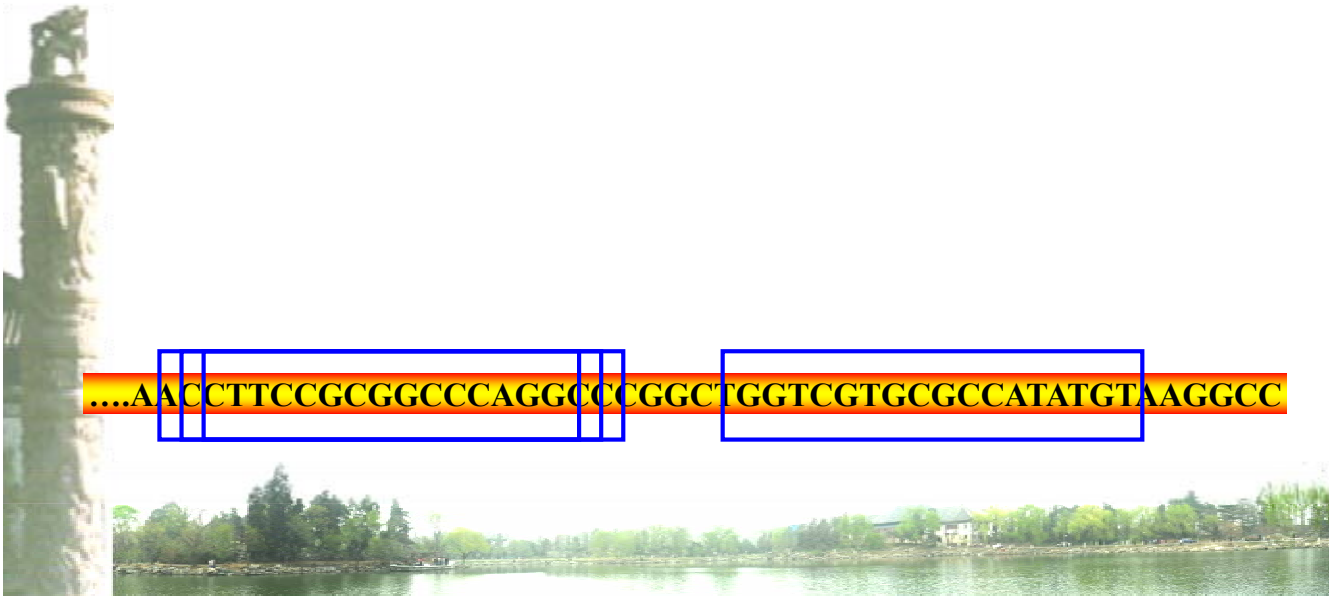
Peking University
Introduction to Computational Biology



据此，可以对某一DNA片段构造判别规则，判定是否CpG岛。

思考：

根据前面的训练集及训练出的参数，如何设计在某一长的DNA序列上识别（注释） CpG岛的大致位置的方法？



Peking University
Introduction to Computational Biology

例子：人类基因组中与启动子相关的CpG islands信号

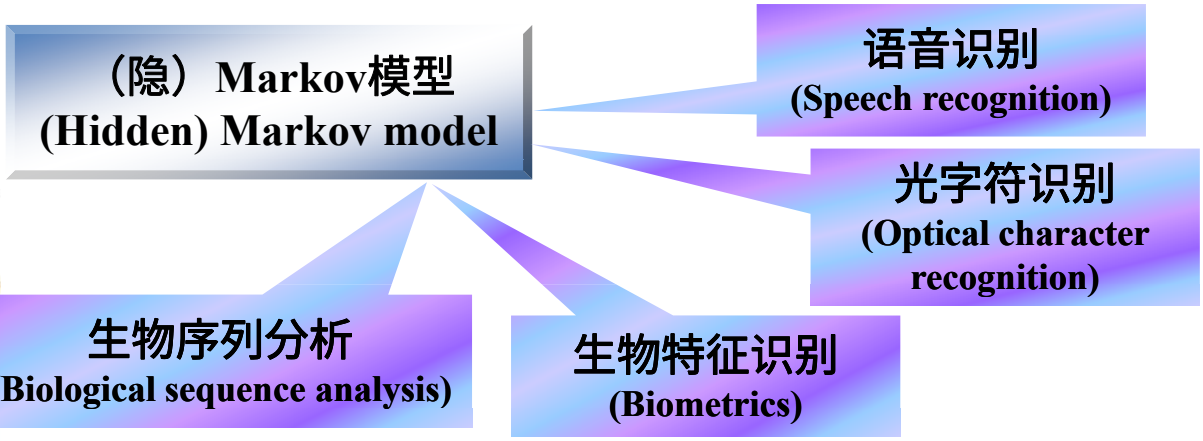
在人类基因组的许多基因的转录起始区域，二核苷酸“ CG”出现的频率往往要高于基因组其它的区域，因此，“ CG”的高含量区（通常长达数百～数千碱基）可能意味着转录启动子的存在，这种“ CG”高含量区被称为CpG岛（CpG islands）。对CpG岛的识别，有助于转录起始信号的识别。

两个问题：

（1）给定一段DNA序列片段，判别它是否CpG islands？
对应于Markov模型问题

（2）给定一段DNA序列，识别其中的CpG islands？
对应于隐Markov模型问题

§7.2 生物序列的隐Markov模型



（1）序列比较与搜寻（尤其是多序列比对）

（2）基因识别、预测（包括DNA编码与非编码区的识别、真核基因剪接位点信号识别、非编码区的转录调控信号识别、跨膜蛋白的识别……）

（3）蛋白质二级结构、家族、超家族预测、分类

……

Peking University
Introduction to Computational Biology

语音识别的HMM方法

根据记录的语音波形，分解成若干段10～20毫秒长的声学片段(帧)，每一帧对应于一个符号（通过HMM训练算法获得的声学特征片段）继而转换成符号序列。再根据已经建立的声学模型对该符号序列进行识别，判断对应的音素、音节、单词。

语音识别的基本问题

（1）语音特征提取：从语音波形中提取出随时间变化的语音特征序列。（模式的构造）

（2）声学模型与模式匹配（识别算法）：对获取的语音特征通过学习算法产生声学模型。在识别时将输入的语音特征同声学模型（模式）进行匹配与比较，得到最佳的识别结果。

（3）语言模型与语言处理：语言模型包括由识别语音命令构成的语法网络或由统计方法构成的语言模型，语言处理可以进行语法、语义分析。对小词表语音识别系统，往往不需要语言处理部分。

Peking University
Introduction to Computational Biology

语音识别的困难

（1）真实语音发声的多样性 → 庞大的语音数据库和语料数据库

（2）真实语音状态驻留时间的多样性 → 复杂的随机过程

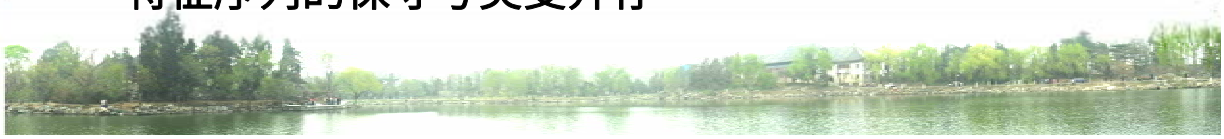
生物序列分析的挑战

（1）分析、识别分子序列（4种核苷酸符号或20种氨基酸符号）的生物学意义；共同之处：“ Based on a sequence of symbols from some alphabet, find out what the sequence represents.”

（2）分子序列的复杂性：

结构的多样性

特征序列的保守与突变并存



例子：人类基因组中与启动子相关的CpG岛信号

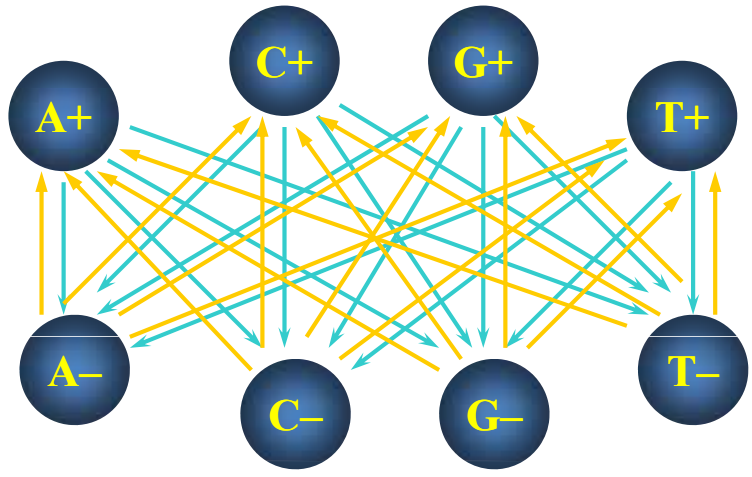
（1）给定一段DNA序列片段，判别它是否CpG岛？
对应于Markov过程问题

（2）给定一段长DNA序列，识别其中的CpG岛？
对应于隐Markov模型问题

构造模型，用于刻画较长DNA序列中的非均匀结构
（CpG岛+非CpG岛区）

- 构造描述CpG islands的“ Model +”和描述非CpG islands的“ Model － ”混合在一条DNA序列中的Markov过程
- 增加CpG islands与非CpG islands交界的状态转移

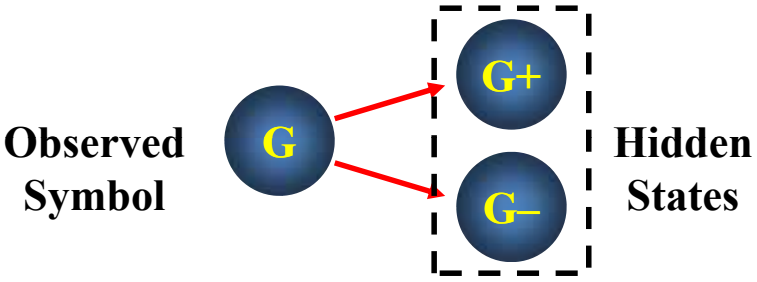
Peking University
Introduction to Computational Biology



Model ＋
(CpG岛)

Model －
(非CpG岛)

刻画含CpG岛结构的DNA序列的HMM模型
（注：除图中所示的状态转移外，“ Model +”和“ Model － ”内的状态转移的定义仍然保留，但图中未标示）



观察的符号与该符号状态的关系

Markov过程： 一对一的关系
HMM: 一对多的关系

Peking University
Introduction to Computational Biology

1. HMM的定义

观察（符号）序列x $\cdots$ C G G C G A T A $\cdots$ Observed

x_{i-1} x_i

状态序列π $\cdots$ C– G+ G+ C+ G+ A+ T– A– $\cdots$ Hidden

π_{i-1} π_i

Notes：观察（符号）序列x和状态序列π，都用Markov过程来描述。状态序列π也称为路径（path）

状态序列π也是Markov过程，故定义状态转移概率 a_{kl}来刻画状态序列π：

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$$

同样，人为添加状态序列的起始和结束状态：

$$a_{0l} = P(\pi_1 = l \mid \pi_0 = START) = P(\pi_1 = l)$$

$$a_{k0} = P(\pi_0 = END \mid \pi_L = k) = P(\pi_L = k)$$

建立了两个Markov链：符号序列$\{x_i\}$、状态序列$\{\pi_i\}$

二者独立

Peking University
Introduction to Computational Biology

为描述符号序列$\{x_i\}$和状态序列$\{\pi_i\}$的关系，定义参数输出概率$e_k(b)$ (emission probability)为：

$$e_k(b) = P(x_i = b \mid \pi_i = k)$$

表示在状态π_i为k时，该状态导致产生符号b（可以是单个符号，也可以是字符串）的概率。

Notes：状态转移概率a_{kl}和输出概率$e_k(b)$的定义都是一种条件概率。

例1：GpC island的HMM模型参数

输出（观察）符号：{A, C, G, T}

状态：{+, –} or {A+, A–, C+, C–, G+, G–, T+, T–}

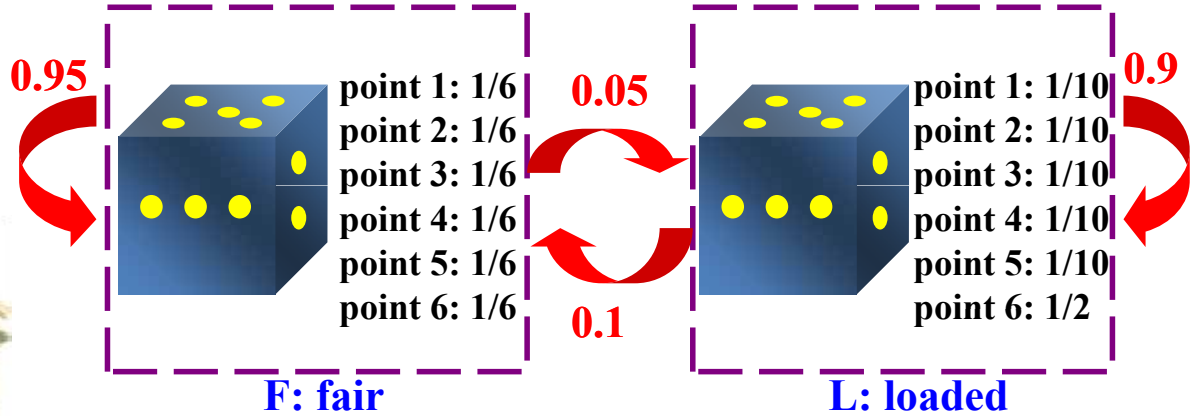
状态转移概率（p为GpC岛占全部DNA序列的比例，且$p+q=1$）：

$\pi_i \backslash \pi_{i+1}$	A^+	C^+	G^+	T^+	A^-	C^-	G^-	T^-
A^+	$0.180p$	$0.274p$	$0.426p$	$0.120p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
C^+	$0.171p$	$0.368p$	$0.274p$	$0.188p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
G^+	$0.161p$	$0.339p$	$0.375p$	$0.125p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
T^+	$0.079p$	$0.355p$	$0.384p$	$0.182p$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$	$\frac{1-p}{4}$
A^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.300q$	$0.205q$	$0.285q$	$0.210q$
C^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.322q$	$0.298q$	$0.078q$	$0.302q$
G^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.248q$	$0.246q$	$0.298q$	$0.208q$
T^-	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$\frac{1-q}{4}$	$0.177q$	$0.239q$	$0.292q$	$0.292q$

输出概率：$e_k(b) = P(x_i = b \mid \pi_i = k)$ 具体值由上述矩阵给出

Peking University
Introduction to Computational Biology

例2：某赌场骰子的HMM模型参数



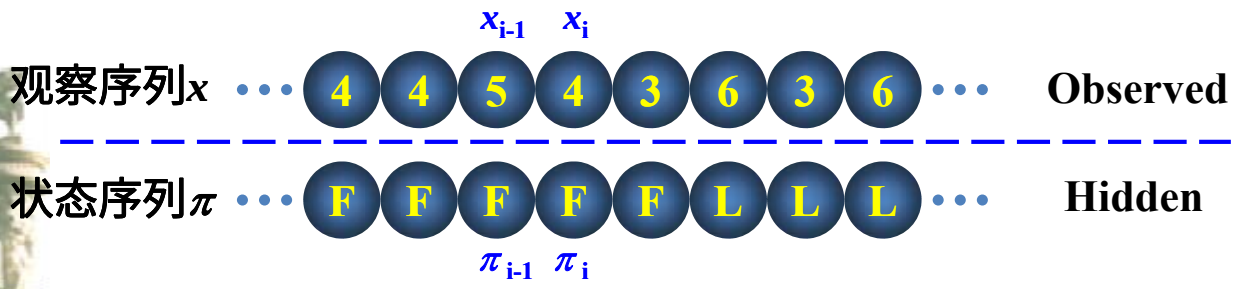
输出（观察）符号：{1, 2, 3, 4, 5, 6}　　状态：{F, L}

$a_{FF} = 0.95$　$a_{FL} = 0.05$　$a_{LL} = 0.90$　$a_{LF} = 0.10$

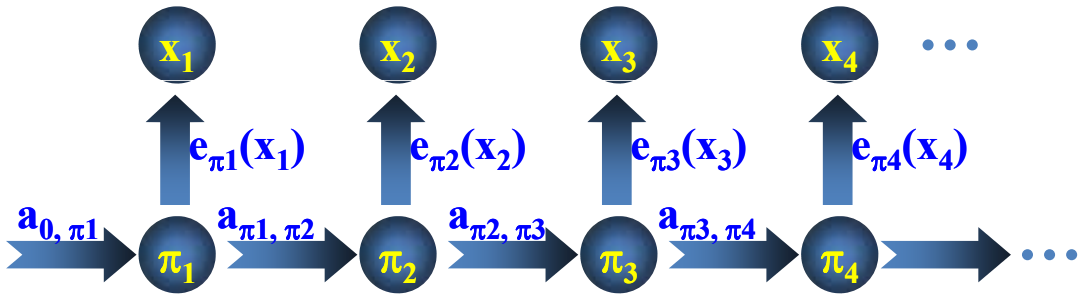
$e_F(1) = e_F(2) = e_F(3) = e_F(4) = e_F(5) = e_F(6) = 1/6$

$e_L(1) = e_L(2) = e_L(3) = e_L(4) = e_L(5) = 0.1$

$e_L(6) = 0.5$



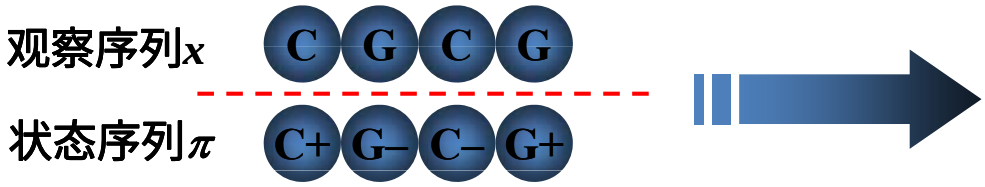
HMM产生观察符号序列的机制：



对于观察序列x和状态路径π，可计算其联合概率：

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

例：



由该状态序列产生该观察序列的概率为：

$$a_{0,C_+}\times(1\times a_{C_+G_-})\times(1\times a_{G_-C_-})\times(1\times a_{C_-G_+})\times(1\times a_{G_+,0})$$

Peking University
Introduction to Computational Biology

One might hope that knowledge of a good statistical model could suggest a physical model, *i.e.* that the hidden Markov model states actually correspond to physical states of the biological system, but this not required.

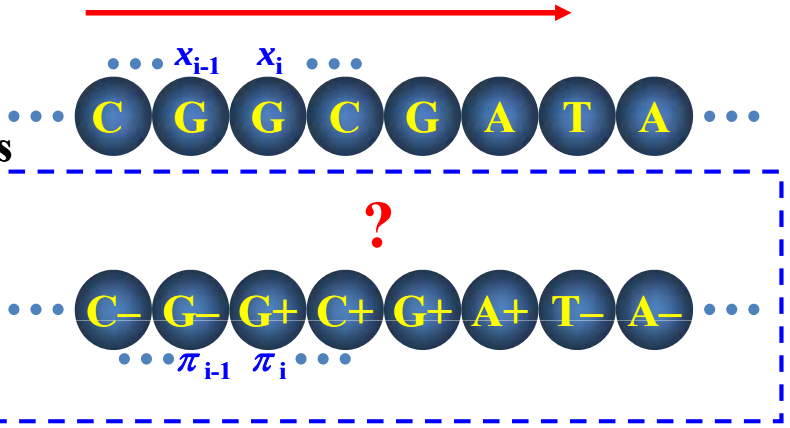
——《Computational Molecular Biology: An Introduction》

本质上说，所有模型都是错误的，但有些是有用的。

—— George Box

2. Viterbi算法——最大可能路径

观察序列
Observation sequences



隐状态序列
Underlying states

译码问题
Decoding

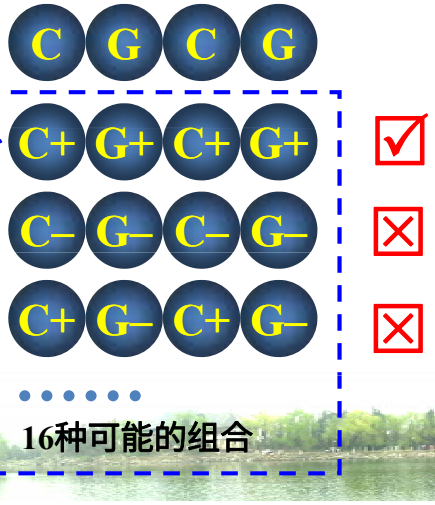
预测隐状态序列的准则

选择最大可能状态路径（Most probable state path），即观察序列$x_1 \ldots x_L$发生概率最高的路径，作为寻求的隐状态路径。

$$\pi^* = \arg\max_{\pi} P(x,\pi) = \arg\max_{\pi} P(\pi \mid x)$$

最大可能路径

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



递归方法与动态规划

根据递归（recursive）方法来求得最大可能路径。

定义$v_k(i)$为最大可能路径延伸到第i步(即产生前i个字串 $x_1 x_2 \ldots x_i$)，且序列第i步为状态k的概率。对各状态k的$v_k(i)$为：

$$v_k(i) = \max_{\{\pi|\pi_i=k\}} P(x_1, x_2, \ldots, x_i, \pi)$$

则沿状态序列，由动态规划原理得到第i+1步的状态为l的概率：

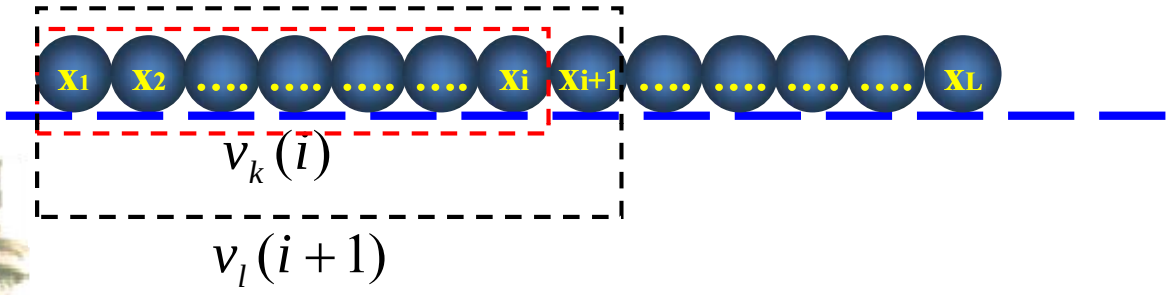
$$v_l(i+1) = e_l(x_{i+1}) \max_k \{v_k(i) a_{kl}\}$$

考虑添加起始状态，有：

$$v_0(0) = 1$$

Peking University
Introduction to Computational Biology

Recursive process



$v_k(i)$

$v_l(i+1)$

$$v_k(i) = \max_{\{\pi|\pi_i=k\}} P(x_1, x_2, ..., x_i, \pi)$$

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

Viterbi算法的基本步骤

Initialization (*i*=0):

$$v_0(0) = 1$$
$$v_k(0) = 0 \qquad for \qquad k > 0$$

Recursion (i=1, 2, …, L):

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

Termination:

$$P(x, \pi^*) = \max_k \{v_k(L) \cdot a_{k0}\}$$

$$\pi_L^* = \arg\max_k \{v_k(L) \cdot a_{k0}\}$$

Traceback (i=L, …, 1):

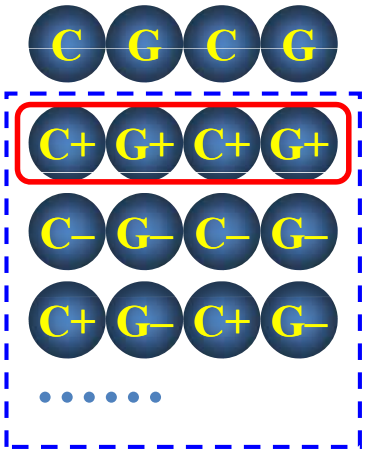
$$\pi_{i-1}^* = ptr_i(\pi_i^*)$$

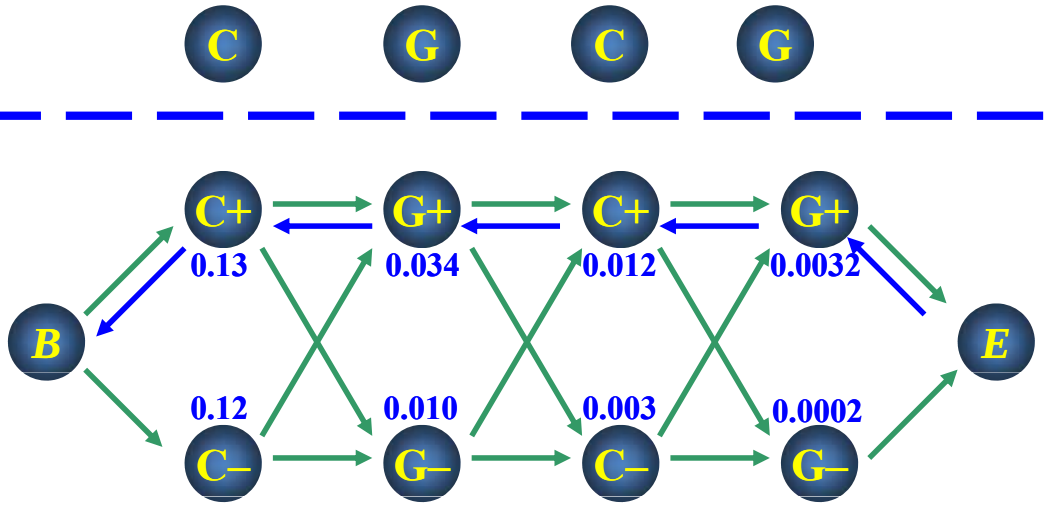
结束

Peking University
Introduction to Computational Biology

Viterbi算法应用举例：CpG岛模型

$v_k(i)$		C	G	C	G
B/E	1	0	0	0	0
A+	0	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	0.12	0	0.0026	0
G–	0	0	0.010	0	0.00021
T–	0	0	0	0	0





Peking University
Introduction to Computational Biology

With logarithmic scores

目的：为避免乘法运算中小数的影响，可采取对数形式。

Initialization (*i*=0):

$$v_0(0) = 0$$
$$v_k(0) = -\infty \qquad for \qquad k > 0$$

Recursion (i=1, 2, …, L):

$$v_l(i) = \log e_l(x_i) + \max_k \{v_k(i-1) + \log(a_{kl})\}$$
$$ptr_i(l) = \arg \max_k \{v_k(i-1) + \log(a_{kl})\}$$

Termination:

$$P(x,\pi^*) = \max_k \{v_k(L) + \log(a_{k0})\}$$

$$\pi_L^* = \arg\max_k \{v_k(L) + \log(a_{k0})\}$$

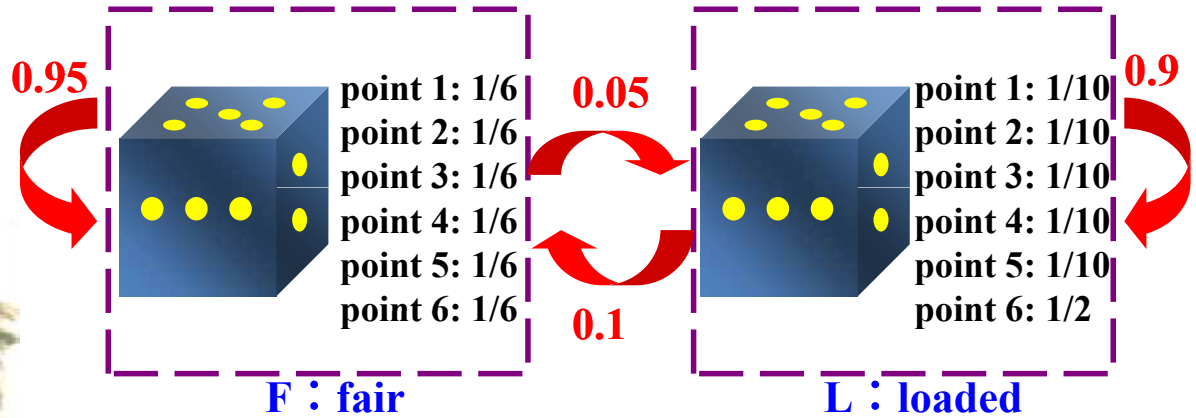
Traceback (i=L, …, 1):

$$\pi_{i-1}^* = ptr_i(\pi_i^*)$$

结束

Peking University
Introduction to Computational Biology

Viterbi算法应用举例：赌场骰子模型



状态：{F, L}　　输出符号：300次连续投掷的记录

$a_{FF} = 0.95$　$a_{FL} = 0.05$　$a_{LL} = 0.90$　$a_{LF} = 0.10$

$e_F(1) = e_F(2) = e_F(3) = e_F(4) = e_F(5) = e_F(6) = 1/6$

$e_L(1) = e_L(2) = e_L(3) = e_L(4) = e_L(5) = 0.1$

$e_L(6) = 0.5$

运用Viterbi算法根据已建立的HMM模型参数，对300次连续投掷的记录进行预测的结果（隐状态序列）之比较：

```
Rolls    315116246446644245311321631164152133625144543631656626566666
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLL

Rolls    651166453132651245636664631636663162326455236266666625151631
Die      LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLFFFLLLLLLLLLLLLLLFFFFFFFFF
Viterbi  LLLLLLFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFF

Rolls    222555441666566563564324364131513465146353411126414626253356
Die      FFFFFFFFLLLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFLL
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFL

Rolls    366163666466232534413661661163252562462255265252266435353336
Die      LLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
Viterbi  LLLLLLLLLLLLFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

Rolls    233121625364414432335163243633665562466662632666612355245242
Die      FFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
Viterbi  FFFFFFFFFFFFFFFFFFFFFFFFFFFFFLLLLLLLLLLLLLLLLLLLLLFFFFFFFFFF
```


3. 前向算法(The Forward Algorithm)

计算某DNA片段取某一Markov过程的可能性

某DNA序列取所有隐Markov状态过程的可能性

$$P(x) = P(x_1)\prod_{i=2}^{L} a_{x_{i-1}x_i}$$

识别CpG岛片段与非CpG岛片段

计算某一DNA序列$x_1 \ldots x_L$取全部可能隐状态路径的可能性。考虑所有可能性之和，即序列$x_1 \ldots x_L$取所有状态路径的全概率：

$$P(x) = \sum_{\pi} P(x, \pi) = \sum_{\pi} P(\pi \mid x)$$

上式计算量随状态路径数目呈指数增长，计算量太大。

假定隐状态路径必定是最大可能状态路径，则可以考虑采用最大可能状态路径的π^*概率来近似代替全概率P(*x*)：

$$\pi^* = \arg\max_{\pi} P(x, \pi) = \arg\max_{\pi} P(\pi \mid x)$$

Peking University
Introduction to Computational Biology

前向算法

实际上，仍可采取类似Viterbi算法的策略，根据动态规划原理，计算每一步的求和，最后得到全概率P(*x*)。

首先定义序列$x_1 \ldots x_i$的第 *i* 个位置上状态π_i为k的概率(全概率)：

$$f_k(i) = P(x_1 x_2 \ldots x_i, \pi_i = k)$$

递归表达式为：

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

前向算法的基本步骤

Initialization (*i*=0):

$$f_0(0) = 1$$

$$f_k(0) = 0 \qquad for \qquad k > 0$$

Recursion (i=1, 2, …, L):

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

Termination:

$$P(x) = \sum_k f_k(L)\cdot a_{k0}$$

结束

4. 后向算法、后验状态概率

Viterbi算法

前向算法

→ 考察产生观察序列$x_1 \ldots x_L$的概率

1、最大概率P(x, π*)
2、全概率P(x)

后向算法

→ 考察产生观察序列$x_1 \ldots x_L$中符号x_i的概率

后验状态概率
(Posterior state probability)
$P(\pi_i = k \mid x)$

间接求解：

首先考察下列概率

$$P(x, \pi_i = k) = P(x_1 \ldots x_L, \pi_i = k)$$
$$= P(x_1 \ldots x_i, \pi_i = k) \cdot P(x_{i+1} \ldots x_L \mid x_1 \ldots x_i, \pi_i = k)$$
$$= P(x_1 \ldots x_i, \pi_i = k) \cdot P(x_{i+1} \ldots x_L \mid \pi_i = k)$$

$f_k(i)$

$b_k(i) = P(x_{i+1} \ldots x_L \mid \pi_i = k)$

$$P(x_1 \ldots x_L, \pi_i = k) = f_k(i) \cdot b_k(i)$$

Peking University
Introduction to Computational Biology

后向算法(Backward algorithm)的基本步骤

Initialization (*i*=L):

$$b_k(L) = a_{k0} \quad \text{for all} \quad k > 0$$

Recursion (i=L-1, L-2, …, 1):

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

Termination:

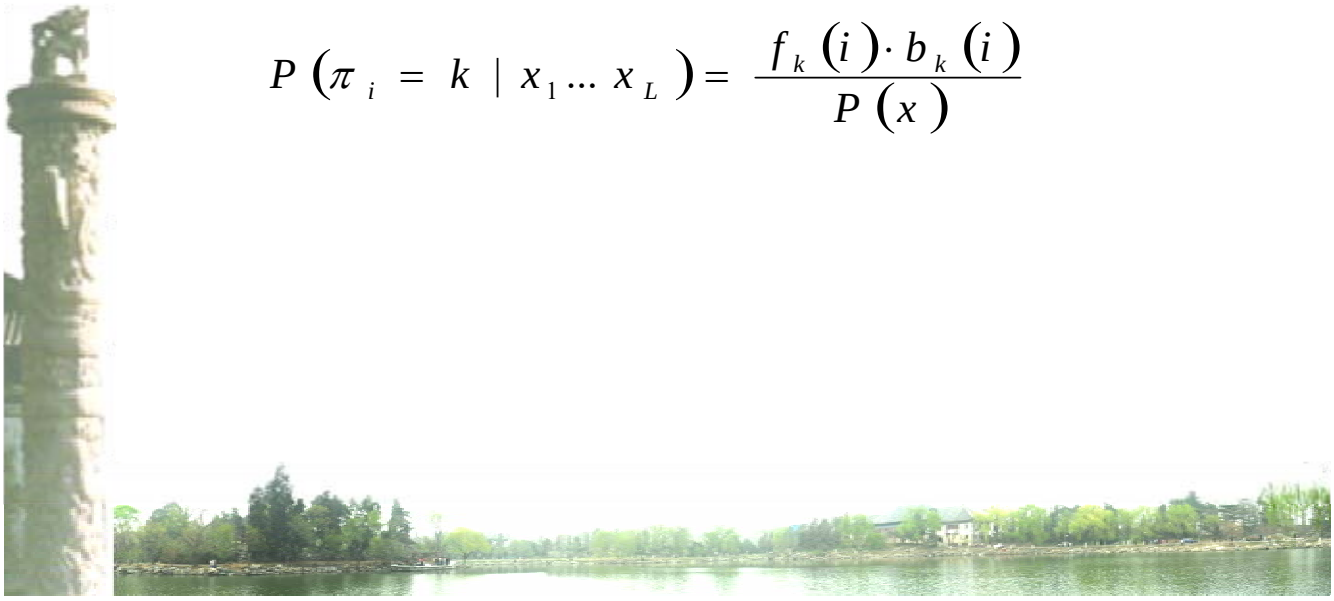
$$P(x) = \sum_l a_{k0} \cdot e_l(x_1) \cdot b_l(1)$$

$$\text{or} = \sum_l f_l(L) \cdot a_{k0}$$

结束

由此可以计算每一步的后验状态概率（Posterior state probability）：

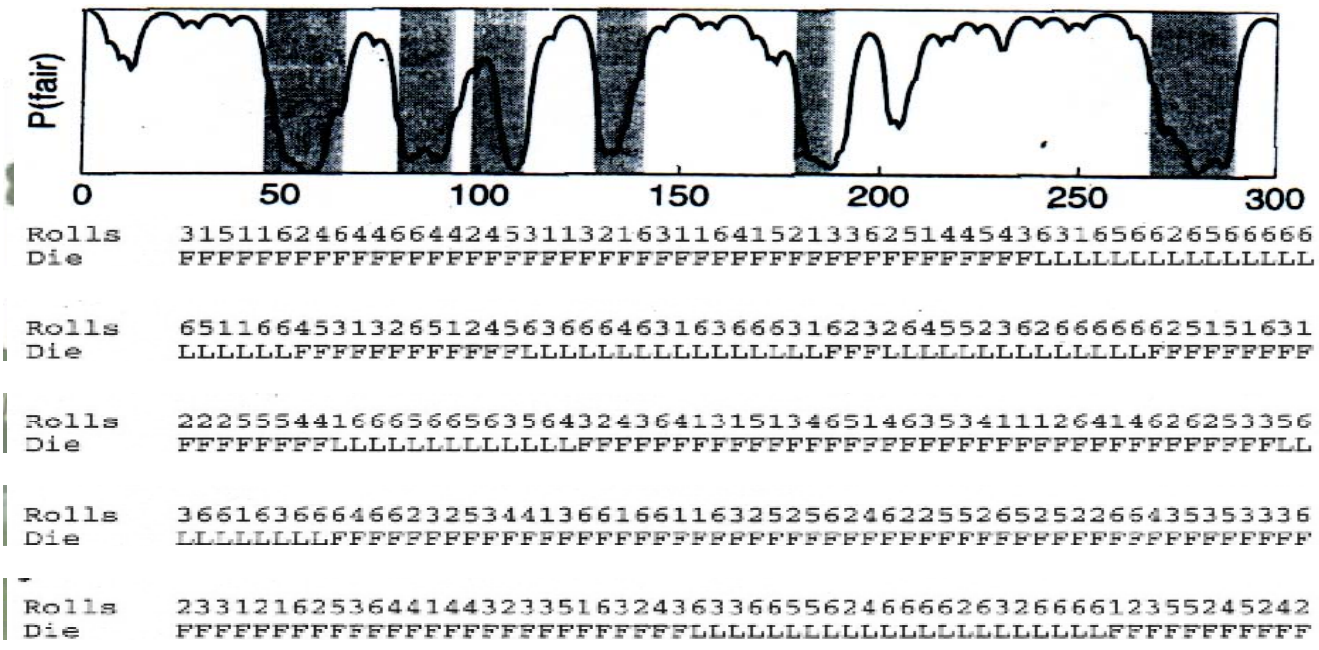
$$P(\pi_i = k \mid x_1 \ldots x_L) = \frac{f_k(i) \cdot b_k(i)}{P(x)}$$



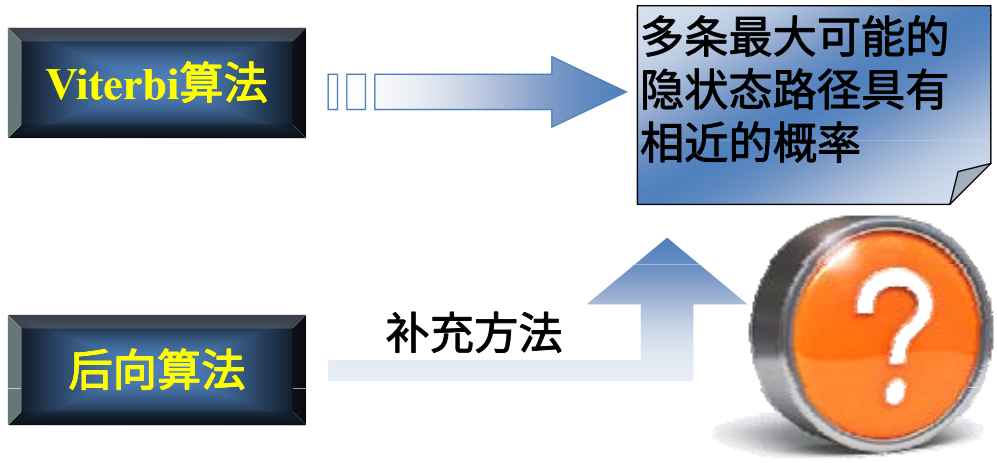
Peking University
Introduction to Computational Biology

后验状态概率算法举例：赌场骰子模型

假设每一次的观察点数都是由fair die投掷产生的，由此可计算已知观察序列中每一步的后验状态概率:



后验状态概率的应用——译码(Decoding)问题



Peking University
Introduction to Computational Biology

由此可以计算由下面每一步后验状态概率决定的序列:

$$\hat{\pi}_i = \arg\max_k P(\pi_i = k \mid x)$$

用以代替全局最大可能状态路径决定的序列：

$$\pi^* = \arg\max_{\pi} P(x, \pi)$$

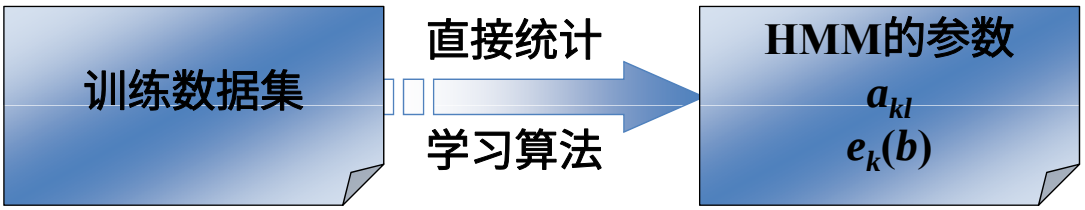
§7.3 隐Markov模型的参数估计与结构设计

状态转移概率a_{kl}：

$$a_{kl} = P(\pi_i = l \mid \pi_{i-1} = k)$$

输出概率$e_k(b)$：

$$e_k(b) = P(x_i = b \mid \pi_i = k)$$



Peking University
Introduction to Computational Biology

1. 直接统计得到模型参数
——训练集状态序列已知的情况

训练数据集（*n*条序列）：$x^1, x^2, \ldots, x^n$

当：

（1）训练集数据足够 ⇒ 满足最大似然估计的良好条件

（2）每一条序列的状态路径已知（给出所有观察序列的状态的注释信息）

可以直接统计得到HMM模型的各个参数。

例：

（1）CpG island：标注CpG岛区域位点（起始、终止）信息的DNA序列

（2）蛋白质二级结构：标注各类二级结构区域位点信息的氨基酸序列

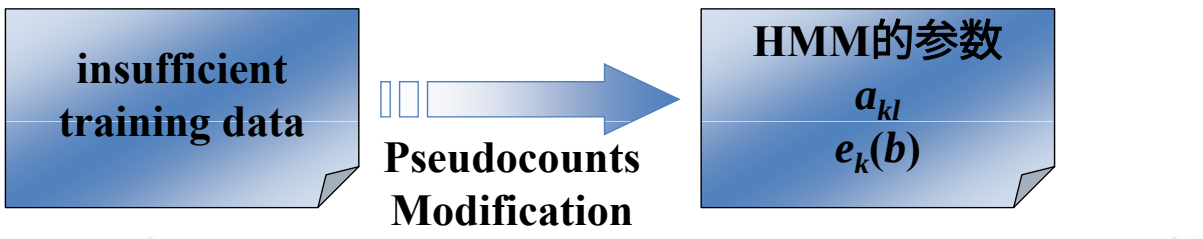
（3）编码基因预测：标注基因结构的各个位点

统计训练集中所有状态转移、符号输出的次数（频数）：

$$A_{kl} \qquad E_k(b)$$

根据最大似然估计原理，直接计算它们的频率作为概率值：

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}} \qquad e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$



Peking University
Introduction to Computational Biology

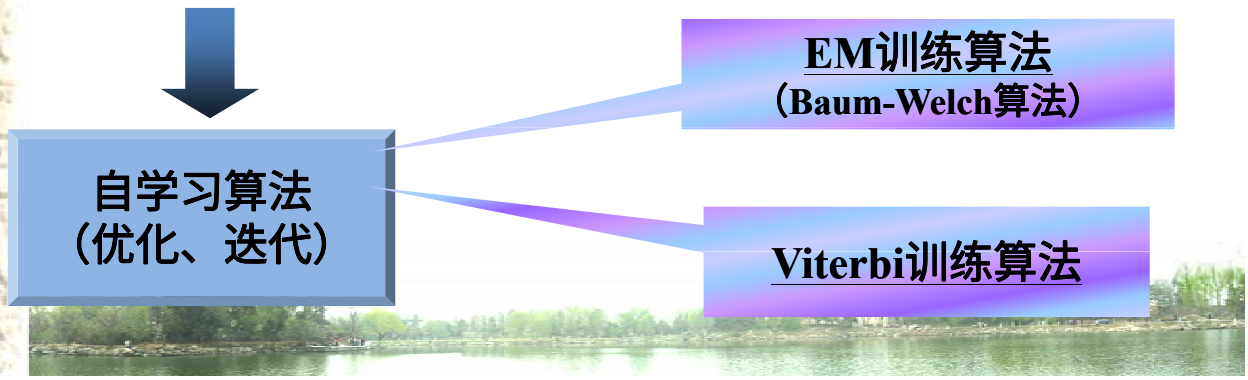
2. 自学习得到模型参数
——训练集状态序列未知的情况

训练数据集（*n*条序列）：$x^1, x^2, \ldots, x^n$

当：

（1）训练集数据足够 ⇒ 满足最大似然估计的良好条件

（2）每一条序列的状态路径未知，无法直接统计得到HMM模型的各个参数。



EM训练算法

主要的概率思想

迭代原理：根据当前已知的状态转移概率a_{kl}和符号输出概率$e_k(b)$，考虑训练集给定序列的所有可能路径，由此计算期望的状态转移和符号输出的次数A_{kl}、$E_k(b)$。再计算新的状态转移概率a_{kl}和符号输出概率$e_k(b)$，重复下一步的运算。

迭代终止。

Peking University
Introduction to Computational Biology

已知：训练数据集（n条序列）：$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$
假设给定当前已知的状态转移概率a_{kl}和符号输出概率$e_k(b)$
对每一条序列，计算下列概率值：

$$P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)}{P(x)}$$

其中θ表示HMM的模型参数。
由此计算期望的状态转移和符号输出的频数。

$$A_{kl} = \sum_{j=1}^{n} \frac{1}{P(x^j)} \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l(x_{i+1}^j) \cdot b_l^j(i+1)$$

$$E_k(b) = \sum_{j=1}^{n} \frac{1}{P(x^j)} \sum_{\{i \mid x_i^j = b\}} f_k^j(i) \cdot b_l^j(i)$$

$$a_{kl} = \frac{A_{kl}}{\sum_{l'} A_{kl'}}$$

$$e_k(b) = \frac{E_k(b)}{\sum_{b'} E_k(b')}$$

$$A_{kl} = \sum_{j=1}^{n} \frac{1}{P(x^j)} \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l\left(x_{i+1}^j\right) \cdot b_l^j(i+1)$$

$$E_k(b) = \sum_{j=1}^{n} \frac{1}{P(x^j)} \sum_{\{i|x_i^j=b\}} f_k^j(i) \cdot b_l^j(i)$$

Peking University
Introduction to Computational Biology

3. 隐Markov模型的结构设计

具体问题具体分析，需要对问题的深刻理解。

（略）

§7.4 隐Markov模型在计算生物学中的应用

- 两序列比对 (pairwise alignment)
- Profile HMM
  分类
  数据库搜索
  基于profile的多序列比对
- 基因识别
  Genscan (Burge *et al*, 1997)
  GeneMark series (Borodovsky *et al*, 1993-2001)
- Motif 搜寻, 信号肽识别, 蛋白质结构预测…

Peking University
Introduction to Computational Biology

1. 两序列比对

给定两条氨基酸序列X, Y:

X: VLSPADK
Y: HLAESK

假定XY是同源序列。进化过程中，由于插入（Insertion）和 缺失（Deletion） 的发生，两序列有多种可能的对齐形式，只有一种是真实的比对（alignment）。

一种可能的对齐形式：

X: V L S P A D - K
Y: H L - - A E S K

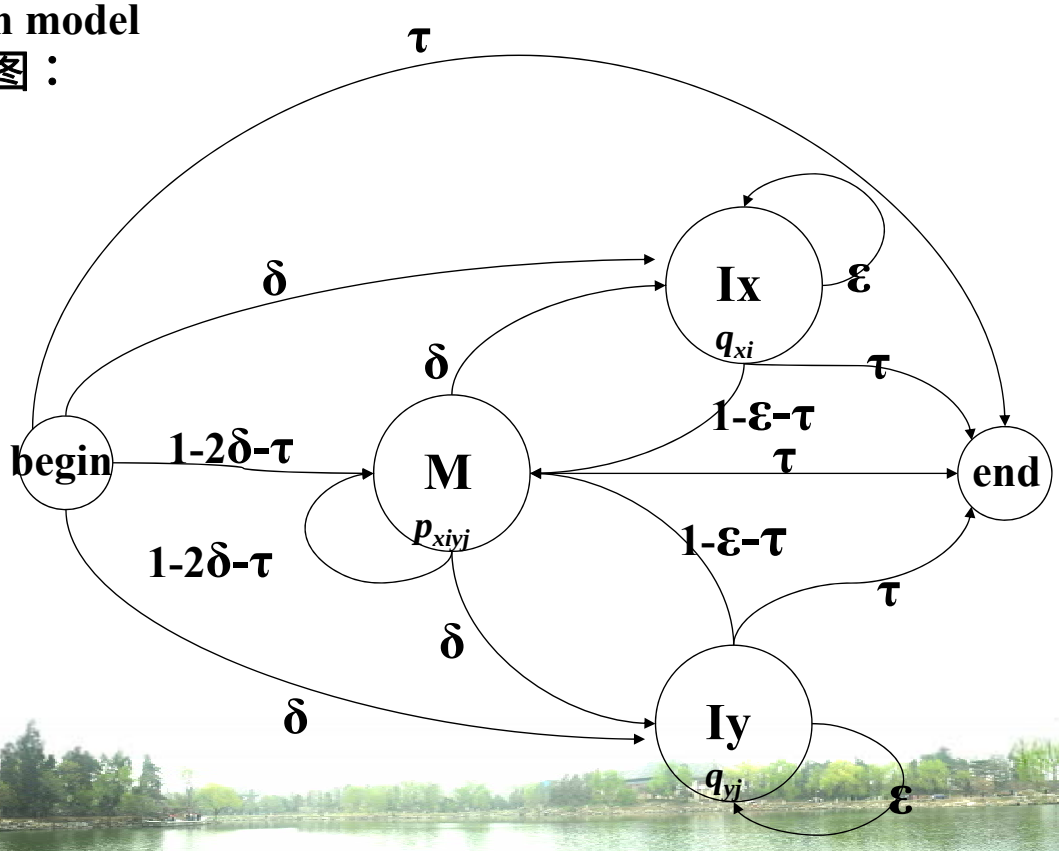
对alignment进行建模 (a match model)：

X: V L S P A D - K
Y: H L - - A E S K
M M Ix Ix M M Iy M

共有三种状态构成了模型的状态空间 {M, Ix, Iy}。
M (match)
Ix (insertion in X)
Iy (insertion in Y)；
任意一个alignment的结果可以看做是对应隐状态链的输出 (emission)。状态之间可以以一定概率转移 (transition)。

X: V L S P A D - K
Y: H L - - A E S K
M → M → Ix → Ix → M → M → Iy → M

Full match model
状态转移图：



回到给定的 amino acid 序列 X 和 Y：

X: VLSPADK
Y: HLAESK

假定是同源的，如何从众多的对齐形式中找到真实的比对？

所建造的模型是一个概率模型，描述了所有同源序列所有可能的对齐形式，并对每一个赋予概率：

$$P(x <> y, \pi)$$

选择概率最大的为最终比对：

$$\pi^* = \arg\max_{\pi} P(x <> y, \pi) = \arg\max_{\pi} \underbrace{P(x <> y \mid \pi)}_{\text{emissions}} \cdot \underbrace{P(\pi)}_{\text{transitions}}$$

Peking University
Introduction to Computational Biology

动态规划算法寻找最优路径 (Viterbi algorithm).

注意: 由于是两条序列，搜索空间为二维（i,j）,i=0,…,7; j=0,…,6.

递归式：

$$v^M(i,j) = p_{x_i y_j} \max \begin{cases} (1-2\delta-t)v^M(i-1,j-1), \\ (1-\varepsilon-t)v^{Ix}(i-1,j-1), \\ (1-\varepsilon-t)v^{Iy}(i-1,j-1); \end{cases}$$

$$v^{Ix}(i,j) = q_{x_i} \max \begin{cases} \delta v^M(i-1,j), \\ \varepsilon v^{Ix}(i-1,j); \end{cases}$$

$$v^{Iy}(i,j) = q_{y_j} \max \begin{cases} \delta v^M(i,j-1), \\ \varepsilon v^{Iy}(i,j-1); \end{cases}$$

再回到序列X, Y:

X: VLSPADK

Y: HLAESK

问题：它们是否同源；或者它们是同源的可能性是多少？

单一的match 模型给出了假定是同源的情况下，X,Y最可能具有的比对形式π^*。以这一比对形式出现的同源序列的概率为：

$$P(x <> y, \pi^* \mid M)$$

现在考察这一比对的significance，需要另外一个模型——随机模型R。R刻画了所有可能的两序列对，并赋予一定的概率

$$P(x, y \mid R)$$

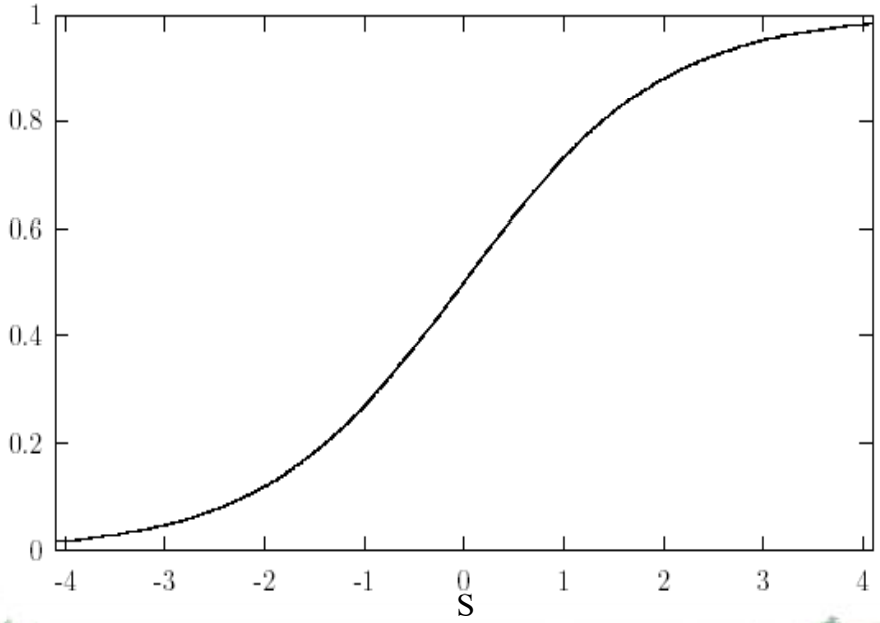
我们考察两概率的log-odd 值：

$$S = \log \frac{P(x <> y, \pi^* \mid M)}{P(x, y \mid R)} + S'$$

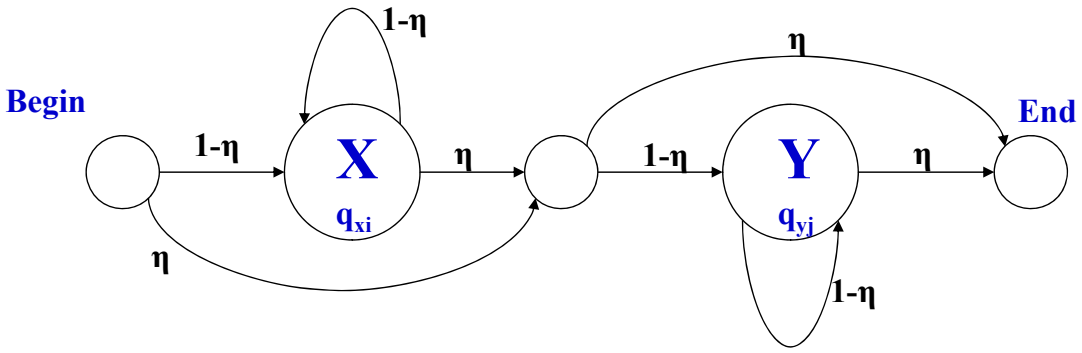
Peking University
Introduction to Computational Biology

Logistic 函数

$$\frac{1}{1+e^{-s}}$$



Full random model
状态转移图:



重新做比对，考虑Random model的影响，即

$$\pi^* = \arg\max_{\pi} \log(\frac{P(x <> y, \pi)}{P(x, y \mid R)}) = \arg\max_{\pi} \log(\frac{P(x <> y \mid \pi) \cdot P(\pi)}{P(x, y \mid R)})$$

根据模型假设：

$$P(x, y \mid R) = \eta(1-\eta)^7 \prod_{i=1}^{7} q_{x_i} \cdot \eta(1-\eta)^6 \prod_{j=1}^{6} q_{y_j}$$

将其各项融进match model 的对应项中：

M: $\log \frac{p_{x_i y_j}}{q_{x_i} q_{y_j}} + \log \frac{a_{lM}}{(1-\eta)^2}$

Ix: $\log \frac{a_{lIx}}{1-\eta}$

Iy: $\log \frac{a_{lIy}}{1-\eta}$

定义：$s(a,b) = \log \frac{p_{ab}}{q_a q_b} + \log \frac{(1-2\delta-\tau)}{(1-\eta)^2}$

$d = -\log \frac{\delta(1-\varepsilon-\tau)}{(1-\eta)(1-2\delta-\tau)}$

$e = -\log \frac{\varepsilon}{1-\tau}$

Log-odd 版本的 Viterbi 算法

递归式：

$$V^M(i,j) = s(x_i, y_j) + \max\begin{cases} V^M(i-1,j-1), \\ V^{Ix}(i-1,j-1), \\ V^{Iy}(i-1,j-1); \end{cases}$$

$$V^{Ix}(i,j) = \max\begin{cases} V^M(i-1,j) - d, \\ V^{Ix}(i-1,j) - e; \end{cases}$$

$$V^{Iy}(i,j) = \max\begin{cases} V^M(i,j-1) - d, \\ V^{Iy}(i,j-1) - e; \end{cases}$$

Peking University
Introduction to Computational Biology

简单情况：
s(a,b) 为替换矩阵
d 为开gap罚分
e为gap 延长罚分

例如：
取BLOSUM50替换矩阵；
affine gap penalty 为
d=10, e=2

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0	-2	-1	-1	-5
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3	-1	0	-1	-5
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3	4	0	-1	-5
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4	5	1	-1	-5
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1	-3	-3	-2	-5
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3	0	4	-1	-5
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3	1	5	-1	-5
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4	-1	-2	-2	-5
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4	0	0	-1	-5
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4	-4	-3	-1	-5
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1	-4	-3	-1	-5
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3	0	1	-1	-5
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1	-3	-1	-1	-5
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1	-4	-4	-2	-5
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3	-2	-1	-2	-5
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2	0	0	-1	-5
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0	0	-1	0	-5
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3	-5	-2	-3	-5
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1	-3	-2	-1	-5
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5	-4	-3	-1	-5
B	-2	-1	4	5	-3	0	1	-1	0	-4	-4	0	-3	-4	-2	0	0	-5	-3	-4	5	2	-1	-5
Z	-1	0	0	1	-3	4	5	-2	0	-3	-3	1	-1	-4	-1	0	-1	-2	-2	-3	2	5	-1	-5
X	-1	-1	-1	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-1	0	-3	-1	-1	-1	-1	-1	-5
*	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	-5	1

比对的分值：

X: V L S P A D - K
Y: H L - - A E S K

S(V,H)+S(L,L)-d-(2-1)e+S(A,A)+S(D,E)-d+S(K,K)
=-4+5-10-2+5+2-10+6
=-8

Peking University
Introduction to Computational Biology

In fact, the search and alignment applications constitute probably the best-known use of HMM for biological sequence analysis. However, HMM theory should be emphasized with its much broader applicability, which goes far beyond that of sequence alignment.

——《Biological sequence analysis》

2. 蛋白质家族的统计描述
（Profile HMM）

一个蛋白家族 (family) 包含有众多的成员蛋白(member)。成员蛋白之间序列相差较大，但结构比较保守，功能也相似或相关，因为它们来自共同的祖先。

蛋白家族的profile即指，对所有成员蛋白共同具有的特征进行统计性的描述。在寻找新的成员蛋白时，用这个profile对数据库中的每一个candidate进行打分，以此判断其是否为该家族成员。

用HMM的方法对某一蛋白家族进行统计描述，称为profile HMM.

Peking University

```
CSRC_HUMAN          1    42   KLGQGCFGEVWMGTW--------NGTTRVAIKTLKP------GTMSPEAF----LQEAQV
CABL_HUMAN          1    43   KLGGGQYGEVYEGVW-------KKYSLTVAVKTLKE------DTMEVEEF----LKEAAV
EPH_HUMAN           1    47   -IGEGEFGEVYRGTLR----LPSQDCKTVAIKTLKDT----SPGGQWWNF----LREATI
FER_HUMAN           1    44   LLGKGNFGEVYKGTL--------KDKTSVAVKTCKED----LPQELKIKF----LQEAKI
IR_HUMAN            1    48   --GQGSFGMVYEGNARDI--IKGEAETRVAVKTVNES----ASLRERIEF----LNEASV
CROS_HUMAN          1    49   --GSGAFGEVYEGTAVDI-LGVGSGEIKVAVKTLKKG----STDQEKIEF----LKEAHL
TRK_HUMAN           1    47   --GEGAFGKVFLAECHNL--LPEQDKMLVAVKALKE-----ASESARQDF----QREAEL
BFGFR_HUMAN         1    50   --GEGCFGQVVLAEAIGLDKDKPNRVTKVAVKMLKSD----ATEKDLSDL----ISEMEM
RET_HUMAN           1    48   --GEGEFGKVVKATAFHL--KGRAGYTTVAVKMLKEN----ASPSELRDL----LSEFNV
EGFR_HUMAN          1    47   --GSGAFGTVYKGLWIP---EGEKVKIPVAIKELREA----TSPKANKEI----LDEAYV
SCH9_YEAST          1    47   LLGKGTFGQVYQVKK-------KDTQRIYAMKVLSKKVI--VKKNEIAHT----IGERNI
S6K_70K_RAT         1    50   -LGKGGYGKVFQVRKV----TGANTGKIFAMKVLKKAMI-VRNAKDTAHT----KAERNI
PKC1_YEAST          1    47   VLGKGNFGKVILSKS-------KNTDRLCAIKVLKKDNI--IQNHDIESA----RAEKKV
CGMPK_HUMAN         1    47   -LGVGGFGRVELVQL------KSEESKTFAMKILKKRHI--VDTRQQEHI----RSEKQI
PRK_RABBIT          1    53   ILGRGVSSVVRRCIH-------KPTCKEYAVKIIDVTGGGSFSAEEVQELREATLKEVDI

CSRC_HUMAN         43    80   M---KKL-R-HEKLVQLYAVVSE-EPIYIVTEYMSKGSLLDFLK
CABL_HUMAN         44    80   M---KEI-K-HPNLVQLLGVCTREPPFYIITEFMTYGNLLDY--
EPH_HUMAN          48    80   M---GQF-S-HPHILHLEGVVTKRKPIMIITEFMENGA------
FER_HUMAN          45    80   L---KQY-D-HPNIVKLIGVCTQRQPVYIIMELVSGGDFLT---
IR_HUMAN           49    80   M---KGF-T-CHHVVRLLGVVSKGQPTLVVMELMAHG-------
CROS_HUMAN         50    80   M---SKF-N-HPNILKQLGVCLLNEPQYIILELMEG--------
TRK_HUMAN          48    80   L---TML-Q-HQHIVRFFGVCTEGRPLLMVFEYMRHGD------
BFGFR_HUMAN        51    80   M---KMIGK-HKNIINLLGACTQDGPLYVIVEYA----------
RET_HUMAN          49    80   L---KQV-N-HPHVIKLYGACSQDGPLLLIVEYAKYG-------
EGFR_HUMAN         48    80   M---ASV-D-NPHVCRLLGICLT-STVQLITQLMPFGCL-----
SCH9_YEAST         48    80   LVTTASK-S-SPFIVGLKFSFQTPTDLYLVTDYMS---------
S6K_70K_RAT        51    80   L---EEV-K-HPFIVDLIYAFQTGGKLYLILEYLS---------
PKC1_YEAST         48    80   FLLATKT-K-HPFLTNLYCSFQTENRIYFAMEFIG---------
CGMPK_HUMAN        48    80   M---QGA-H-SDFIVRLYRTFKDSKYLYMLMEACLGGE------
PRK_RABBIT         54    80   L---RKV-SGHPNIIQLKDTYETNTFFFLVF-------------
```

An alignment of 15 kinase: human, rabbit, rat & yeast; retrieved from PAPIA server.

怎样建profile HMM?

首先对已知的成员蛋白做多序列比对 (基于结构或基于序列).
注意：必须保证得到的比对是正确的，因为下一步将对其进行建模，然后搜寻新的成员蛋白。

然后对比对好的多重序列进行建模 (consensus modeling)。对于序列的每一个位置都有三种可能的状态：match、insertion、deletion。
Match状态（或者说main状态）处于哪些位置必须确定下来，以此确定模型的拓扑结构。
比如：

```
WMGTW--------NGTTRVA
YEGVW-------KKYSLTVA
YRGTLR---LPSQDCKTVA
YKGTL--------KDKTSVA
YEGNARDI--IKGEAETRVA
YEGTAVDI-LGVGSGEIKVA
FLAECHNL--LPEQDKMLVA
VLAEAIGLDKDKPNRVTKVA
VKATAFHL--KGRAGYTTVA
YKGLWIP---EGEKVKIPVA
YQVKK------KDTQRIYA
FQVRKV---TGANTGKIFA
ILSKS------KNTDRLCA
ELVQL------KSEESKTFA
RRCIH------KPTCKEYA
******     **********
```


profile HMM 的状态转移图

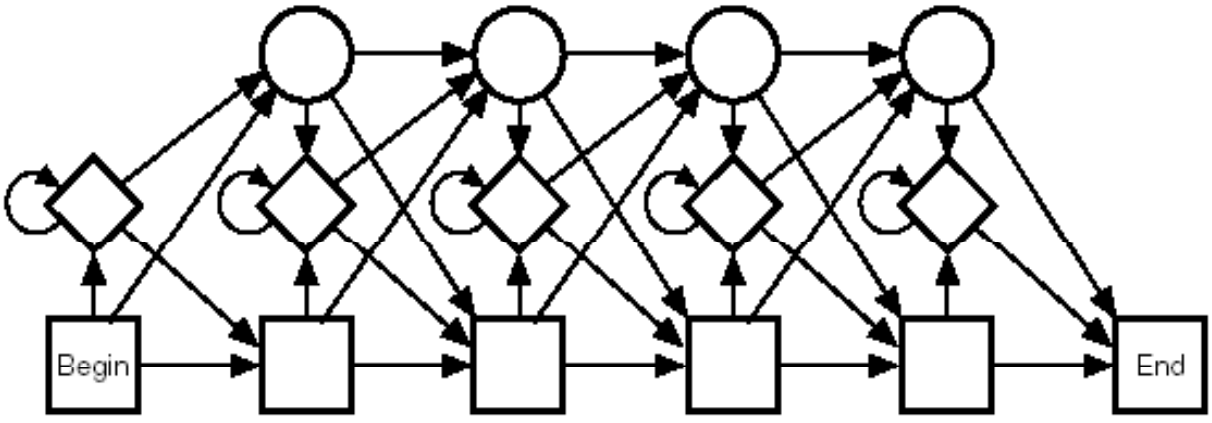


Figure from *Computational methods in molecular biology*, Chap. 4, by Anders Krogh.

: Match　　: Insertion　　: Deletion

参数估计

(Transition probabilities and emission probabilities)

直接从多道比对中，通过统计得出。

由于样本有限，一些允许的事件出现的次数为零——运用Laplace准则: 对每一个可能的事件都事先加一(pseudo-count=1)。

例如：

```
WMGTW--------NGTTRVA
YEGVW-------KKYSLTVA
YRGTLR---LPSQDCKTVA
YKGTL--------KDKTSVA
YEGNARDI--IKGEAETRVA
YEGTAVDI-LGVGSGEIKVA
FLAECHNL--LPEQDKMLVA
VLAEAIGLDKDKPNRVTKVA
VKATAFHL--KGRAGYTTVA
YKGLWIP---EGEKVKIPVA
YQVKK-------KDTQRIYA
FQVRKV---TGANTGKIFA
ILSKS-------KNTDRLCA
ELVQL------KSEESKTFA
RRCIH------KPTCKEYA
******     **********
position   1 2 3 4 5...
```

Peking University
Introduction to Computational Biology

$e_{M_1}(W) = 2/35;$ $a_{M_1M_2} = 16/18;$

$e_{M_1}(Y) = 8/35;$ $a_{M_1I_1} = 1/18;$

$e_{M_1}(F) = 3/35;$ $a_{M_1D_2} = 1/18.$

$e_{M_1}(V) = 3/35;$

$e_{M_1}(I) = 2/35;$

$e_{M_1}(E) = 2/35;$

$e_{M_1}(R) = 2/35;$

其它的为 $1/35.$ …

数据库搜寻

寻找新的成员蛋白(significant matches to the profile):

首先运用动态规划方法将每一个可能的匹配 (match) 同已有的比对 (profile) 对齐。

$$\pi^* = \arg\max_{\pi} P(x, \pi \mid profile)$$

这个匹配 是家族成员的似然性(likelihood)即是：

$$P(x, \pi^* \mid profile)$$

为了考察匹配的significance，需要一随机模型进行比较——log-odd 版本的Viterbi算法.

一个标准的随机模型产生这条序列 (match) 的概率为：

$$P(x \mid R) = \prod_i q_{x_i}$$

Peking University
Introduction to Computational Biology

Log-odd 版本的Viterbi 递归式：

$$V_j^M(i) = \log\frac{e_{M_j}(x_i)}{q_{x_i}} + \max\begin{cases} V_{j-1}^M(i-1) + \log a_{M_{j-1}M_j}, \\ V_{j-1}^I(i-1) + \log a_{I_{j-1}M_j}, \\ V_{j-1}^D(i-1) + \log a_{D_{j-1}M_j}; \end{cases}$$

$$V_j^I(i) = \log\frac{e_{I_j}(x_i)}{q_{x_i}} + \max\begin{cases} V_j^M(i-1) + \log a_{M_jI_j}, \\ V_j^I(i-1) + \log a_{I_jI_j}, \\ V_j^D(i-1) + \log a_{D_jI_j}; \end{cases}$$

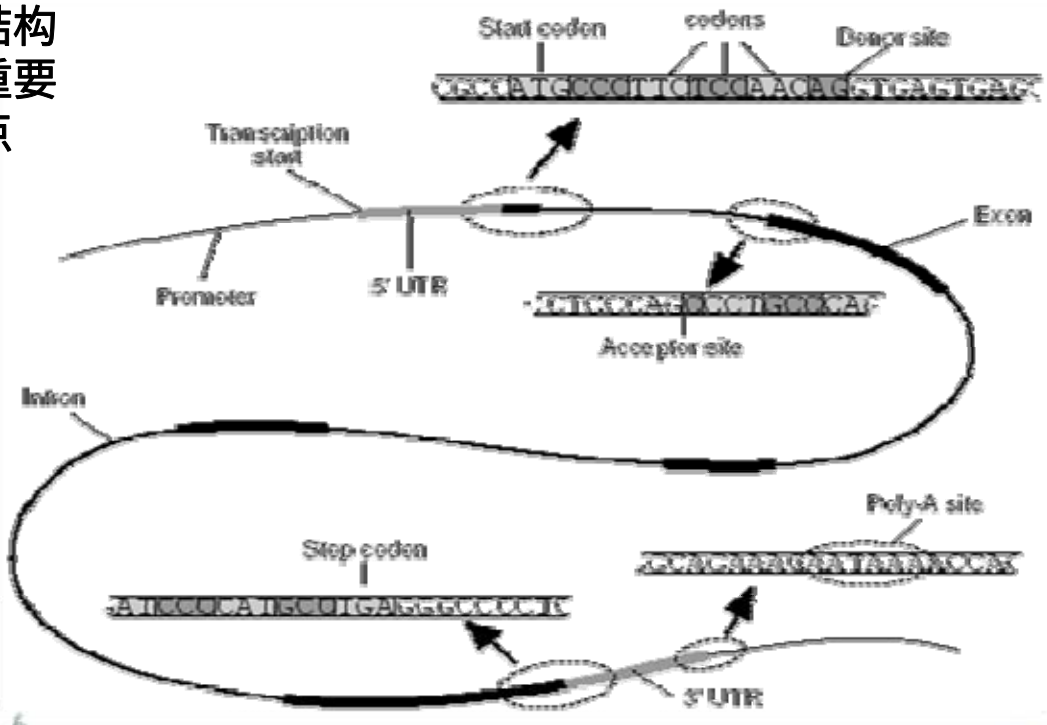
$$V_j^D(i) = \max\begin{cases} V_{j-1}^M(i) + \log a_{M_{j-1}D_j}, \\ V_{j-1}^I(i) + \log a_{I_{j-1}D_j}, \\ V_{j-1}^D(i) + \log a_{D_{j-1}D_j}; \end{cases}$$

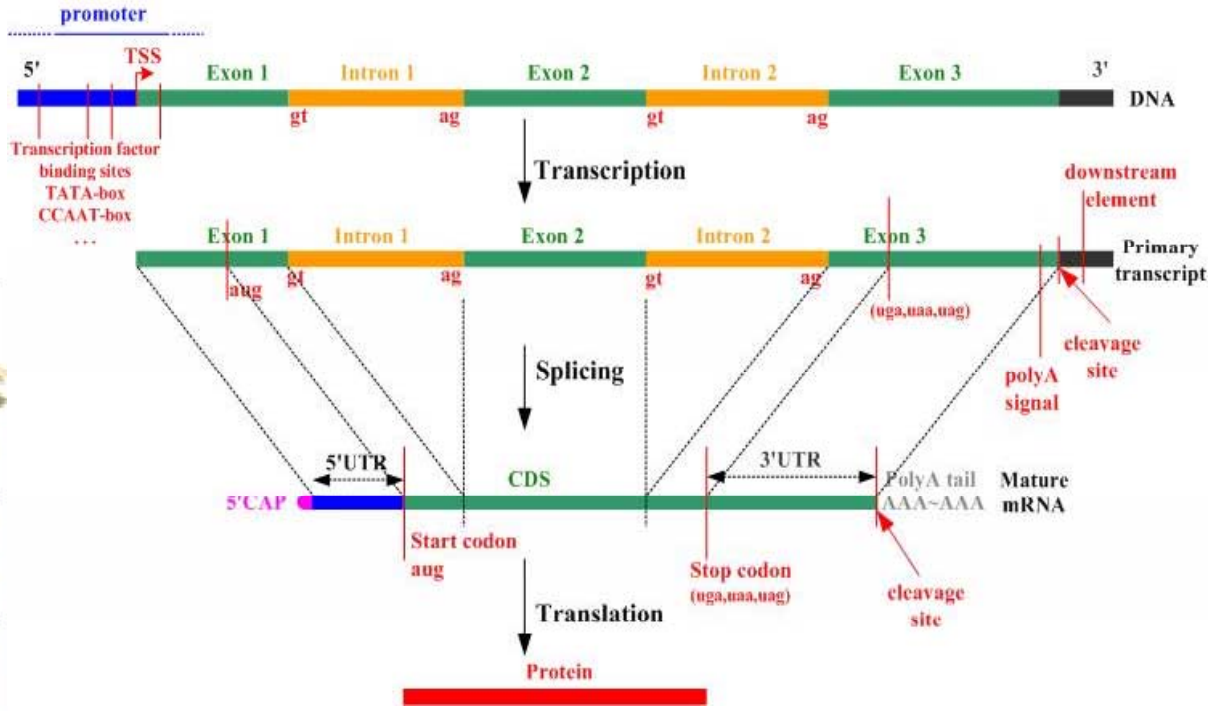
3. 基于HMM的基因识别与预测

- GENSCAN (Burge 1997)
- FGENESH (Solovyev 1997)
- HMMgene (Krogh 1997)
- GENIE (Kulp 1996)
- GENMARK (Borodovsky & McIninch 1993)
- VEIL (Henderson, Salzberg, & Fasman 1997)

Peking University
Introduction to Computational Biology

真核基因结构以及一些重要的调控位点





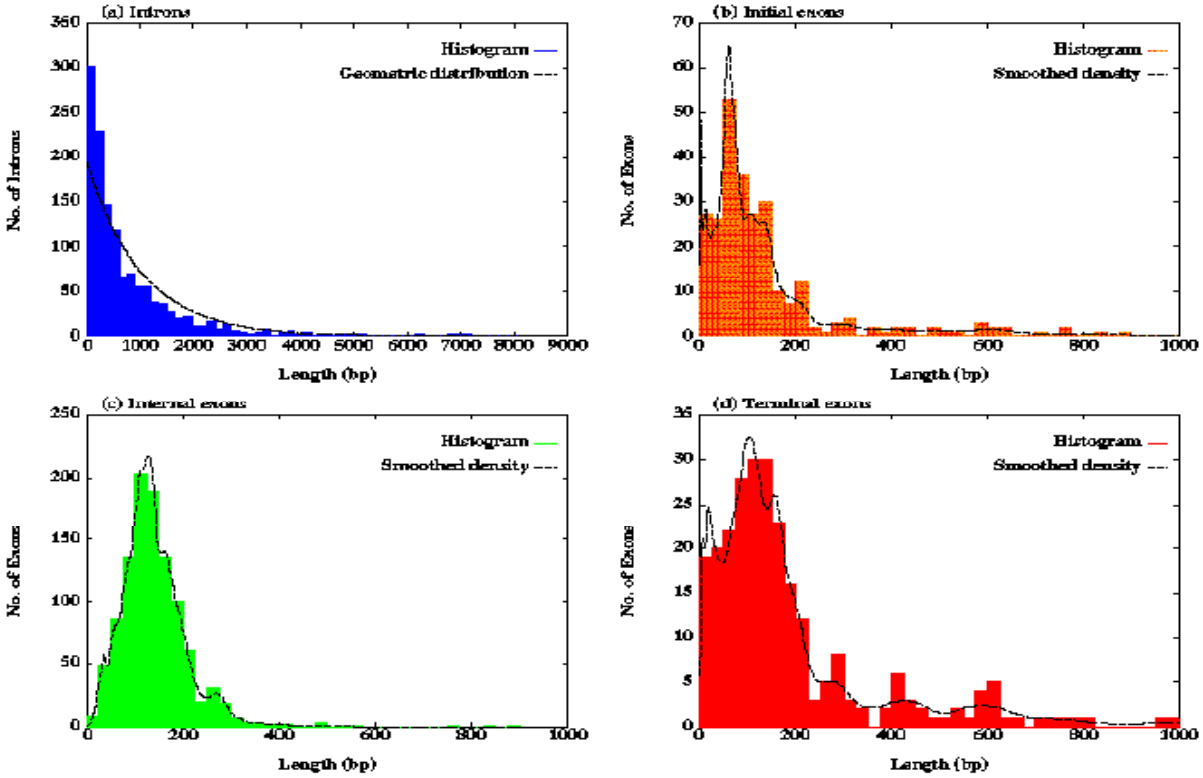
http://online.itp.ucsb.edu/online/infobio01/burge/

Peking University
Introduction to Computational Biology

初步统计

- 脊椎动物基因平均长度为 30kb。
- 编码区大约长 1kb。
- 外显子长度变化很大，从几十到几千个碱基。
- 5’非翻译区 (5’ UTR) 大约长 750 bp。
- 3’非翻译区 (3’UTR) 大约长 450 bp。

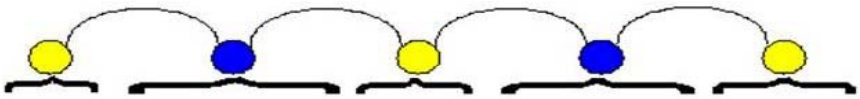
人基因的内含子和三类外显子（第一个、中间的、以及最后一个）的长度分布图



Peking University
Introduction to Computational Biology

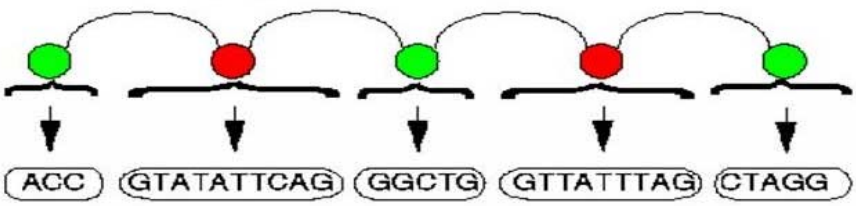
半马氏以及半隐马氏

半马氏链



状态链为Markov链；
每个状态有特定的停留时间。

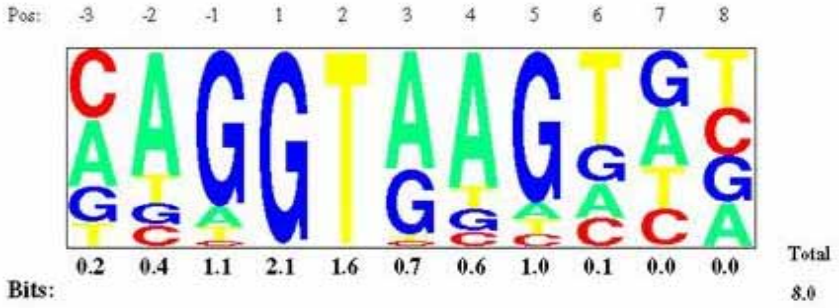
半隐马氏链



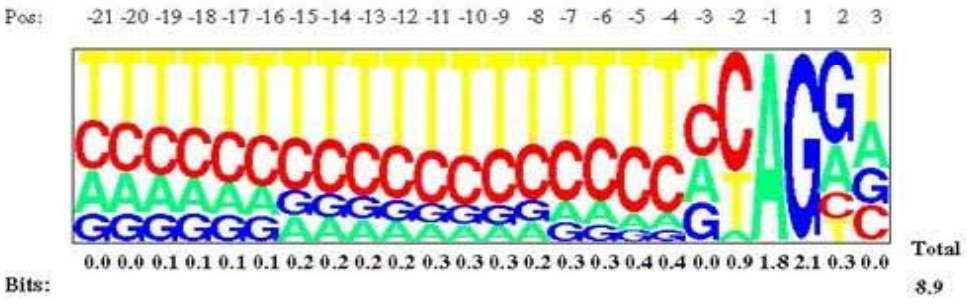
状态不能被观察到；
观察序列由隐状态产生，长度同隐状态对应的停留时间。

人类基因剪接位点信号

5’ 剪接信号

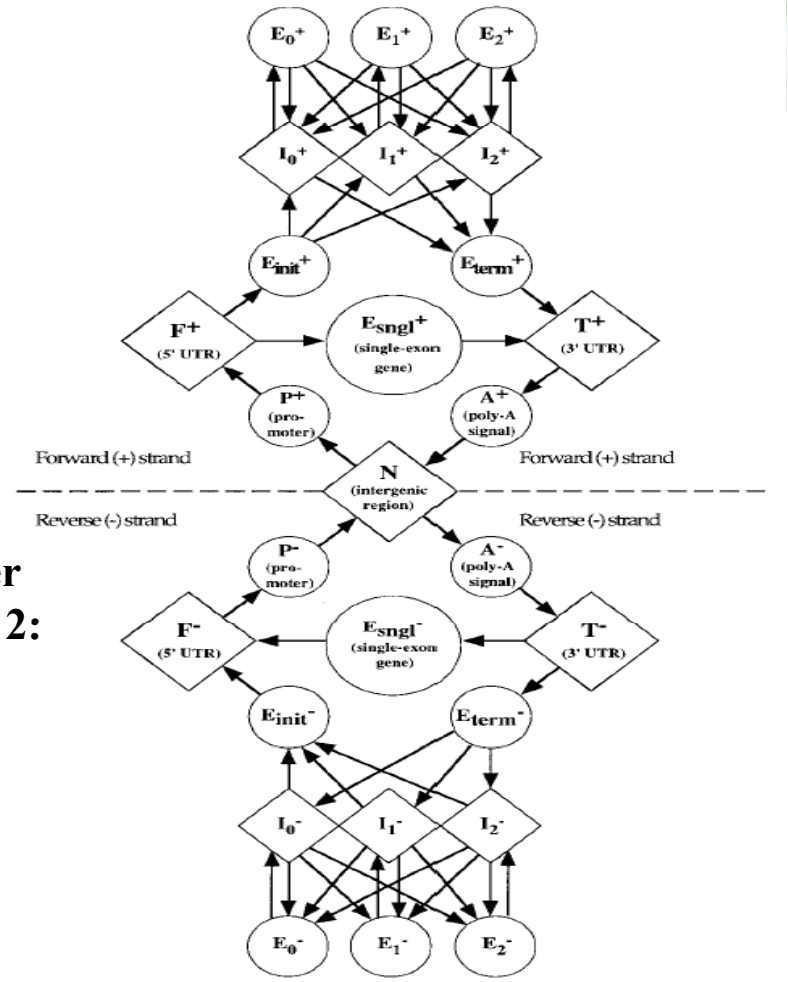


3’ 剪接信号



- N – 基因间区
- P – 启动子
- F - 5’ 非翻译区
- E_{sngl} – 单外显子基因
- E_{init} – 第一个外显子
- E_k – k相位的中间外显子
- E_{term} – 最后一个外显子
- I_k – k相位的内含子

(0:between codons; 1:after the first base of a codon; 2: after the second base of a codon)



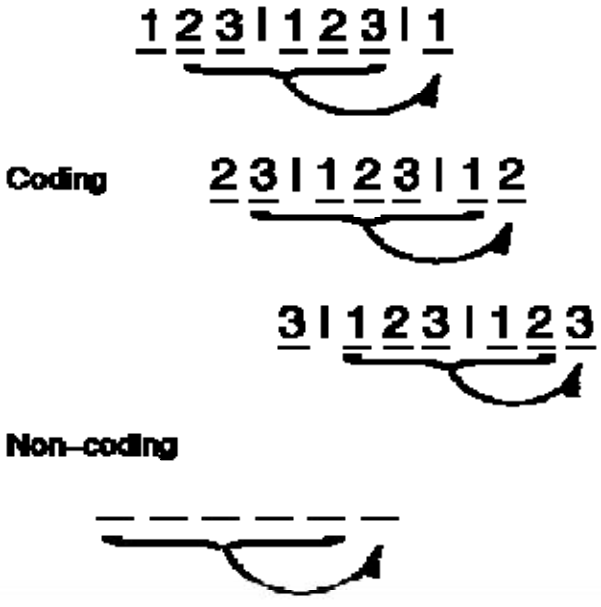
Burge, C. & Karlin, S, *J. Mol. Biol.* (1997) 268, 78-94.

GenScan 信号建模

- WMM (Weight Matrix 模型)
  - Poly A;
  - 翻译起始和终止位点;
  - 启动子;
- WAM (Weight Array 模型)
  acceptor 剪接位点 (2阶)
- MDD (Maximal Dependence Decomposition)
  donor 剪切位点

Peking University
Introduction to Computational Biology

编码与非编码区的建模



GenScan 的特点

1、对两条链（正，负）同时建模;

2、每一个状态可以产生一串的核苷酸序列;

3、考虑内含子与外显子的长度分布;

4、对剪切位点做单独的建模;

5、直接从带有标注（annotation）的学习集中提取参数。

Peking University
Introduction to Computational Biology

4. 隐Markov模型的讨论

HMM的局限

自由参数过多、对学习集要求较高

模型的物理意义？

HMM的优势

坚实的统计学基础

有效的学习算法