Markov模型及其在生物信息学中的应用

王 萌

北京大学生命科学学院生物信息中心(CBI)

马尔可夫模型

- 问题起源：一个事件是独立发生的，还是与之前的事件有关？

- Markov(俄) 于1906年首先提出马尔可夫过程。

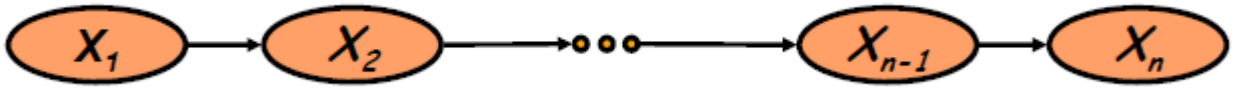
- 天气预报：用今天、昨天、前天…的天气预测明天的天气

- 马尔可夫性

  在一个随机过程中，如果未来某个事件发生的概率仅与其前一个事件有关，而与这之前的其他事件都无关，则该随机过程具有马尔可夫性。

马尔可夫链

- 具有马尔可夫性的离散随机过程



$$P(x_i|x_1, x_2, x_3, \dots, x_{i-1}) = P(x_i|x_{i-1})$$

- K阶马尔可夫过程

$$P(x_i|x_1, x_2, x_3, \dots, x_{i-1}) = P(x_i|x_{i-1}, x_{i-2}, \dots, x_{i-k})$$

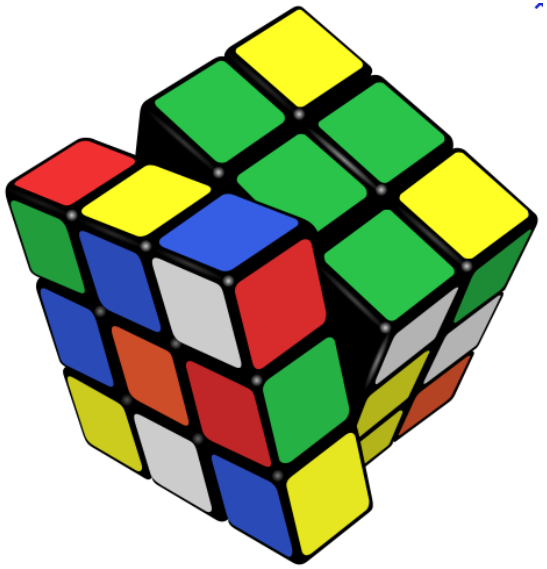
Markov模型应用

马尔可夫模型及后面要讲的隐马尔可夫模型广泛应用在

- 生物信息
- 人工智能
- 语音识别
- 机器翻译
- 汉字输入

在生物信息学中目前主要应用于

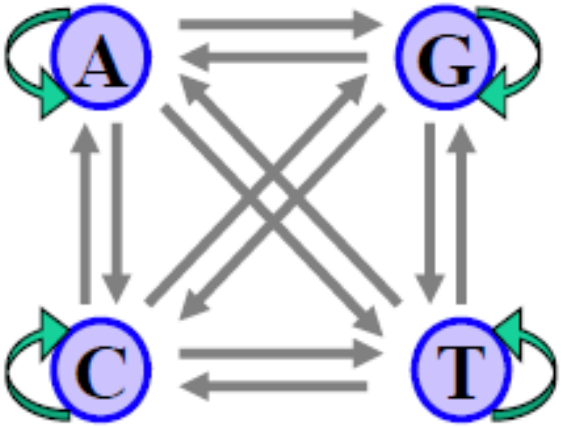
- 基因预测
- 找CpG岛
- 找Motif
- 剪切位点识别
- 序列比对

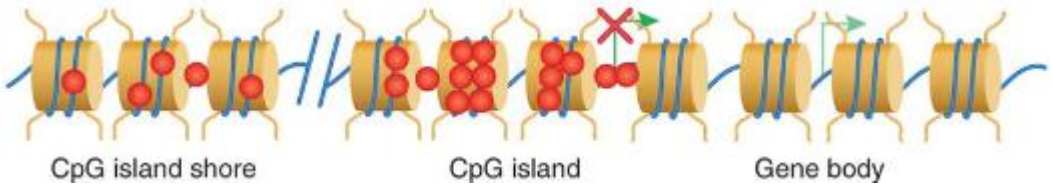


CpG岛分析

- 问题：给定一段DNA序列，判断该序列是否来自CpG岛。
- 模型：为每种碱基(A, G, C, T)建立一个状态，碱基s指向碱基t的边的权值为经统计得到的前一个碱基为s当前碱基为t的概率，称为**状态转移概率**，即

$$a_{st} = \mathrm{P}(\mathrm{x_i} = \mathrm{t} | \mathrm{x_{i-1}} = \mathrm{s})$$





CpG岛分析 – 模型建立

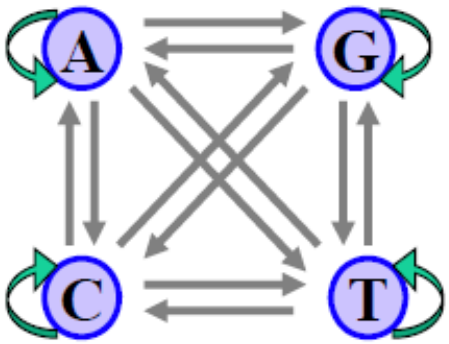
- 建模：给定一段DNA序列，观察到这段序列的概率为

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

- 若马尔可夫性成立，上述公式变为

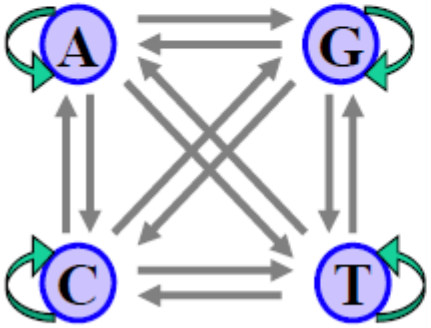
$$P(x) = P(x_L \mid x_{L-1}) P(x_{L-1} \mid x_{L-2}) \cdots P(x_2 \mid x_1) P(x_1)$$

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



CpG岛分析 – 模型训练

```
AGTAAAAAATAAATATGTTTAATTTGTGAACTGAT
TACCATCAGAATTGTACTGTTCTGTATCCCACCAG
CAATGTCTAGGAATGCCTGTTTCTCCACAAAGTGT
TTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAG
CCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGC
CTGGCCATGTGTATTTTTTTAAATTTCCACTGATG
ATTTTGCTGCATGGCCGGTGTTGAGAATGACTGCG
CAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATG
TAGTTTAAACGAGATTGCCAGCACCGGGTATCATT
CACCATTTTTCTTTTCGTTAACTTGCCGTCAGCCT
. . . . . .
```



Transition probabilities matrix P-
Based on known non-CpG Island sequences

Negative	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

Transition probabilities matrix P+
Based on known CpG Island sequences

Positive	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

CpG岛分析

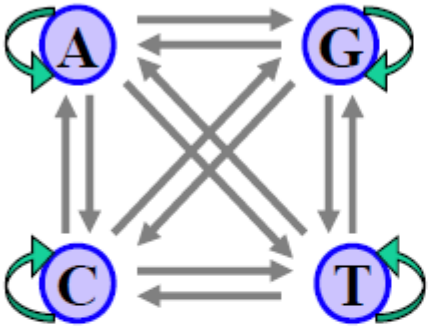
- Given a new sequence $X=(x_1,\ldots, x_L)$, we can compute the ratio

$$\mathrm{RATIO} = \frac{p(\boldsymbol{x} \mid +\mathrm{model})}{p(\boldsymbol{x} \mid -\mathrm{model})} = \frac{\prod_{i=0}^{L-1} p_+(x_{i+1} \mid x_i)}{\prod_{i=0}^{L-1} p_-(x_{i+1} \mid x_i)}$$

- RATIO>1 $\Rightarrow$ sequence from *CpG* island regions
- RATIO<1 $\Rightarrow$ sequence from non-*CpG* island regions

Positive+	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

Negative-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292



$$\text{RATIO} = \frac{p(\boldsymbol{x} \mid +\text{model})}{p(\boldsymbol{x} \mid -\text{model})} = \frac{\prod_{i=0}^{L-1} p_+(x_{i+1} \mid x_i)}{\prod_{i=0}^{L-1} p_-(x_{i+1} \mid x_i)}$$

New sequence: **x=TGCAGCG**

$$P(x|+model) = P_+(G|C) * P_+(C|G) * P_+(G|A) * P_+(A|C) * P_+(C|G) * P_+(G|T)$$
$$= 0.274 * 0.339 * 0.426 * 0.171 * 0.339 * 0.384$$
$$= 0.000880819444$$

$$P(x|-model) = P_-(G|C) * P_-(C|G) * P_-(G|A) * P_-(A|C) * P_-(C|G) * P_-(G|T)$$
$$= 0.078 * 0.246 * 0.285 * 0.322 * 0.246 * 0.292$$
$$= 0.00012648773$$

RATIO= P (x|+model)/ P (x|-model) = 0.0008808/ 0.0001265 = 6.96 > 1

=> Sequence x more likely comes from CpG island regions

更复杂的问题

- Given a (stretch of a) genomic sequence, where are the CpG islands?

  CGAAACTTTGCGCGGATTTCGCTTTGCTGTTCCTGCATGTAACCCT

- Given a (stretch of a) genomic sequence, where are the coding regions and where are noncoding regions?

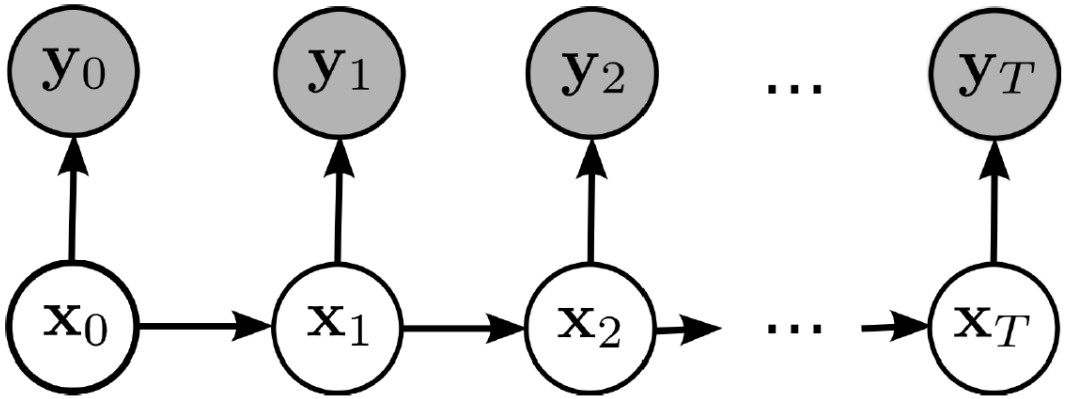
  ACCCTAACCCTAACCCTCGCGGTACCCTCAGCCCGAAAAAATCG

➔Partition problem: segment a given sequence into different parts based on its sequence features.

➔Need to guess the "state" of each position (e.g., CpG/non-CpG, coding/noncoding)

隐马尔可夫模型 (HMM)

- 在正常的马尔可夫模型中，状态对于观察者来说是直接可见的。这样状态的转换概率便是全部的参数。而在隐马尔可夫模型中,状态并不是直接可见的，但受状态影响的某些变量则是可见的。每一个状态在可能输出的符号上都有一概率分布。因此输出符号的序列能够透露出状态序列的一些信息。



隐马尔可夫模型 (HMM)

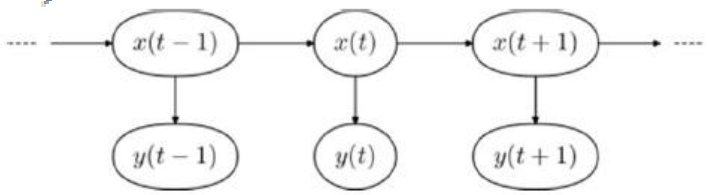
- 每个yi是我们观察到的符号，它是由其对应的隐状态xi以一定的概率生成的，这些隐状态对应观察者来说是不可见的，但他们是在背后真正发挥作用的，这些隐状态服从普通的马尔可夫模型

- **状态转移概率**

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

- 在每个隐状态中，每个符号出现的概率称为**生成概率**，可用如下条件概率表示

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



隐马尔可夫模型 (HMM)

在一个隐马尔可夫模型中，一个符号序列可以看成是如下步骤生成的：

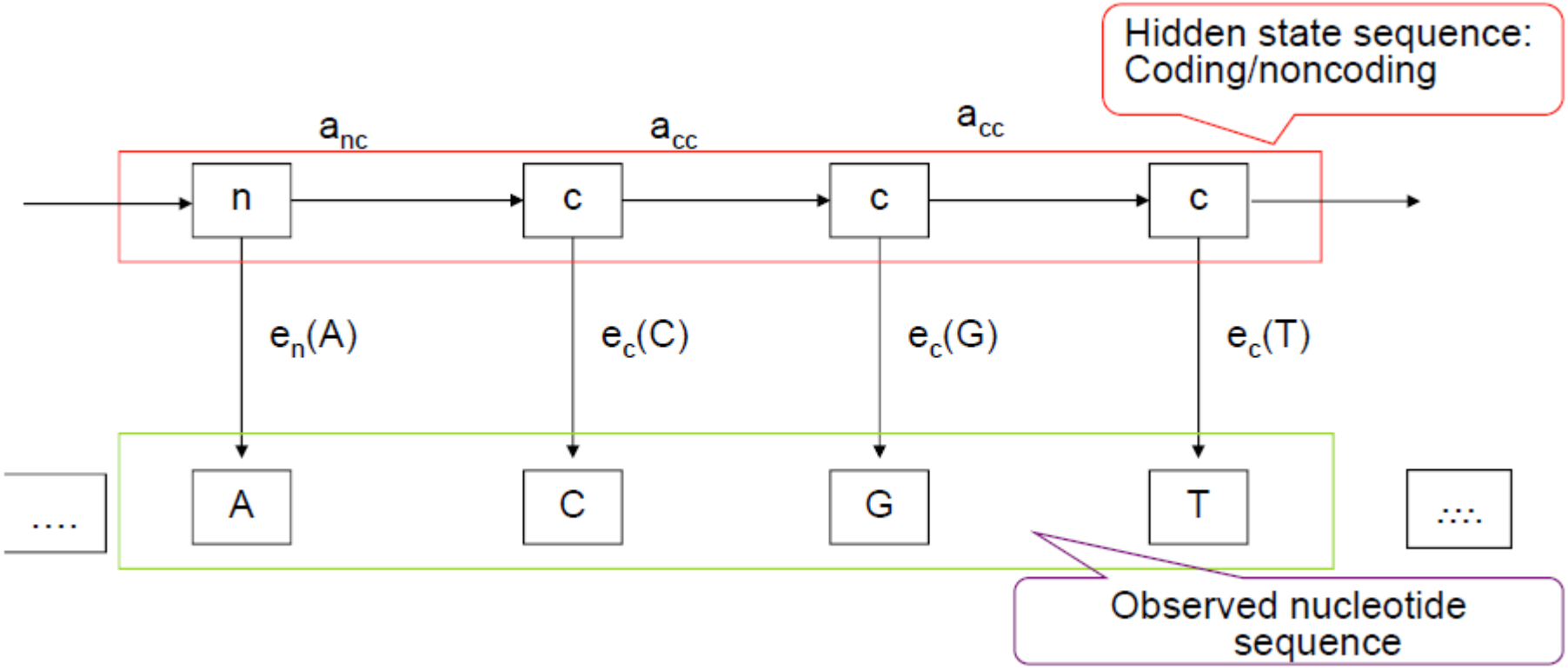
- 首先，选择一个状态作为起始，该概率称为起始概率
- 在一个状态π_i中，生成一个符号x_i，该生成概率为$e_{\pi_i}(x_i)$
- 然后转移到下一个状态π_j，该状态转移的概率为$a_{\pi_i\pi_j}$
- 重复上述第2步和第3步，直到生成整个序列

综上，观察到一个序列 $X=(x_1,\ldots,x_L)$ 的概率为：

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

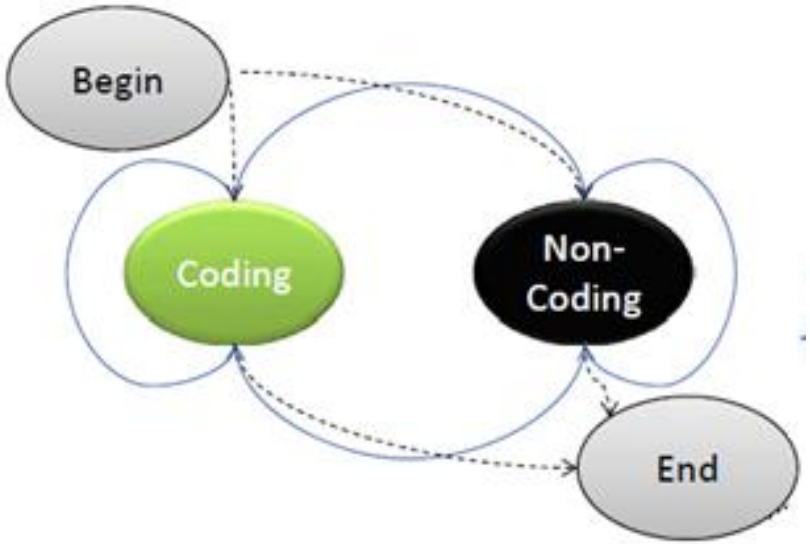
基于HMM的基因预测

给定一段序列，找出其中的基因，就是区分哪些段是Coding的(用C表示)，哪些是Noncoding(用N表示)。用HMM模型表示如下：



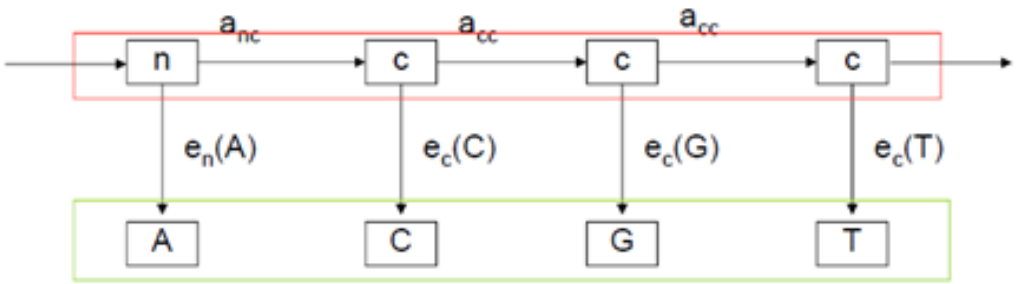
基于HMM的基因预测

HMM分为两部分，**状态转移(Transition)**和**生成(Emission)**。其中的状态转移模型如下：



模型训练

了上述模型，我们仅需同在之前的马尔可夫模型中一样，用已知的数据对该模型进行训练。即收集已知的基因序列和非基因序列，从中统计**状态转移概率矩阵**和在每个状态中的**生成概率矩阵**。



	n	c
n	0.8	0.2
c	0.4	0.6

(1) 状态转移概率矩阵

	A	C	G	T
n	0.2	0.3	0.3	0.2
c	0.4	0.2	0.2	0.2

(2)生成概率矩阵

基因预测

- 给定一段序列 CGAAAAAATCG ， 设对于Coding和Noncoding的起始概率分别为0.8和0.2，使用上述模型对编码区和非编码区进行预测。
- 目标是使P(X, π)最大化。这个问题可以用动态规划(Dynamic Programing)来快速地解决

动态规划

Initialization (i=0)

$$v_0(0) = 1, v_k(0) = 0 \text{ for all } k > 0$$

Recursion:

$$v_{coding}(i+1) = e_{coding}(x_{i+1}) \max_{k \in (coding, noncoding)} (v_k(i) a_{k->coding})$$

$$v_{noncoding}(i+1) = e_{noncoding}(x_{i+1}) \max_{k \in (coding, noncoding)} (v_k(i) a_{k->noncoding})$$

Termination:

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

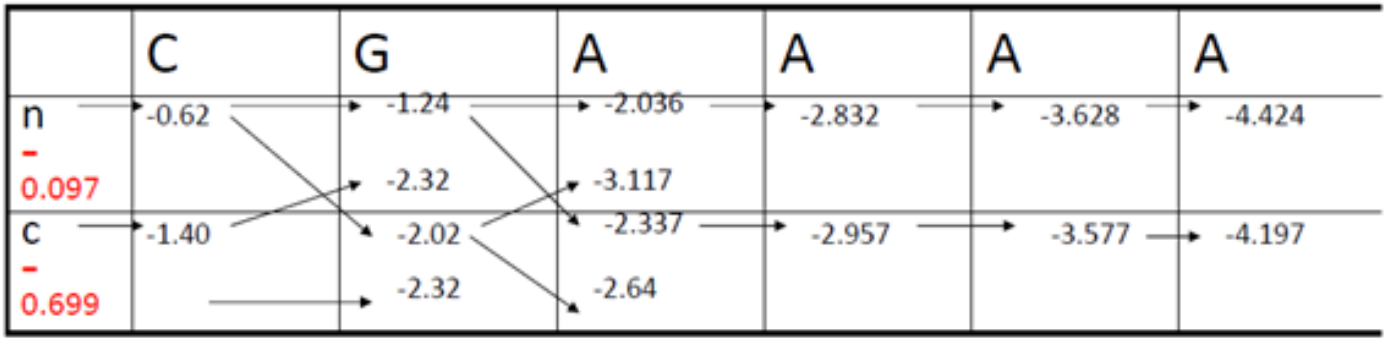
$$\pi_L^* = \arg\max_k (v_k(L) a_{k0})$$

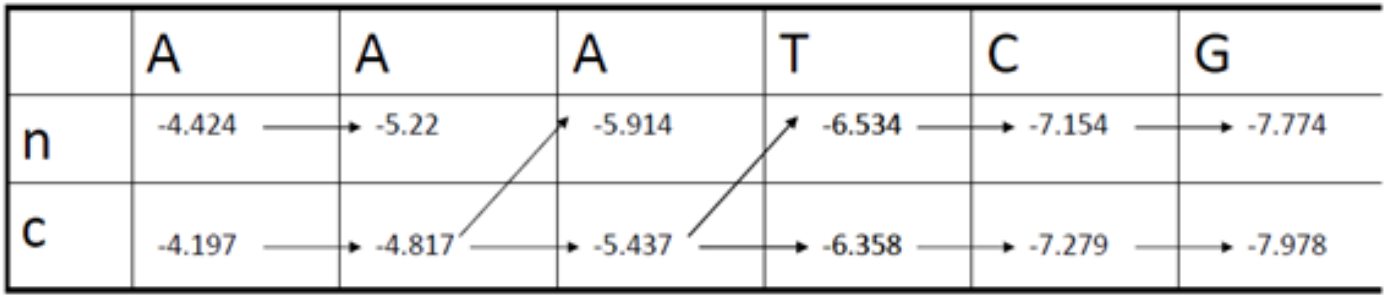
动态规划

- 计算过程中将所有的数都取log，从而使乘法运算变为加法运算。计算过程如下

	n	c
n	-0.097	-0.699
c	-0.398	-0.222

	A	C	G	T
n	-0.699	-0.523	-0.523	-0.699
c	-0.398	-0.699	-0.699	-0.699





回溯

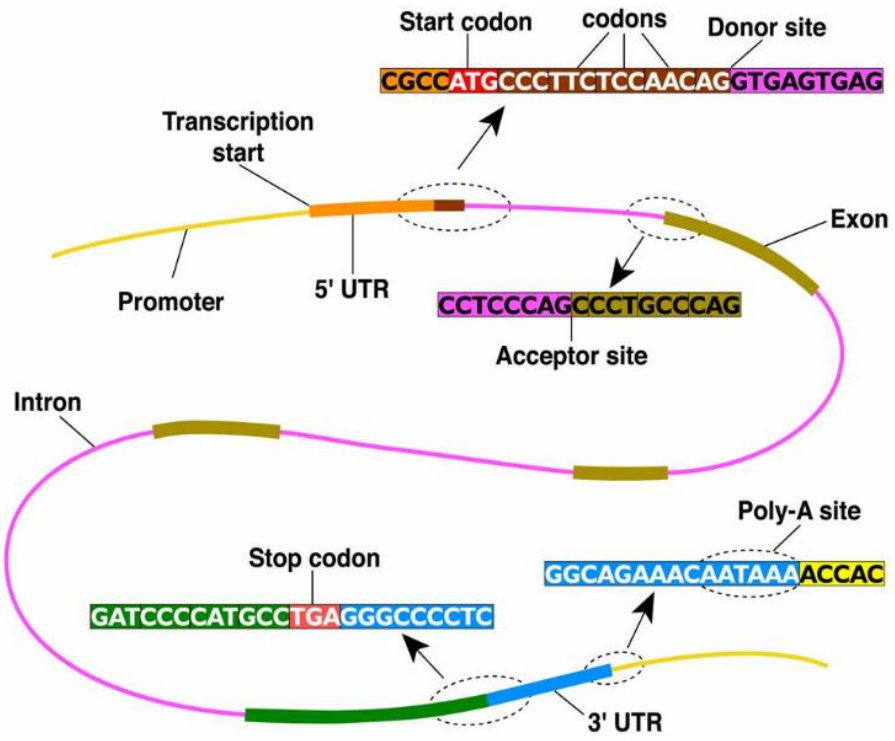
- 通过对上述结果回溯可以得到使P(X, π)最大时coding和noncoding的分布情况：

CGAAAAAATCG

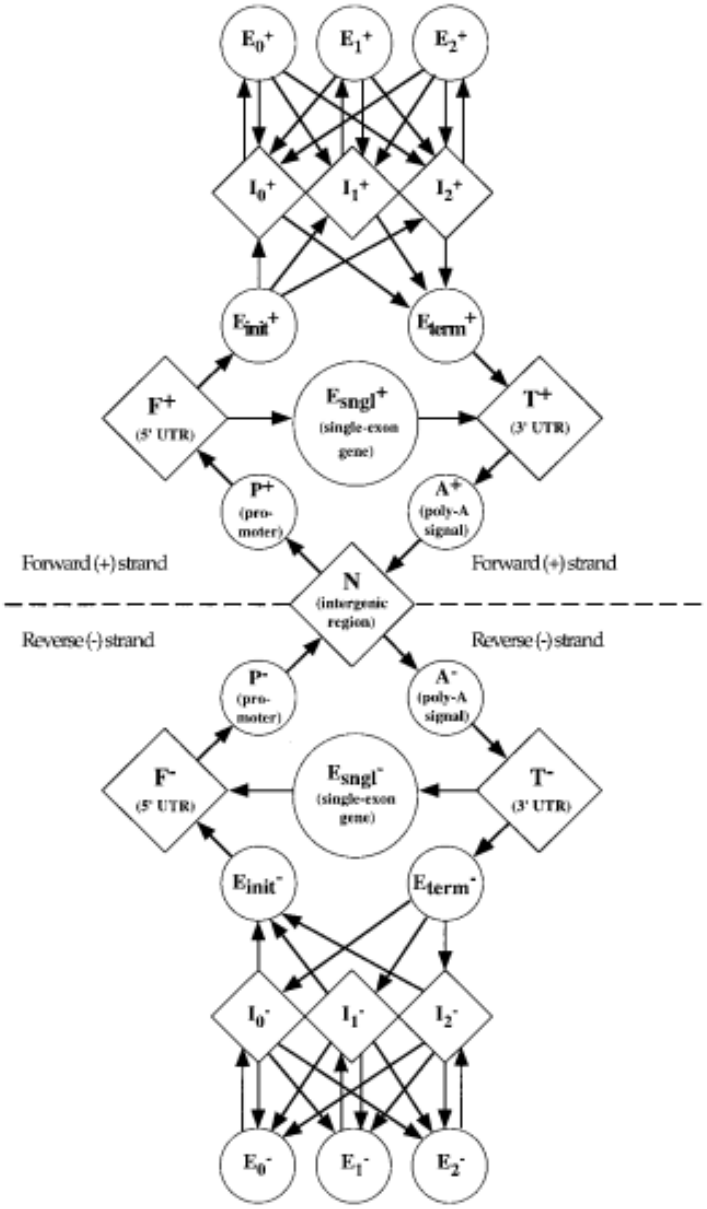
NNCCCCCCCNNN

基因预测

在实际的基因预测中，不仅要考虑编码、非编码，还要考虑内含子、外显子、启动子、5’UTR、3’UTR等。但整个模型和计算过程完全同上，只是状态转移矩阵中多了这些行。



基因预测 - GeneScan



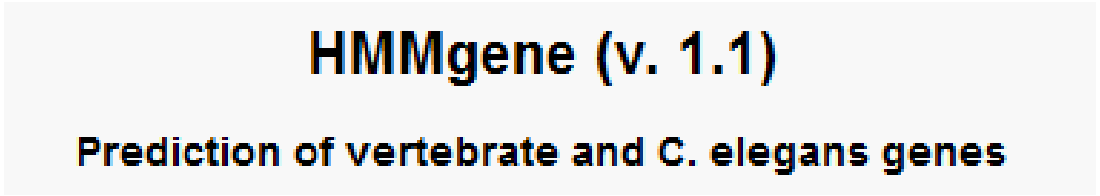
- N – intergenic region
- P – promoter
- F – 5’ UTR
- Esngl – single-exon gene
- Einit – initial exon
- Ek (0=<k<=2) – phase k internal exon
- Eterm – terminal exon
- T – 3’ UTR
- A – polyadenylation signal
- Ik (0=<k<=2) – phase k intron

演示

以HBA1基因(位于chr16:226,679-227,520)及两侧各延伸1k所得到的序列为例进行分析

- GeneScan

- HMMGene



参考文献

- Gao Ge, Lectures in MIB(Methods in Bioinformatics), 2012
- http://mib.cbi.pku.edu.cn
- http://en.wikipedia.org/wiki/Markov_model
- http://en.wikipedia.org/wiki/Hidden_Markov_model
- Burge, C. and S. Karlin, Prediction of complete gene structures in human genomic DNA. J Mol Biol, 1997. 268(1): p. 78-94.

致谢

- 感谢罗老师在ABC课上的指导
- 感谢我们小组成员

姬亚朋
王　嵛
梁　鹏

谢谢！