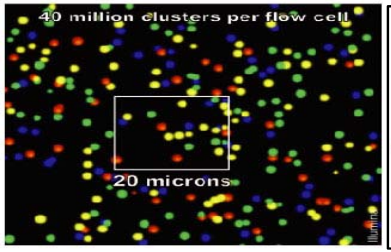
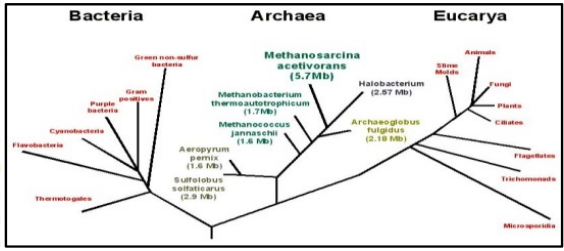
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



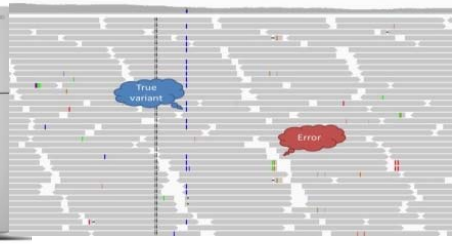


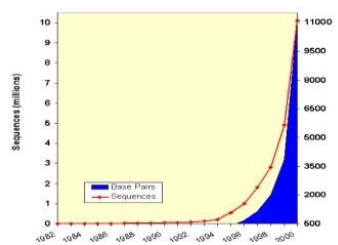
Unit 3:
Predict with Hidden Markov Model

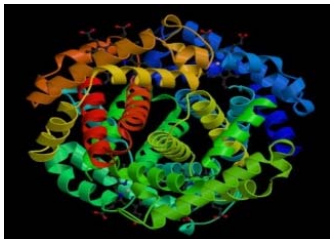
北京大学生物信息学中心 高歌

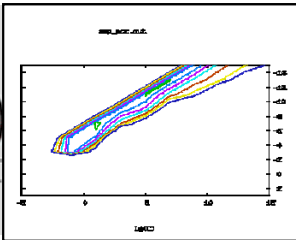
Ge Gao, Ph.D.

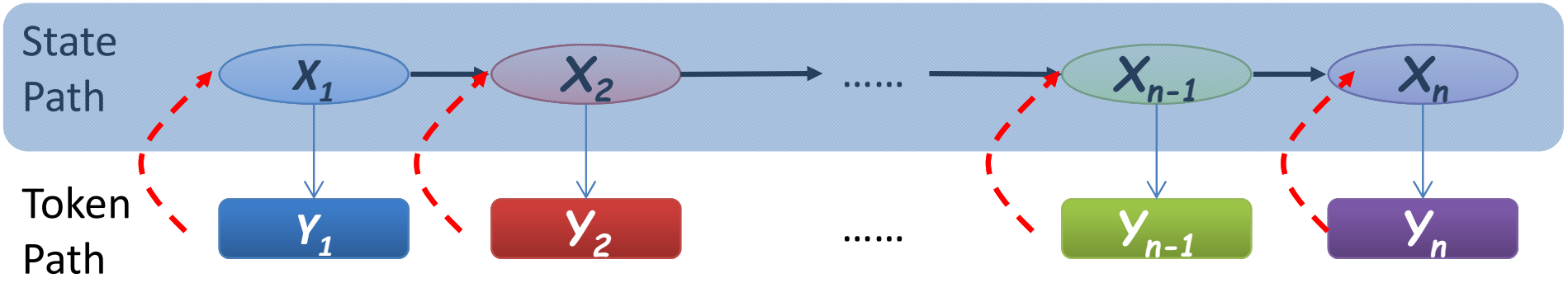
Center for Bioinformatics, Peking University









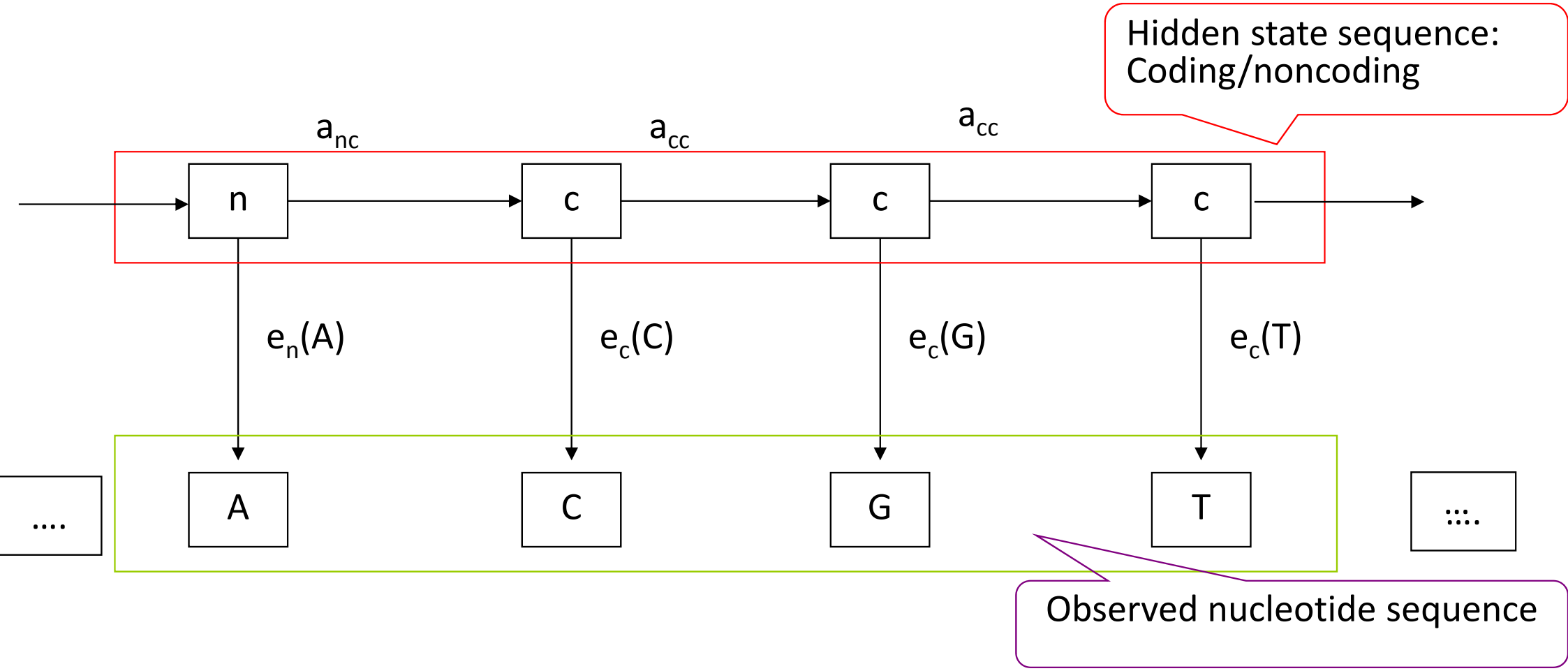


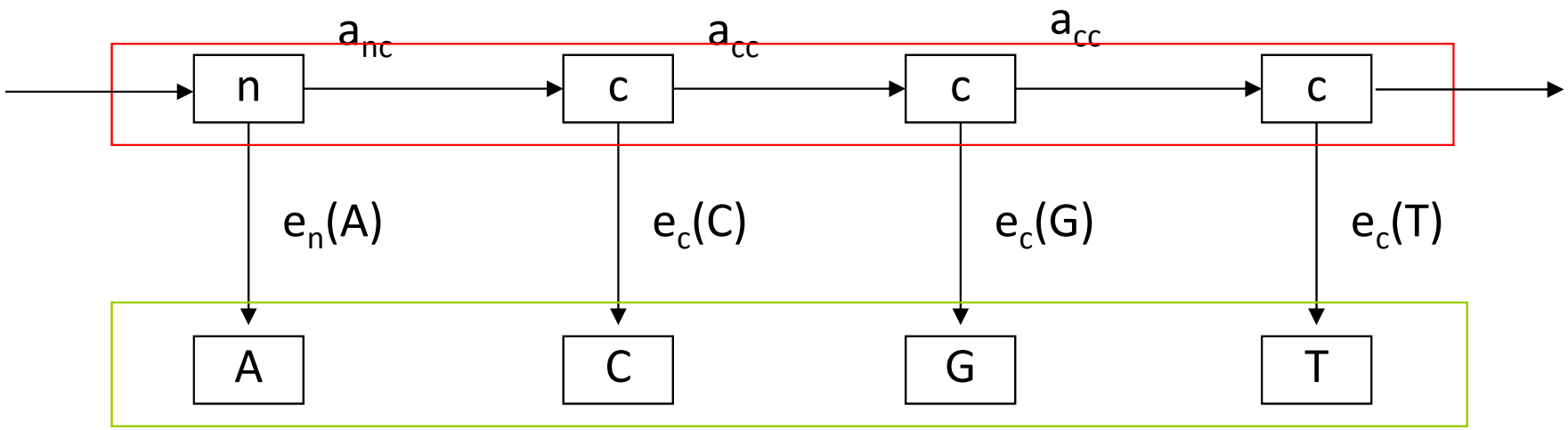
Hidden Markov Model: as a predictor

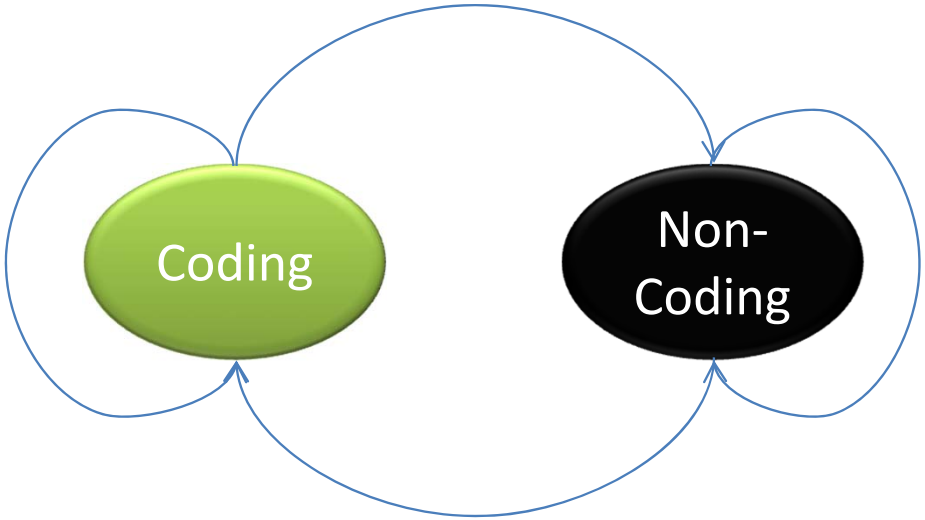
The Most Simple Gene Predictor (MSGP)

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

ACCCTAACCCTAACCCTCGCGGTACCCTCAGCCCGAAAAAATCG



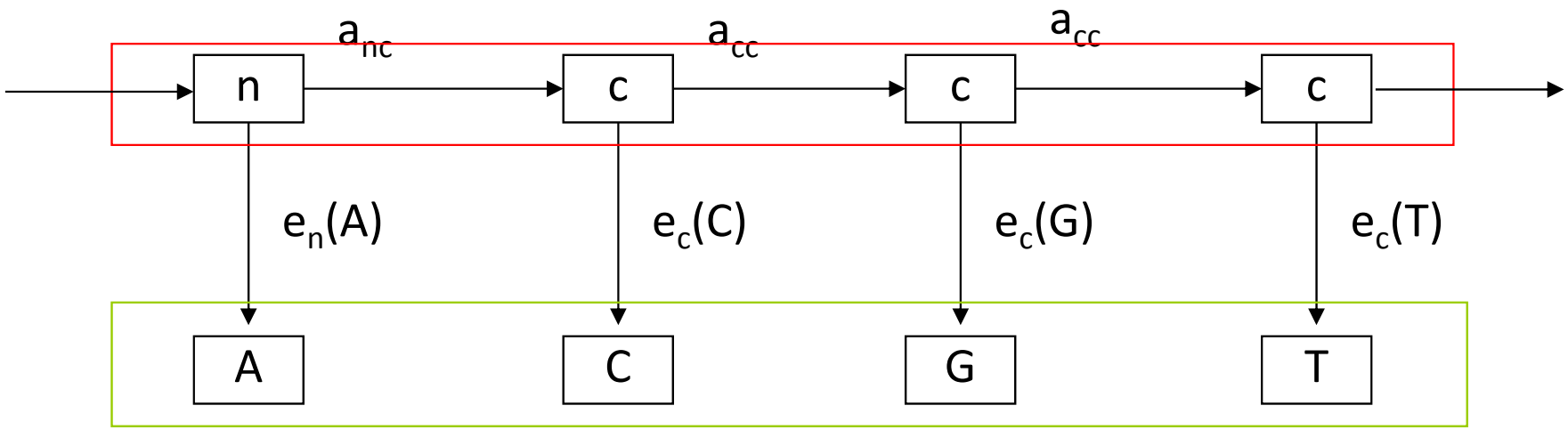




Transition Probability

	n	c
n		
c		

$$a_{kl} = P(x_t = S_l \mid x_{t-1} = S_k)$$



Emission Probability

$$e_k(b) = P(y_i = b \mid x_i = S_k)$$

	A	C	G	T
Coding				

	A	C	G	T
Non-coding				

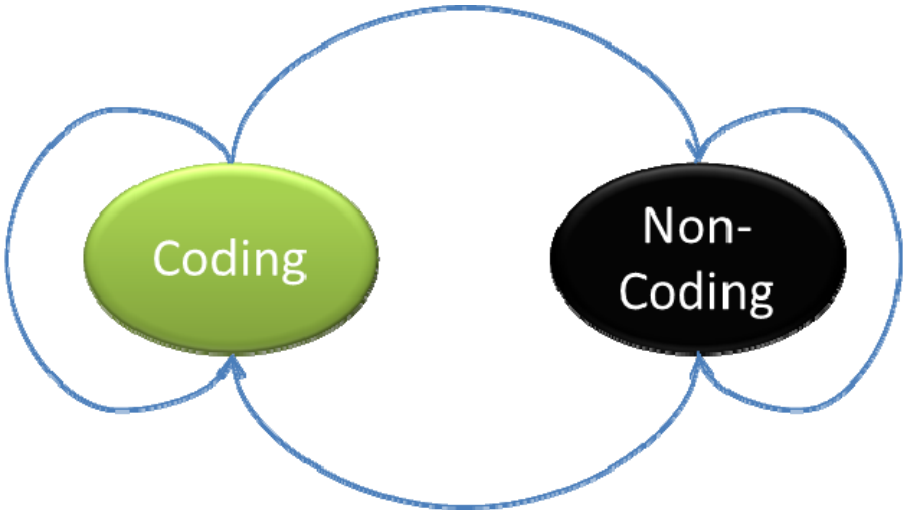
Training the model

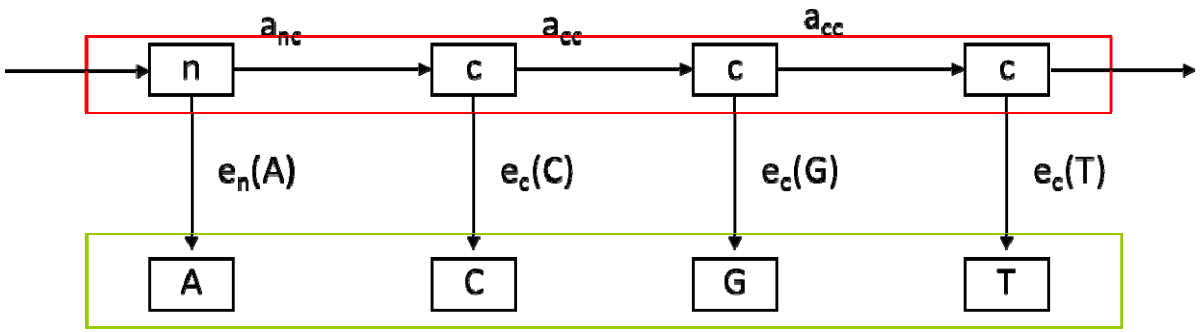
- What we need to train?
  - Transition Probabilities between states
  - Emission Probabilities for each state
- Estimate Probabilities from known "Training set"
  - An annotated genomic region, with coding/noncoding sequences labeled.

Token: ACGCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGATGT……
State: CCCCCCCCCCNNNNNCCCCCCCCCCNNNNNNNNNCCCCCCCCCCCCNNN……

$$\hat{a}_{kl} = \frac{a_{kl}}{\sum_{l'} a_{kl'}}$$

$$\hat{e}_k(b) = \frac{e_k(b)}{\sum_{b'} e_k(b')}$$



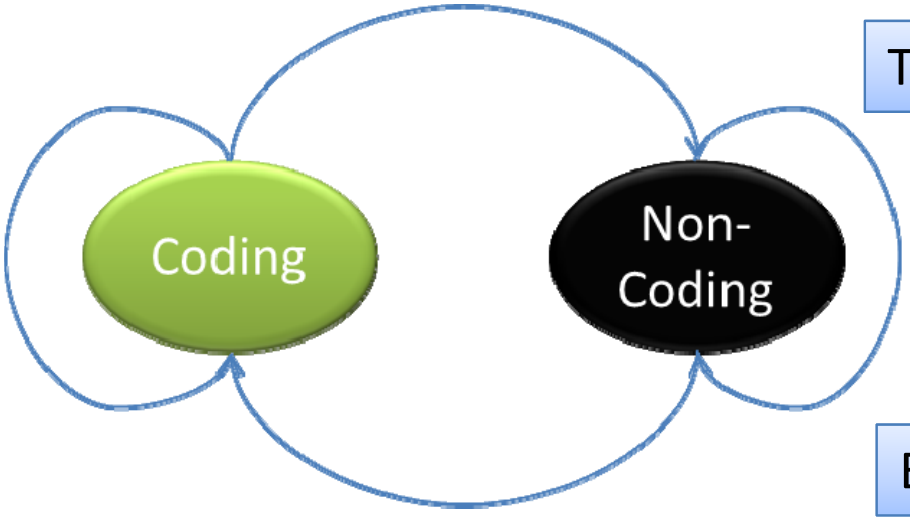


Transition Probability

	n	c
n	0.8	0.2
c	0.4	0.6

Emission Probability

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



Transition Probability

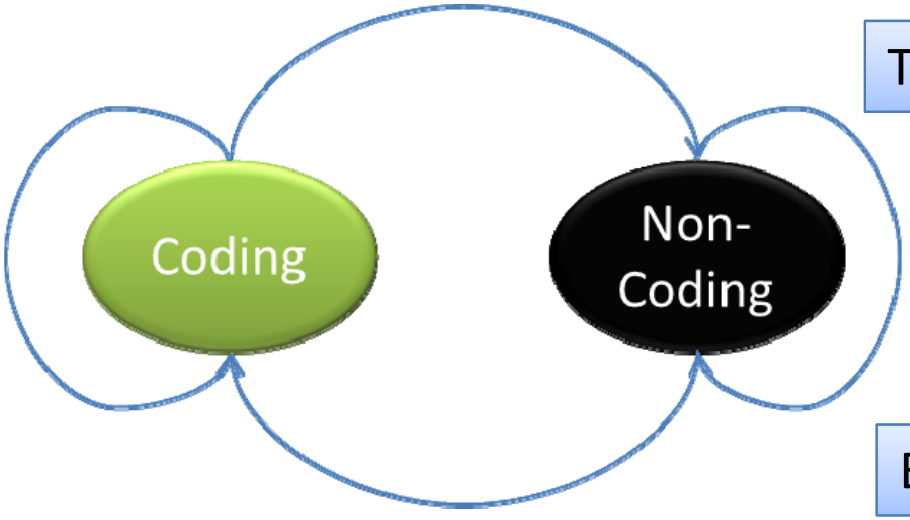
	n	c
n	0.8	0.2
c	0.4	0.6

Emission Probability

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

Given a sequence $X=X_1X_2X_3...X_n$, let $S=S_1S_2S_3...S_n$ represent its hidden states (i.e. coding, non-coding annotation), we need the the best S:

$$S^* = \arg\max_S P(S \mid X)$$



Transition Probability

	n	c
n	0.8	0.2
c	0.4	0.6

Emission Probability

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

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

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

$$P(X, S) = \max(P_{coding}(n), P_{noncoding}(n))$$

Logarithmic transformation: Ease calculation

$$\text{Log}\,(a*b) = \text{Log}(a) + \text{Log}\,(b)$$

	n	c
n	0.8	0.2
c	0.4	0.6

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

Log10

	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

Testing Sequence:

CGAAAAAATCG

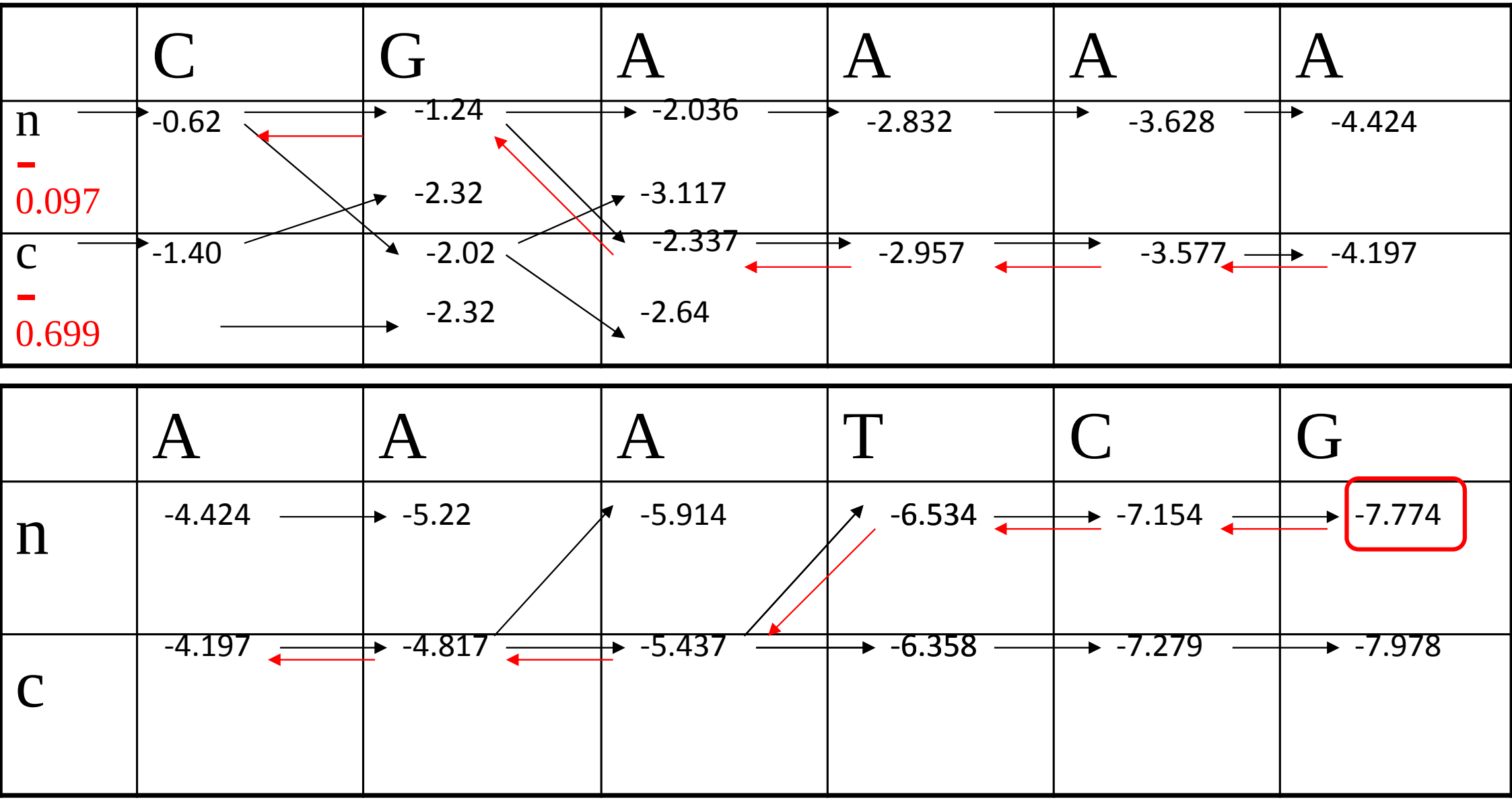
$$P_l(i+1) = e_l(x_{i+1}) \max_k (P_k(i) a_{kl})$$

	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

	C	G	A	A	A	A
n -0.097	-0.62	-1.24 -2.32	-2.036 -3.117	-2.832	-3.628	-4.424
c -0.699	-1.40	-2.02 -2.32	-2.337 -2.64	-2.957	-3.577	-4.197

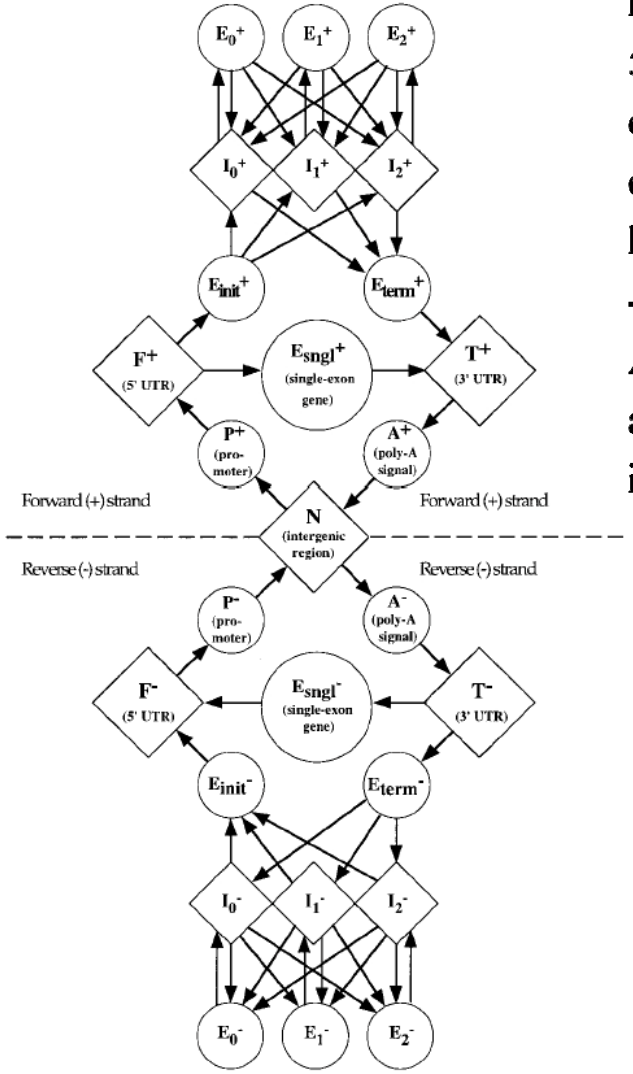
	A	A	A	T	C	G
n	-4.424	-5.22	-5.914	-6.534	-7.154	-7.774
c	-4.197	-4.817	-5.437	-6.358	-7.279	-7.978



The Most Simple Gene Predictor (MSGP)

CGAAAAAATCG

NNCCCCCCCNNN

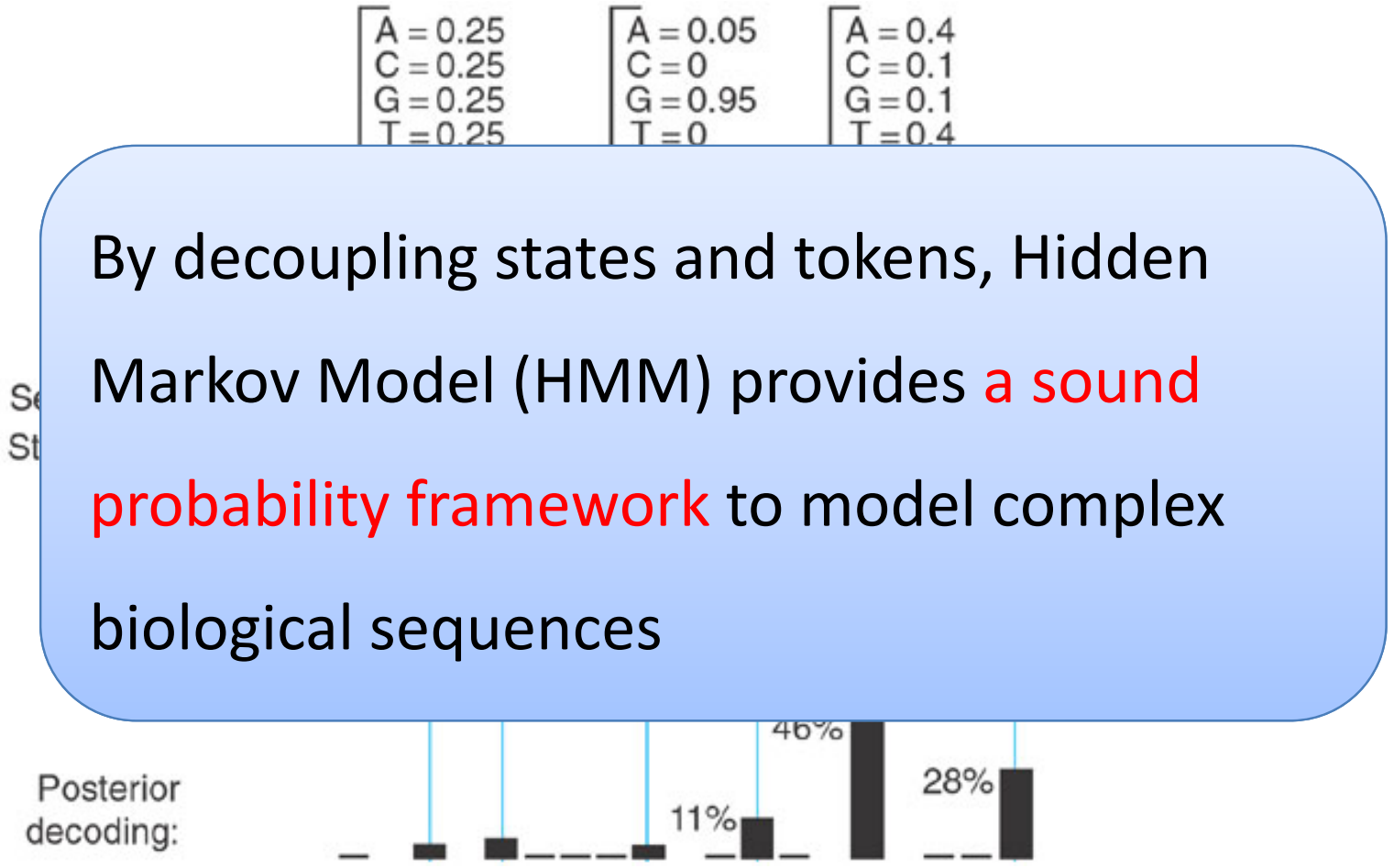


N, intergenic region; P,promotor; F, 5'UTR; E_{sngl} , single-exon gene; E_{init} , initial exon; $E_k(0 \le k \le 2)$ phase k internal exon; E_{term} ,ter -minal exon; T, 3'UTR; A,polyadenylation signal; and, $I_k(0 \le k \le 2)$, phase k intron.) strand.

GenScan:

- Chris Burge (1996): A 27-state semi-HMM
- A simpler model: 19-state
- A model taking UTR introns into account: 35-state

5' splice site recognition



By decoupling states and tokens, Hidden Markov Model (HMM) provides a sound probability framework to model complex biological sequences

(*Nature Biotechnology* **22**, 1315 - 1316 (2004))

生物信息学：导论与方法
Bioinformatics: Introduction and Methods