7.91 / 7.36 / BE.490

Lecture #4

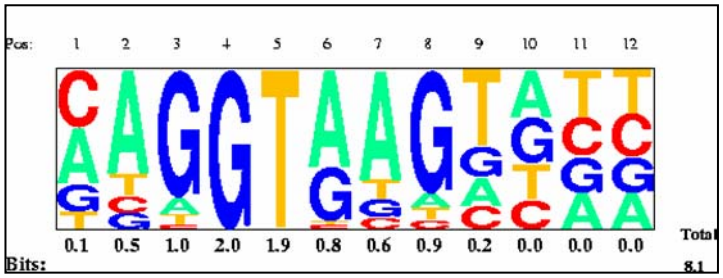
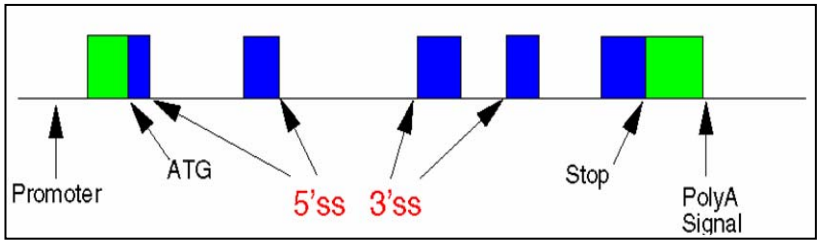
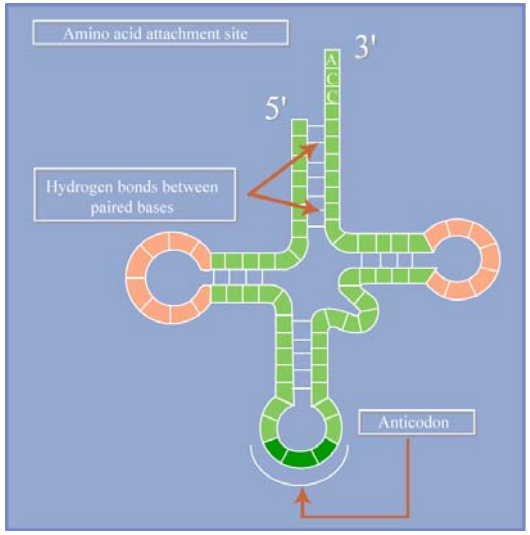
Mar. 4, 2004

Markov & Hidden Markov Models

for DNA Sequence Analysis

Chris Burge

Organization of Topics

Lecture	Object	Model	Dependence Structure
3/2		Weight Matrix Model	Independence
3/4		Hidden Markov Model	Local Dependence
3/9		Energy Model, Covariation Model	Non-local Dependence

Markov & Hidden Markov Models for DNA

- Markov Models for splice sites
- Hidden Markov Models
    - looking under the hood
- The Viterbi Algorithm
- Real World HMMs

See Ch. 4 of Mount

Review of DNA Motif Modeling & Discovery

- WMMs for splice sites
- Information Content of a Motif
- The Motif Finding/Discovery Problem
- The Gibbs Sampler

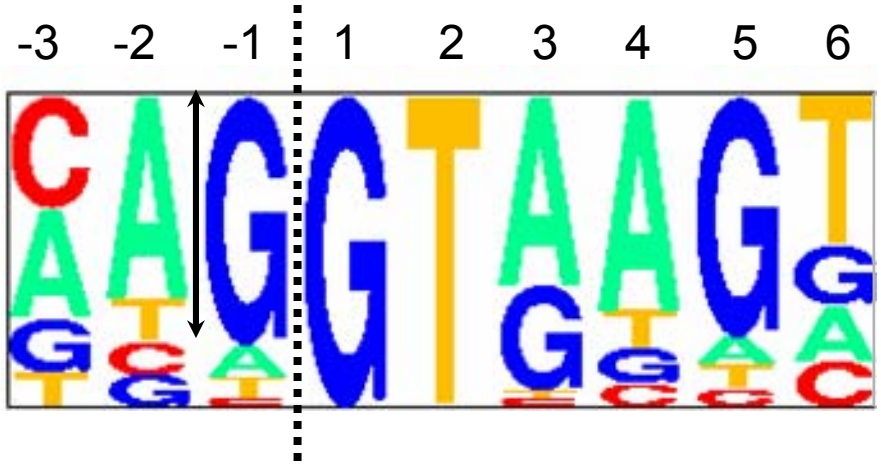
The Gibbs Sampling Algorithm Multimedia Experience

- Motif Modeling - Beyond Weight Matrices

See Ch. 4 of Mount

Information Content of a DNA/RNA Motif

f_k = freq. of nt *k* at position



Shannon Entropy $H(\vec{f}) = -\sum_k f_k \log_2(f_k)$ (bits)

Information/position

$$I(\vec{f}) = 2 - H(\vec{f}) = 2 + \sum_k f_k \log_2(f_k) = \sum_k f_k \log_2\left(\frac{f_k}{1/4}\right) \text{ (bits)}$$

Motif containing m bits of info. will occur approximately once per 2^m bases of random sequence

Variables Affecting Motif Finding

```
gcggaagagggcactagcccatgtgagagggcaaggacca
atctttctcttaaaaataacataattcagggccaggatgt
gtcacgagctttatcctacagatgatgaatgcaaatcagc
taaaagataatatcgaccctagcgtggcgggcaaggtgct
gtagattcgggtaccgttcataaaagtacgggaatttcgg
tatacttttaggtcgttatgttaggcgagggcaaaagtca
ctctgccgattcggcgagtgatcgaagagggcaatgcctc
aggatggggaaaatatgagaccaggggagggccacactgc
acacgtctagggctgtgaaatctctgccgggctaacagac
gtgtcgatgttgagaacgtaggcgccgaggccaacgctga
atgcaccgccattagtccggttccaagagggcaactttgt
ctgcgggcggcccagtgcgcaacgcacagggcaaggttta
tgtgttgggcggttctgaccacatgcgagggcaacctccc
gtcgcctaccctggcaattgtaaaacgacggcaatgttcg
cgtattaatgataaagagggggtaggaggtcaactcttc
aatgcttataacataggagtagagtagtgggtaaactacg
tctgaaccttctttatgcgaagacgcgagggcaatcggga
tgcatgtctgacaacttgtccaggaggaggtcaacgactc
cgtgtcatagaattccatccgccacgcggggtaatttgga
tcccgtcaaagtgccaacttgtgccgggggctagcagct
acagcccgggaatatagacgcgtttggagtgcaaacatac
acgggaagatacgagttcgatttcaagagttcaaaacgtg
cccgataggactaataaggacgaaacgagggcgatcaatg
ttagtacaaacccgctcacccgaaaggagggcaaatacct
agcaaggttcagatatacagccaggggagacctataactc
gtccacgtgcgtatgtactaattgtggagagcaaatcatt
…
```

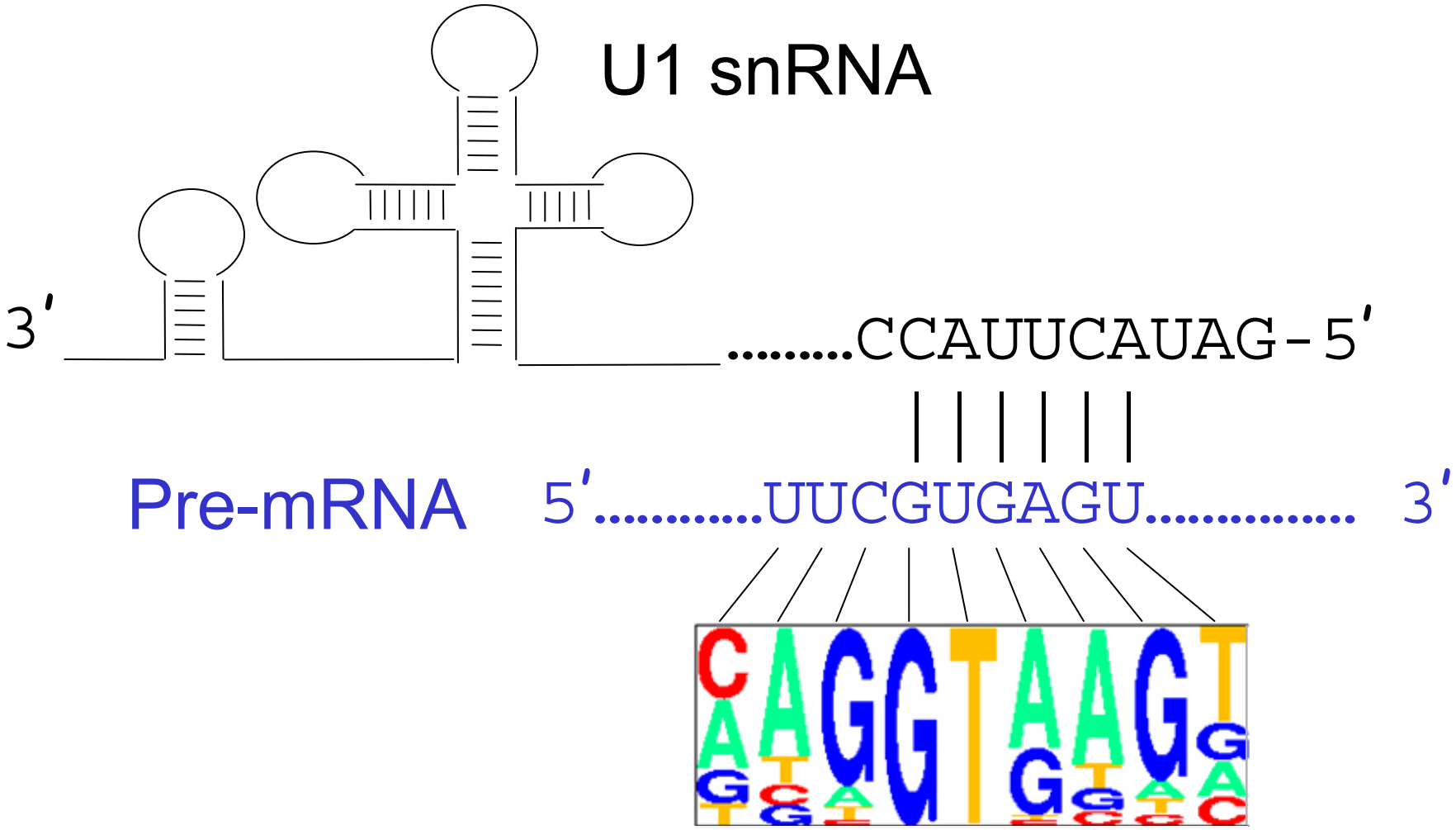
L = avg. sequence length

N = no. of sequences

I = info. content of motif

W = motif width

How is the 5′ss recognized?

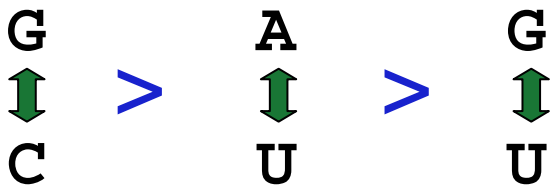


RNA Energetics I

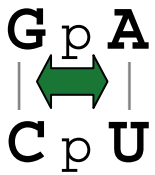
```
   ...CCAUUCAUAG-5'
      ||||||
5'...CGUGAGU.....  3'
```

Free energy of helix formation derives from:

- base pairing:



- base stacking:



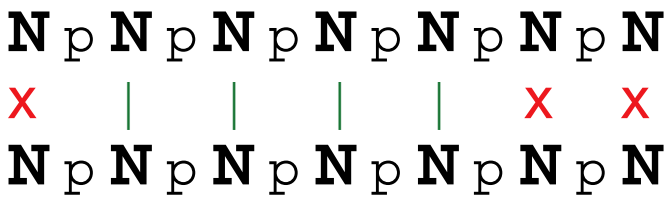
Doug Turner's Energy Rules:

```
                5' --> 3'
                   UX
                   AY
                3' <-- 5'

                      X
Y       A         C         G         U
A       .         .         .     -1.30
C       .         .     -2.40         .
G       .     -2.10         .     -1.00
T   -0.90         .     -1.30         .
```

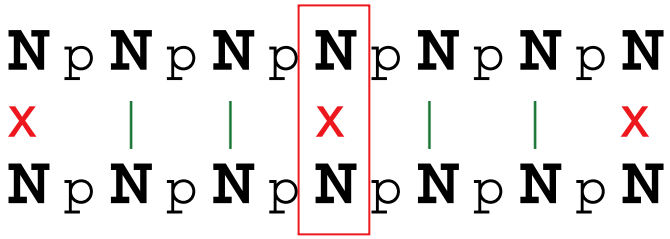
RNA Energetics II

A)



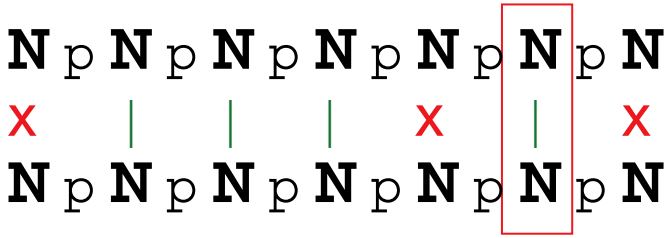
Lots of consecutive base pairs - good

B)



Internal loop - bad

C)



Terminal base pair not stable - bad

Generally A will be more stable than B or C

Conditional Frequencies in 5′ss Sequences

5′ss which have G at +5

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

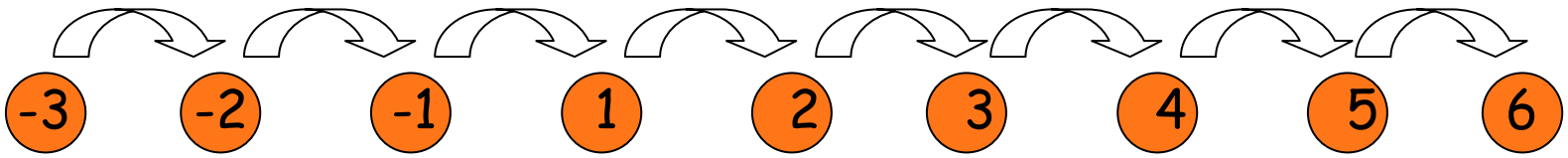
5′ss which lack G at +5

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

Data from Burge, 1998 “Computational Methods in Molecular Biology”

What kind of model could incorporate interactions between positions?

A Markov Model



Terminology

Random Variable (RV):

A quantity which may assume any one of a set of values, each with a definite probability of occurrence

Examples: X = the outcome of rolling a die

$P(X=1) = \frac{1}{6}$ $P(X=2) = \frac{1}{6}$ … $P(X=6) = \frac{1}{6}$

The craps process: $X_1, X_2, X_3, \ldots$ successive dice rolls

Stochastic Process:

a random process

or a sequence of Random Variables

What is a *Markov* Model (aka *Markov* Chain)?

Classical Definition

A discrete stochastic process $X_1, X_2, X_3, \ldots$
which has the Markov property:

$$P(X_{n+1} = j \mid X_1=x_1, X_2=x_2, \ldots X_n=x_n) = P(X_{n+1} = j \mid X_n=x_n)$$

(for all x_i, all j, all n)

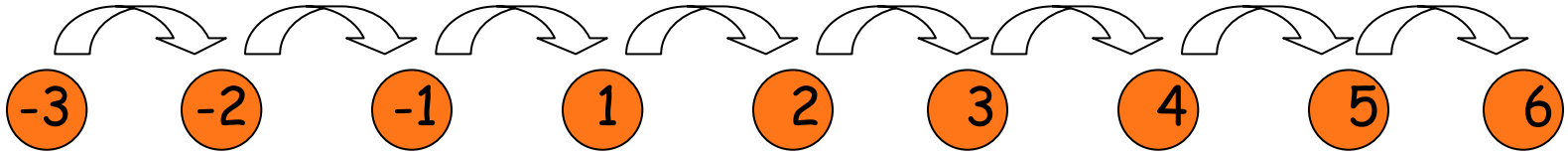
In words:

A random process which has the property that the future (next state) is conditionally independent of the past given the present (current state)

Markov - a Russian mathematician, ca. 1922

Inhomogeneous 1st-Order Markov Model

$$P_{-2}(A|C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$

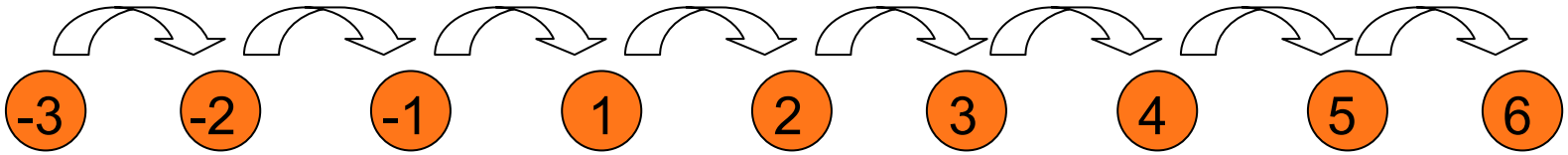


$$S = S_1\, S_2\, S_3\, S_4\, S_5\, S_6\, S_7\, S_8\, S_9$$

$$R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2|S_1)P_{-1}(S_3|S_2)\cdots P_{6}(S_9|S_8)}{P_{bg}(S_1)P_{bg}(S_2|S_1)P_{bg}(S_3|S_2)\cdots P_{bg}(S_9|S_8)}$$

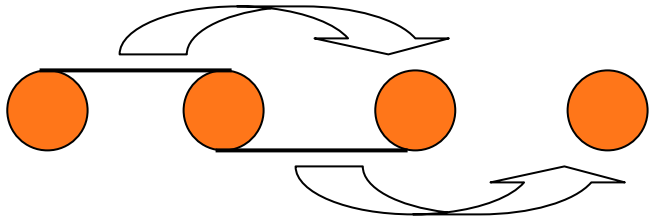
Estimating Parameters for a Markov Model

$$P_{-2}(A \mid C) = \frac{N_{CA}^{(-3,-2)}}{N_{C}^{(-3)}}$$

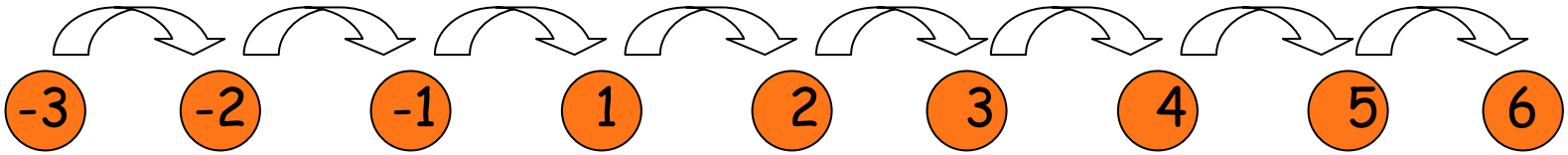


What about longer-range dependence?

- k-order Markov model



~4^{k+1} parameters / position for Markov model of order k



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

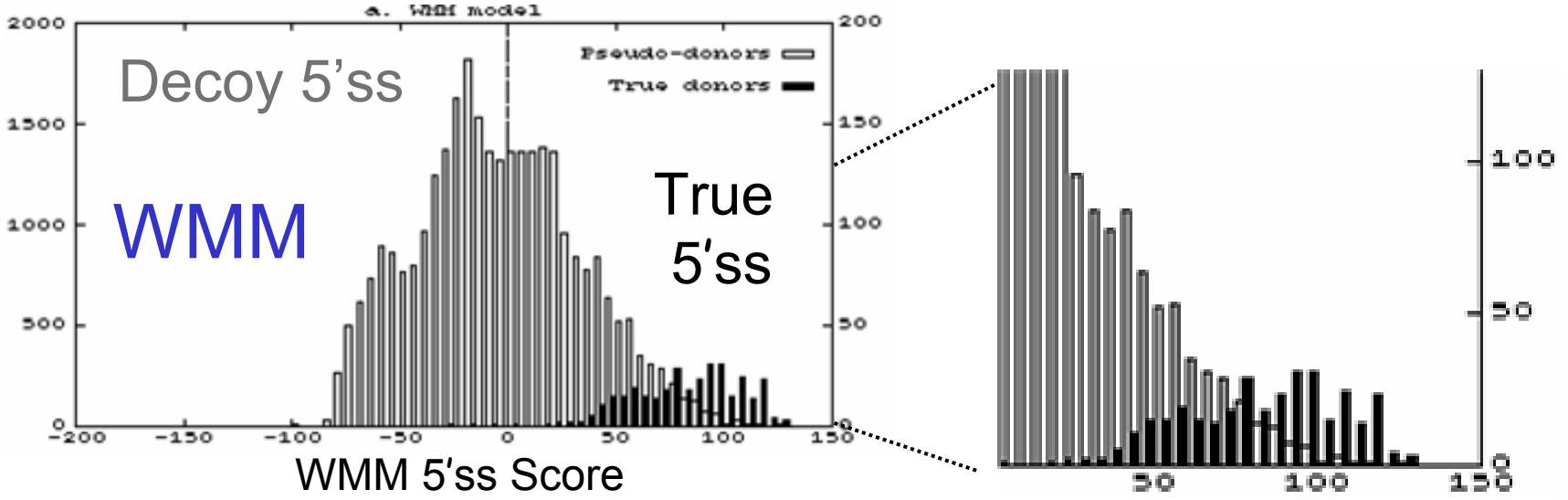
Inhomogeneous

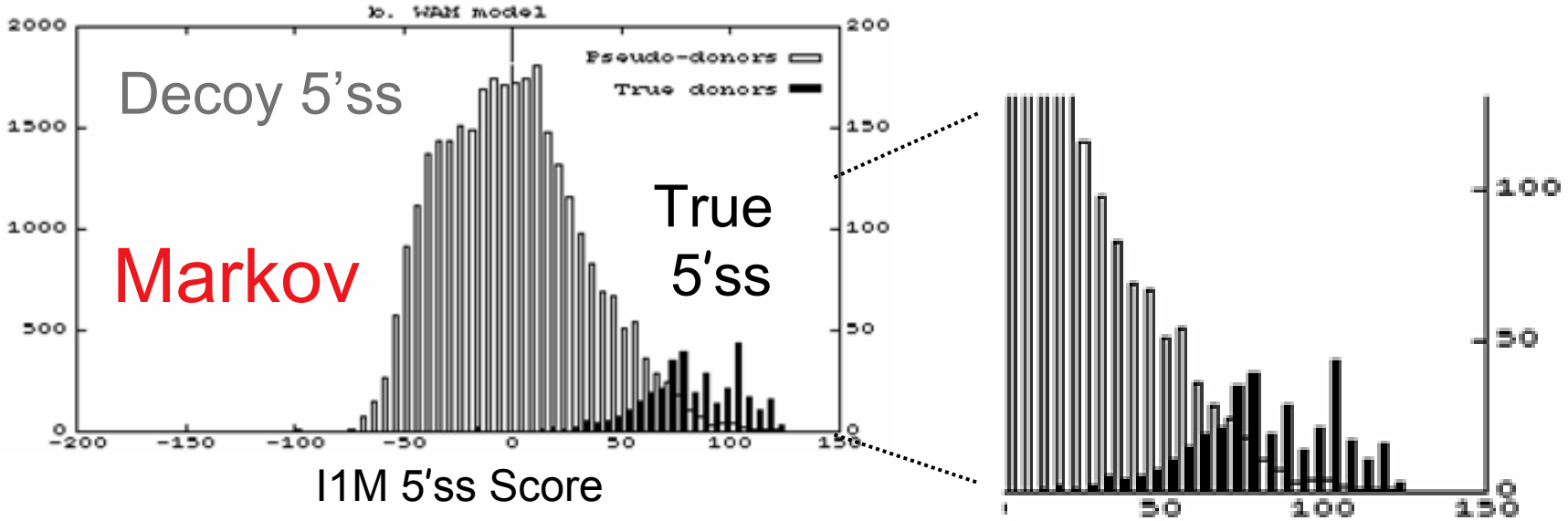
$$R = \frac{P(S|+)}{P(S|-)} = \frac{P_{-3}(S_1)P_{-2}(S_2|S_1)P_{-1}(S_3|S_2)\bullet\bullet\bullet P_6(S_9|S_8)}{P_{bg}(S_1)P_{bg}(S_2|S_1)P_{bg}(S_3|S_2)\bullet\bullet\bullet P_{bg}(S_9|S_8)}$$

Homogeneous

$$s = \log_2 R$$

WMM vs 1st-order Markov Models of Human 5′ss

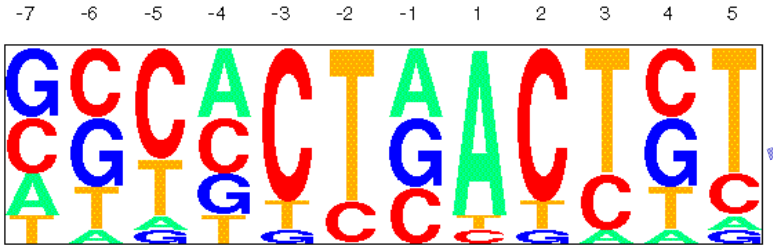




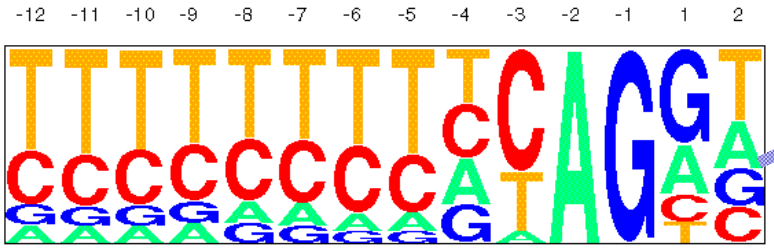
Splicing Model I

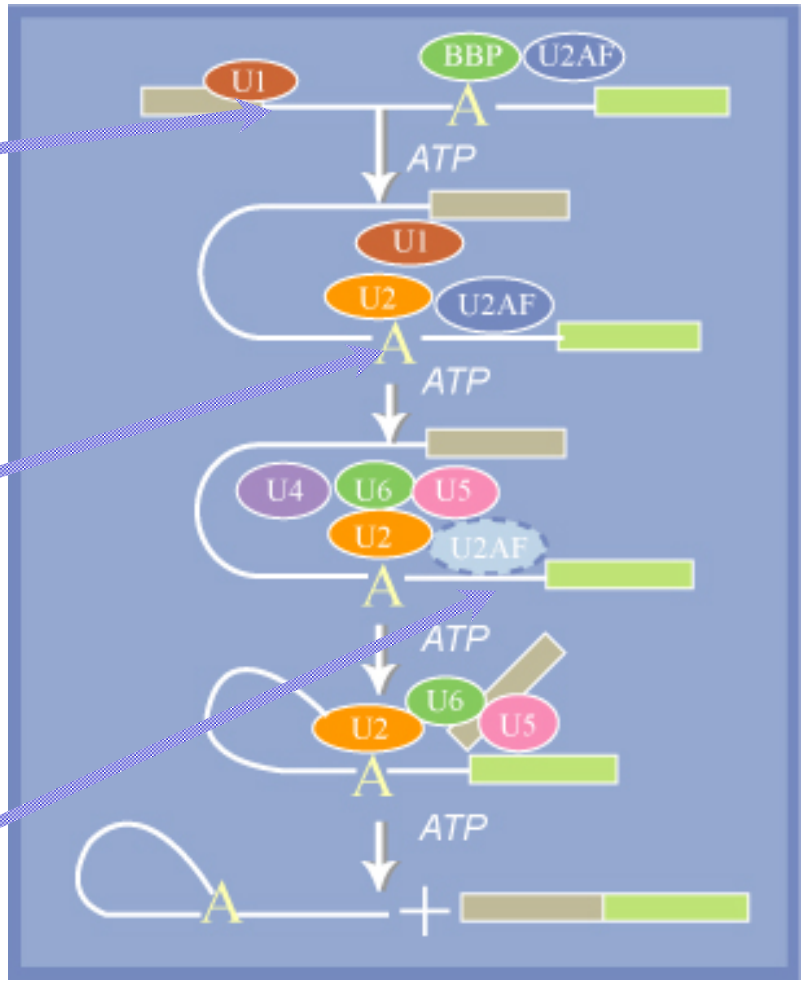
5′ splice site

branch site

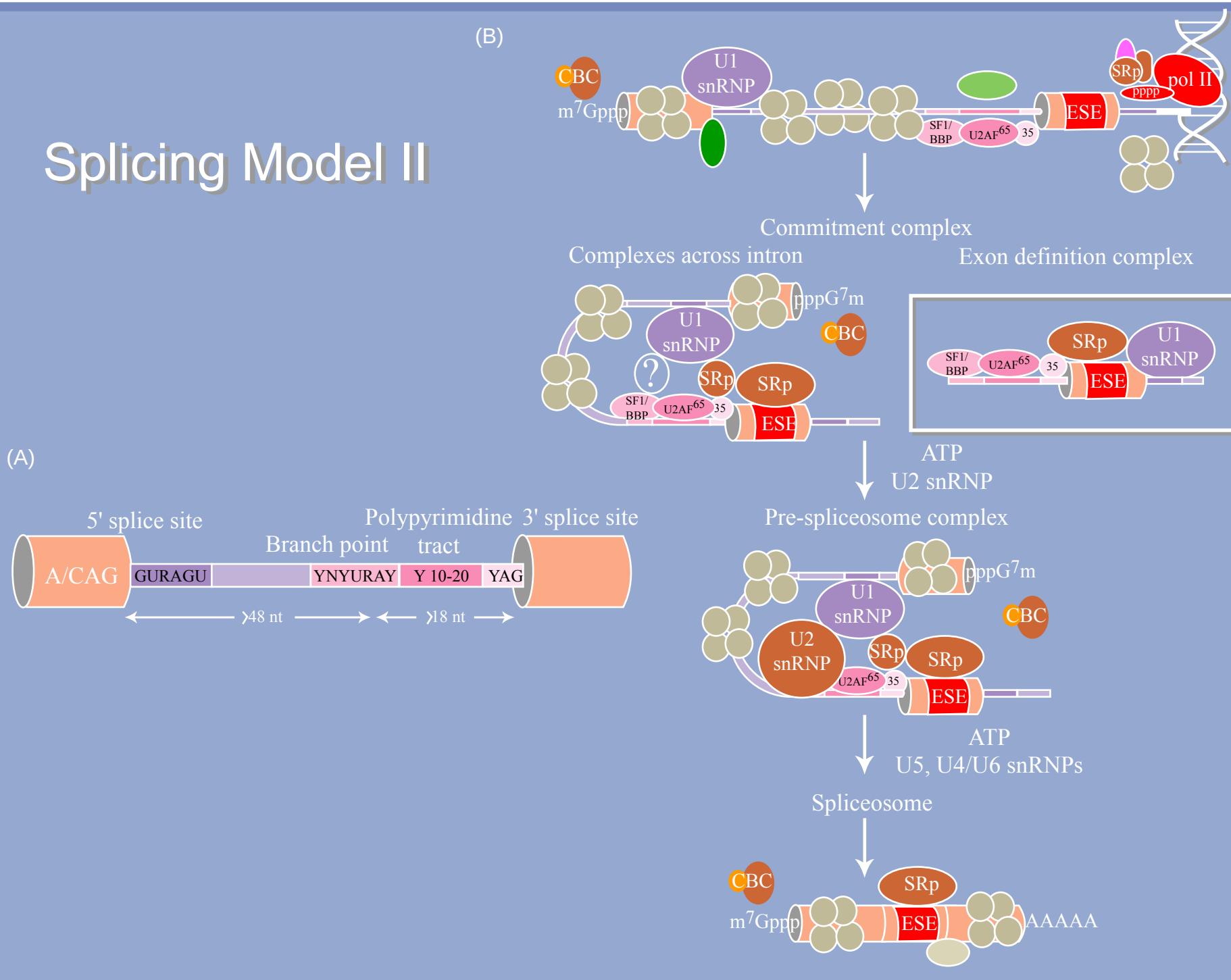


3′ splice site

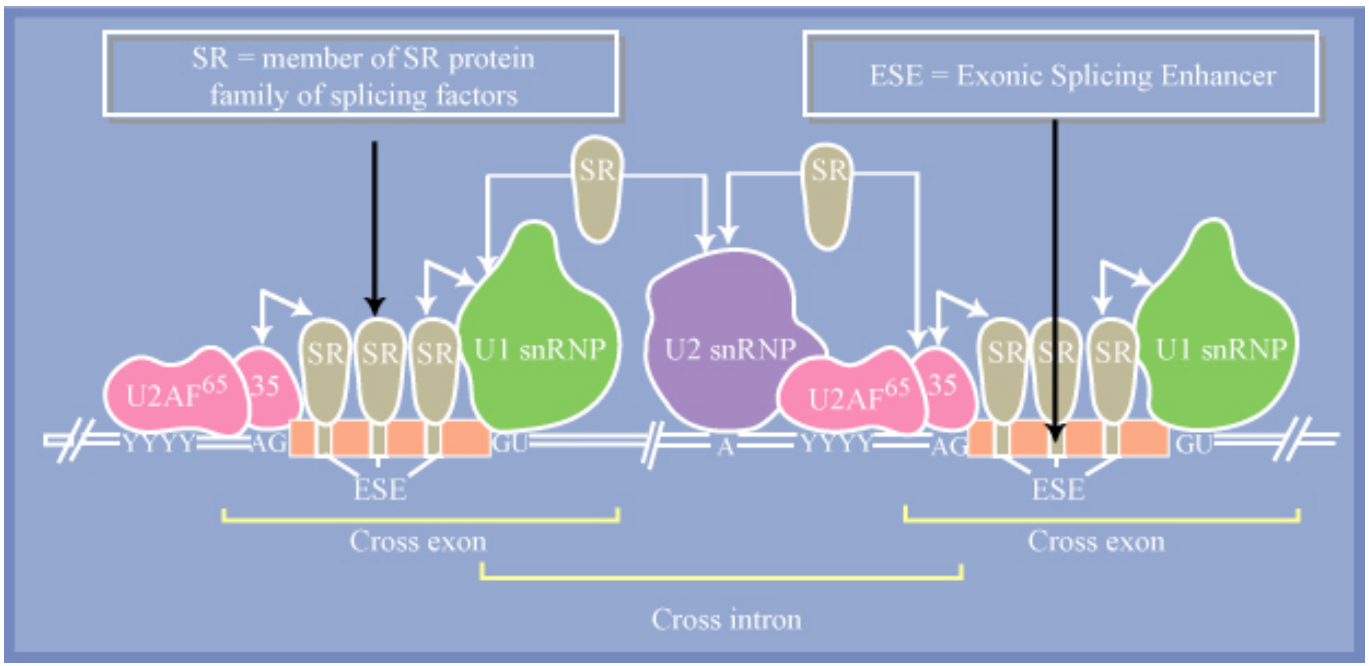




Splicing Model II



A Recent Model of Human Pre-mRNA Splicing



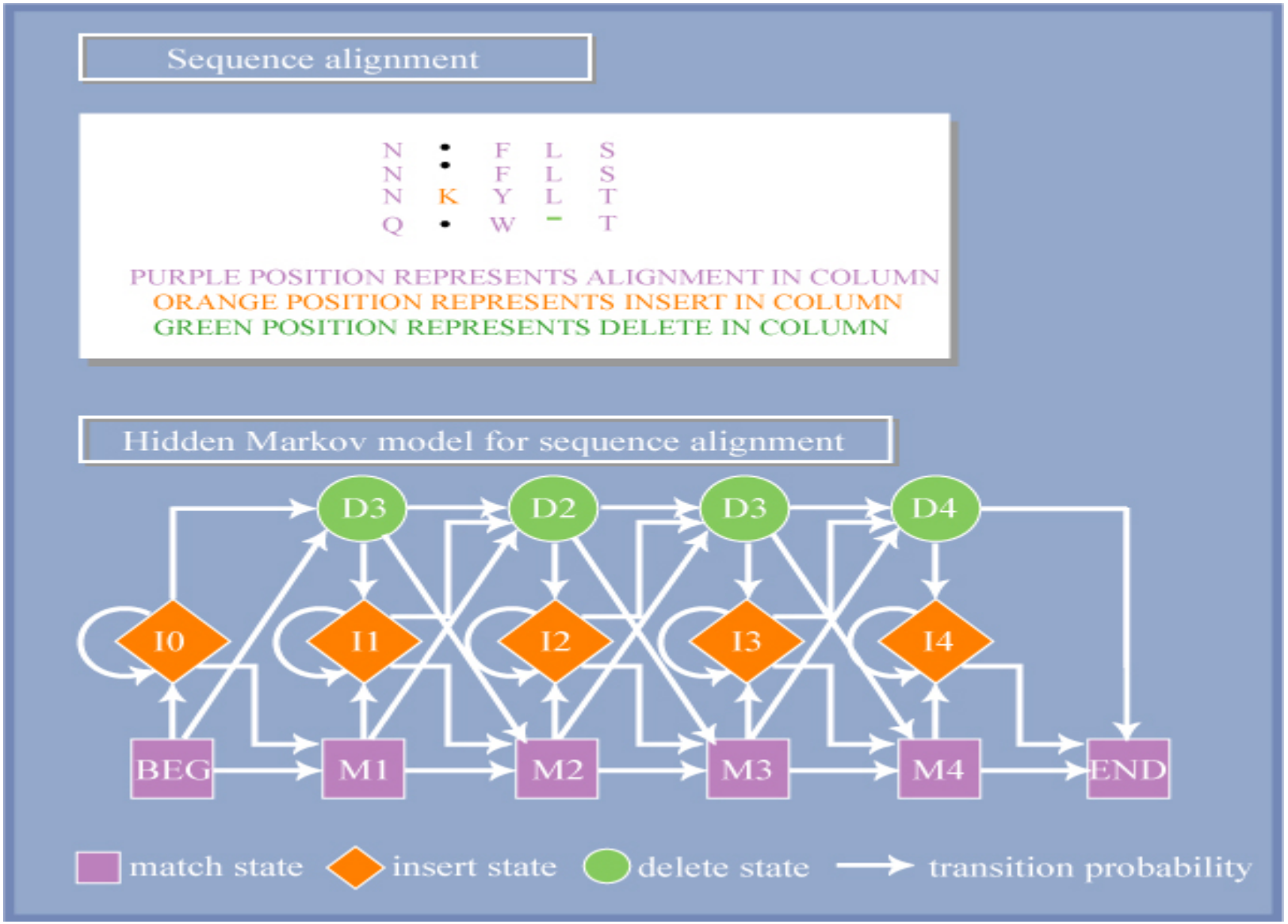
ESEs are short motifs that enhance recognition of adjacent splice sites in both constitutive and alternatively spliced exons - precise sequence requirements not well characterized

Hidden Markov Models

aka HMMs

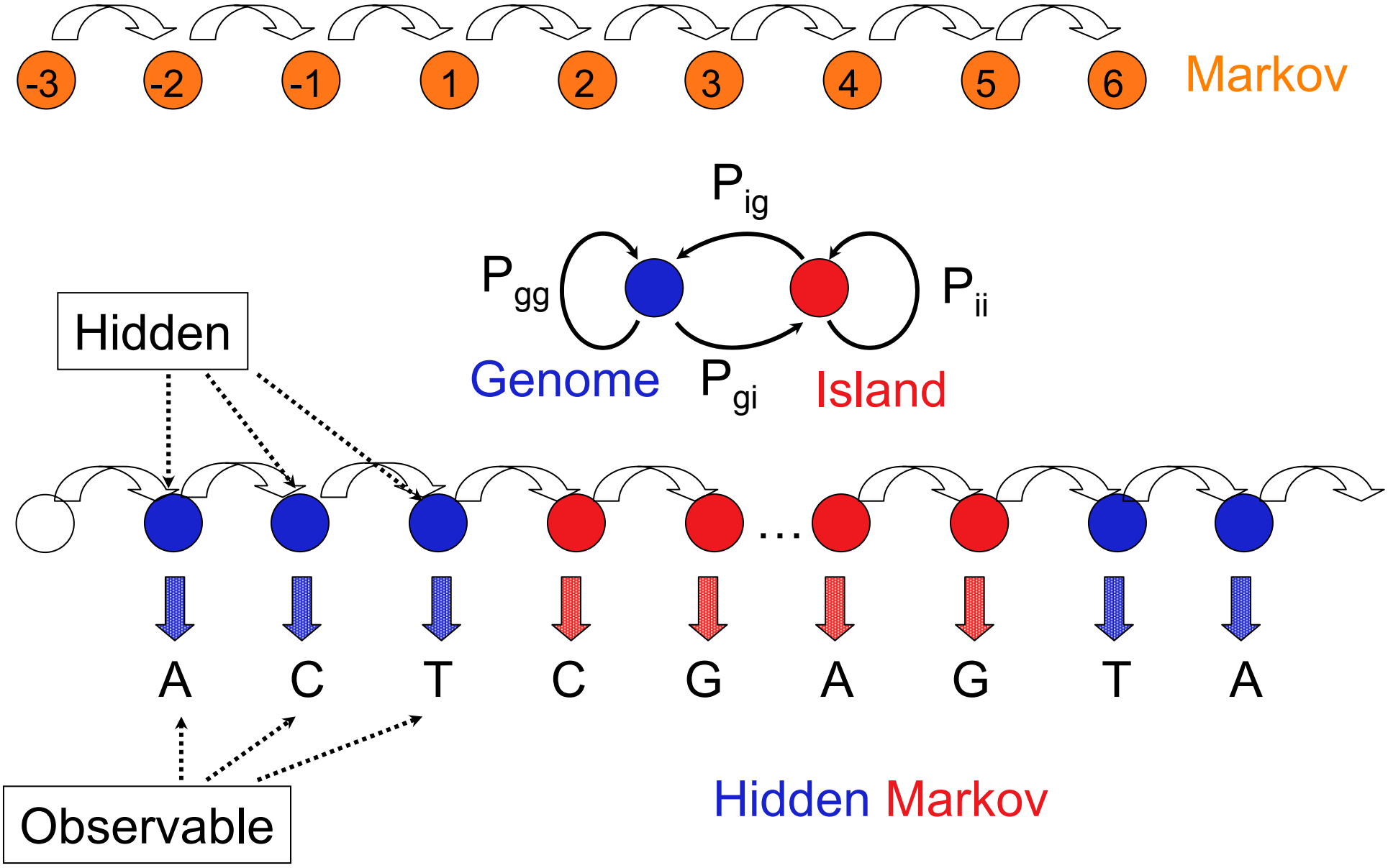
A later development, developed in E. E. for applications to voice recognition

Markov Models

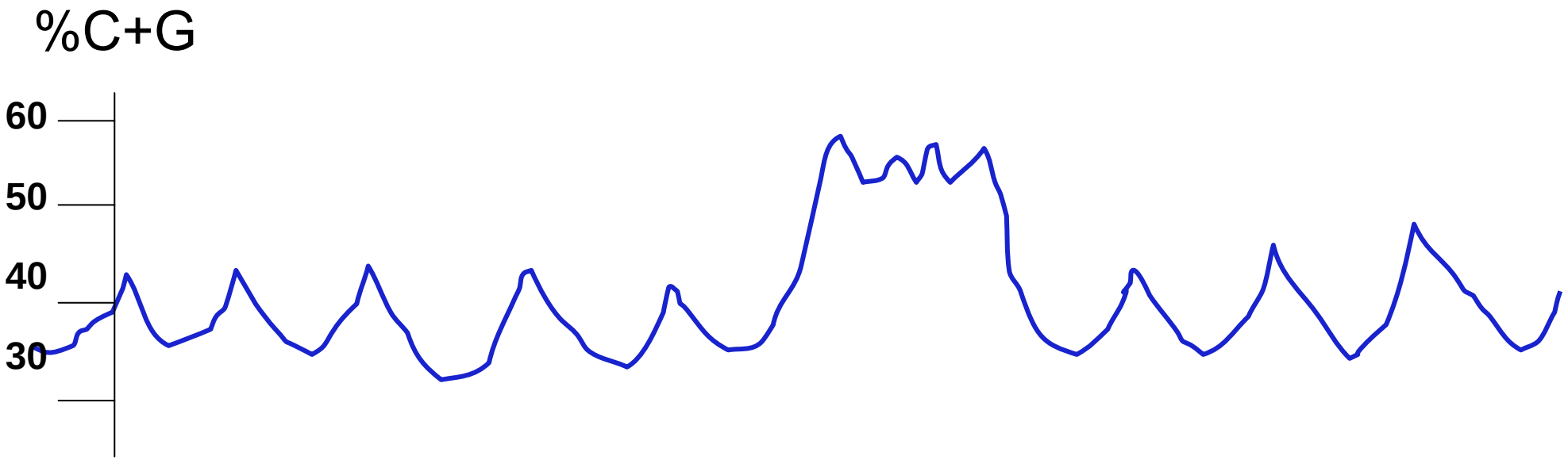


Courtesy of M. Yaffe

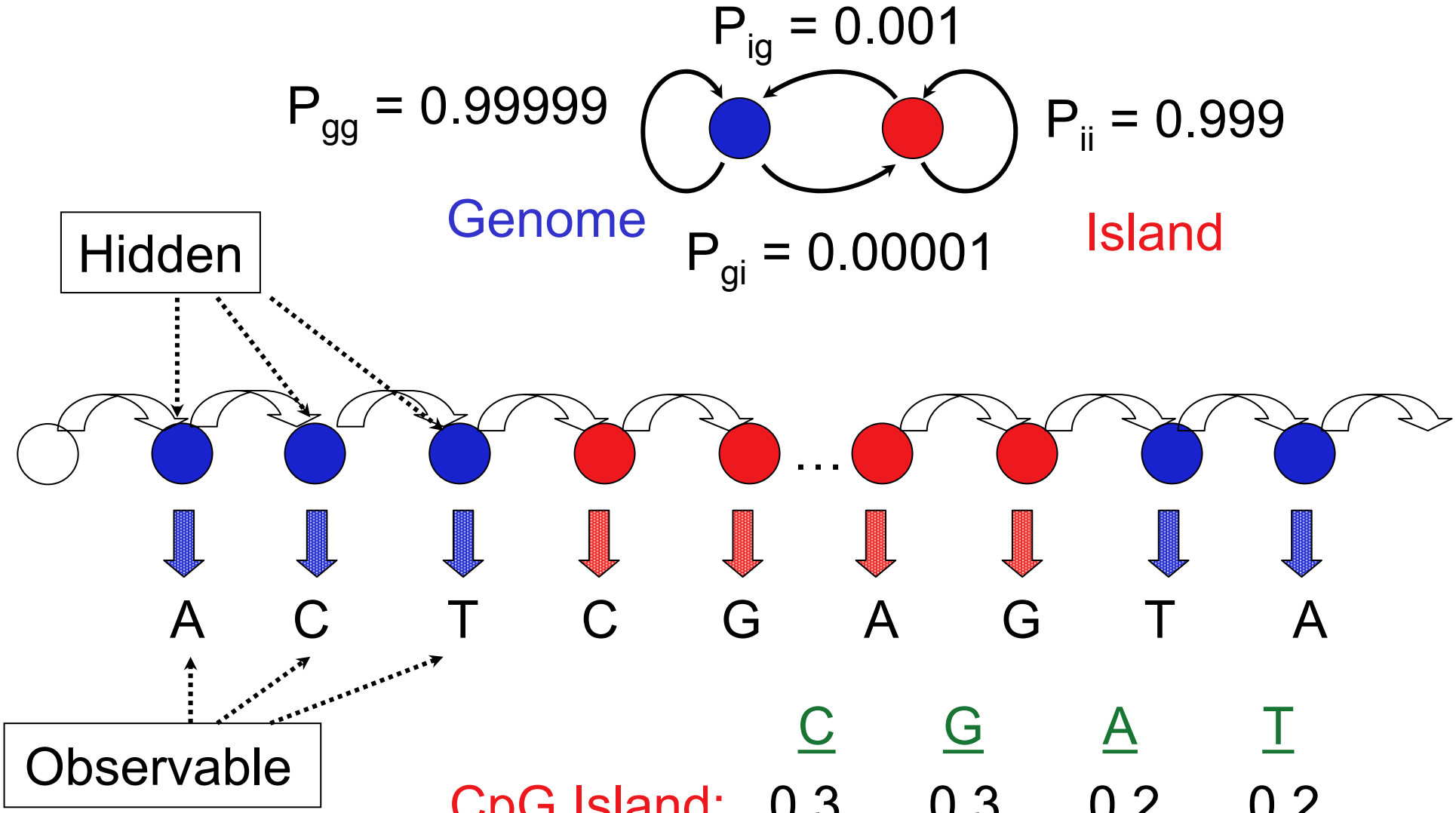
Markov and Hidden Markov Models



CpG Islands

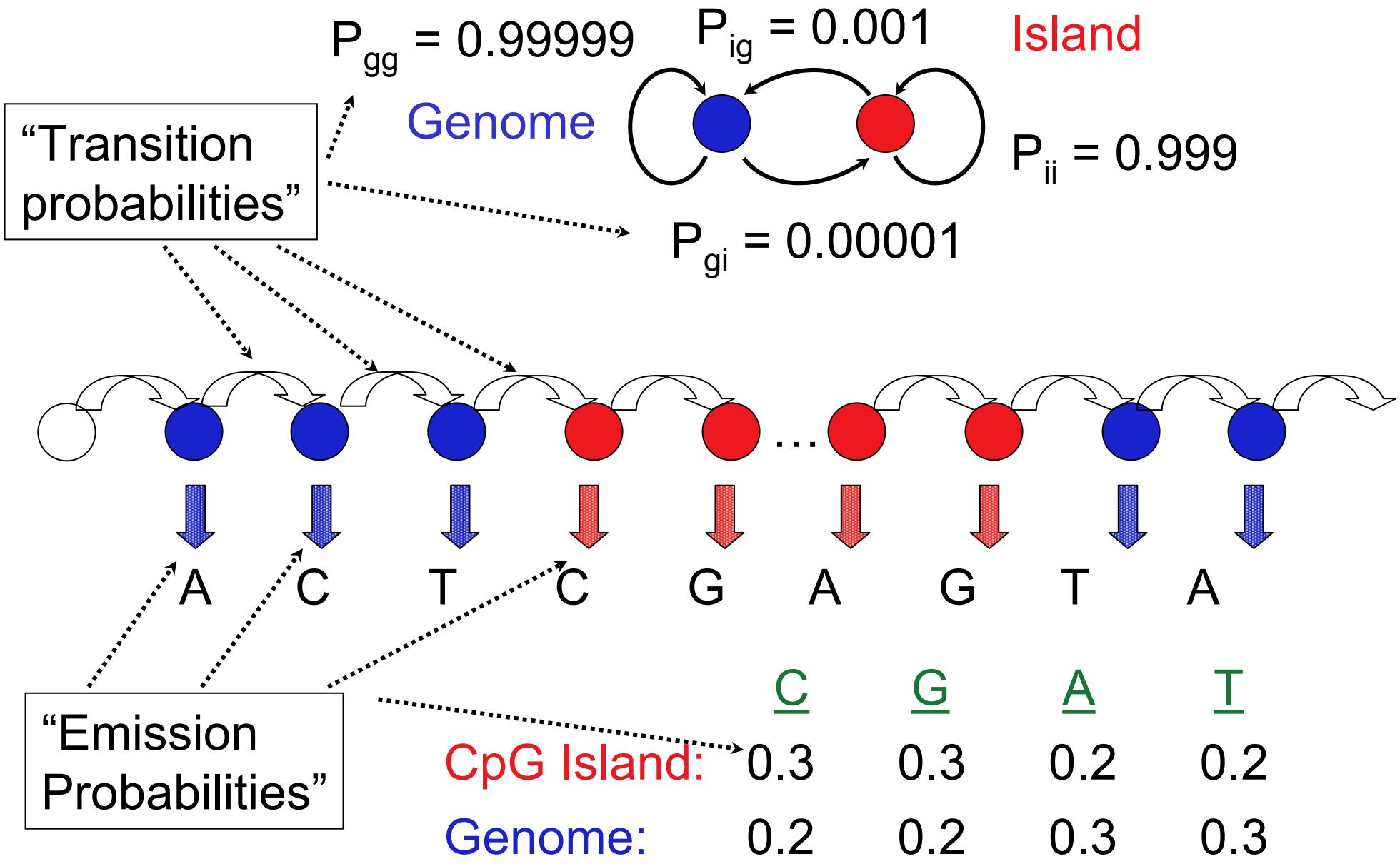


CpG Island Hidden Markov Model

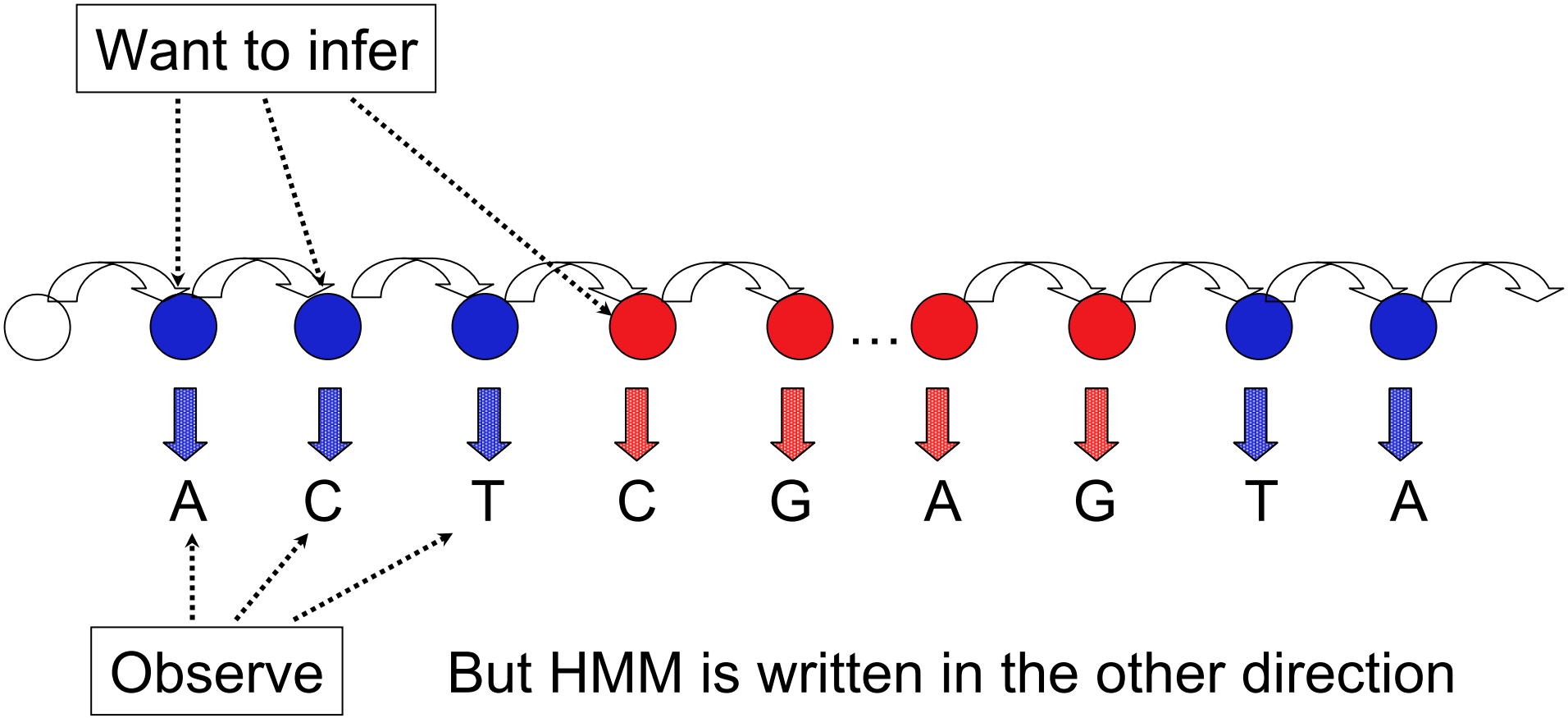


	C	G	A	T
CpG Island:	0.3	0.3	0.2	0.2
Genome:	0.2	0.2	0.3	0.3

CpG Island HMM II



CpG Island HMM III



But HMM is written in the other direction
(observable depends on hidden)

Inferring the Hidden from the Observable (Bayes' Rule)

$$P(H = h_1, h_2, ..., h_n \mid O = o_1, o_2, ..., o_n)$$

Conditional Prob:
P(A|B) = P(A,B)/P(B)

$$= \frac{P(H = h_1, ..., h_n, O = o_1, ..., o_n)}{P(O = o_1, ..., o_n)}$$

$$= \frac{P(H = h_1, ..., h_n) P(O = o_1, ..., o_n \mid H = h_1, ..., h_n)}{P(O = o_1, ..., o_n)}$$

$P(O = o_1, ..., o_n)$ somewhat difficult to calculate

But notice:

$$P(H = h_1, ..., h_n, O = o_1, ..., o_n) > P(H = h'_1, ..., h'_n, O = o_1, ..., o_n)$$

implies $P(H = h_1, ..., h_n \mid O = o_1, ..., o_n) > P(H = h'_1, ..., h'_n \mid O = o_1, ..., o_n)$

so can treat $P(O = o_1, ..., o_n)$ as a constant

Finding the Optimal "Parse" (Viterbi Algorithm)

Want to find sequence of hidden states $H^{opt} = h_1^{opt}, h_2^{opt}, h_3^{opt}, \ldots$

which maximizes joint probability: $P(H = h_1, \ldots, h_n, O = o_1, \ldots, o_n)$

(optimal "parse" of sequence)

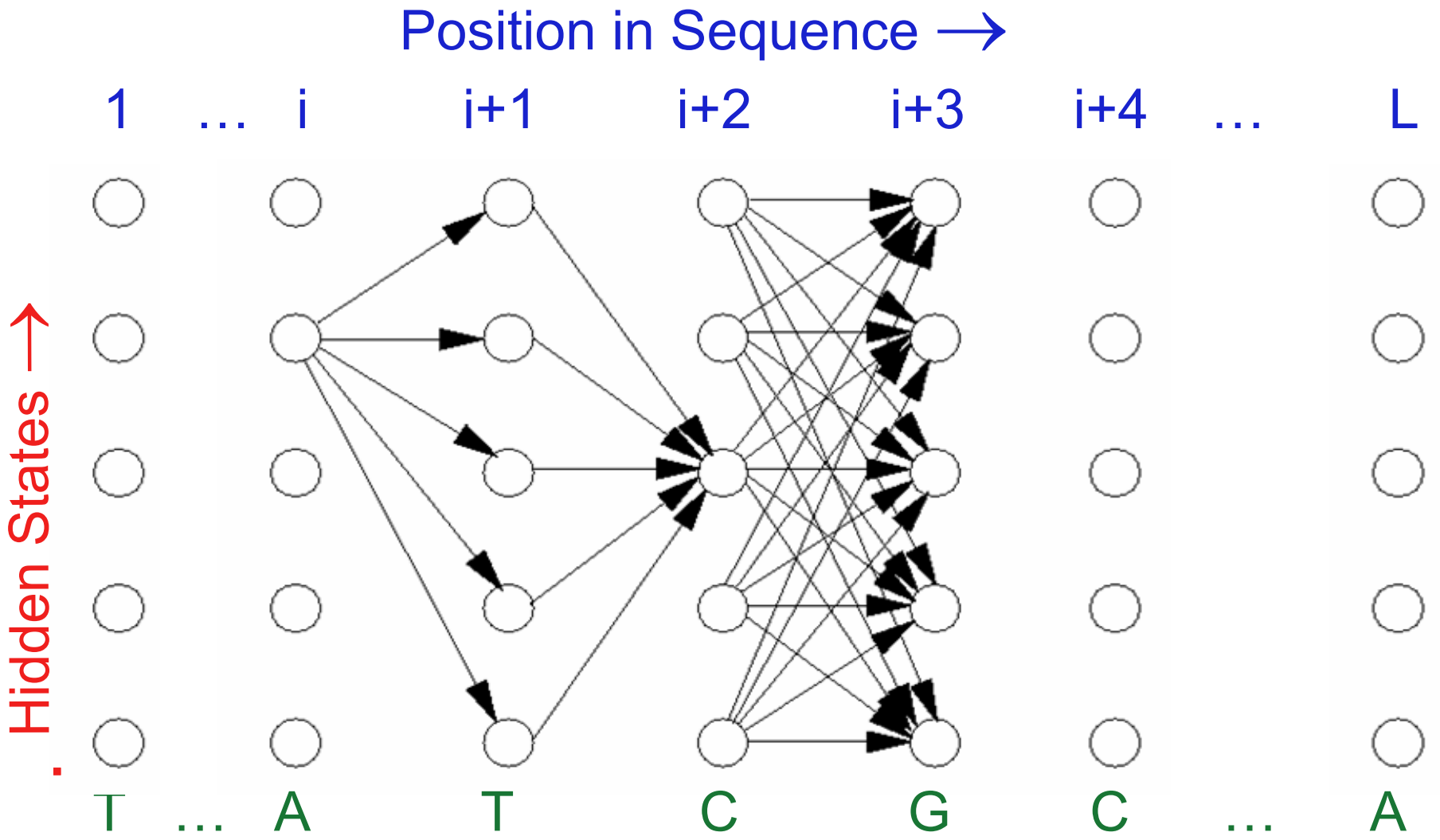
Solution:

Define $R_i^{(h)}$ = probability of optimal parse of the subsequence 1..i ending in state h

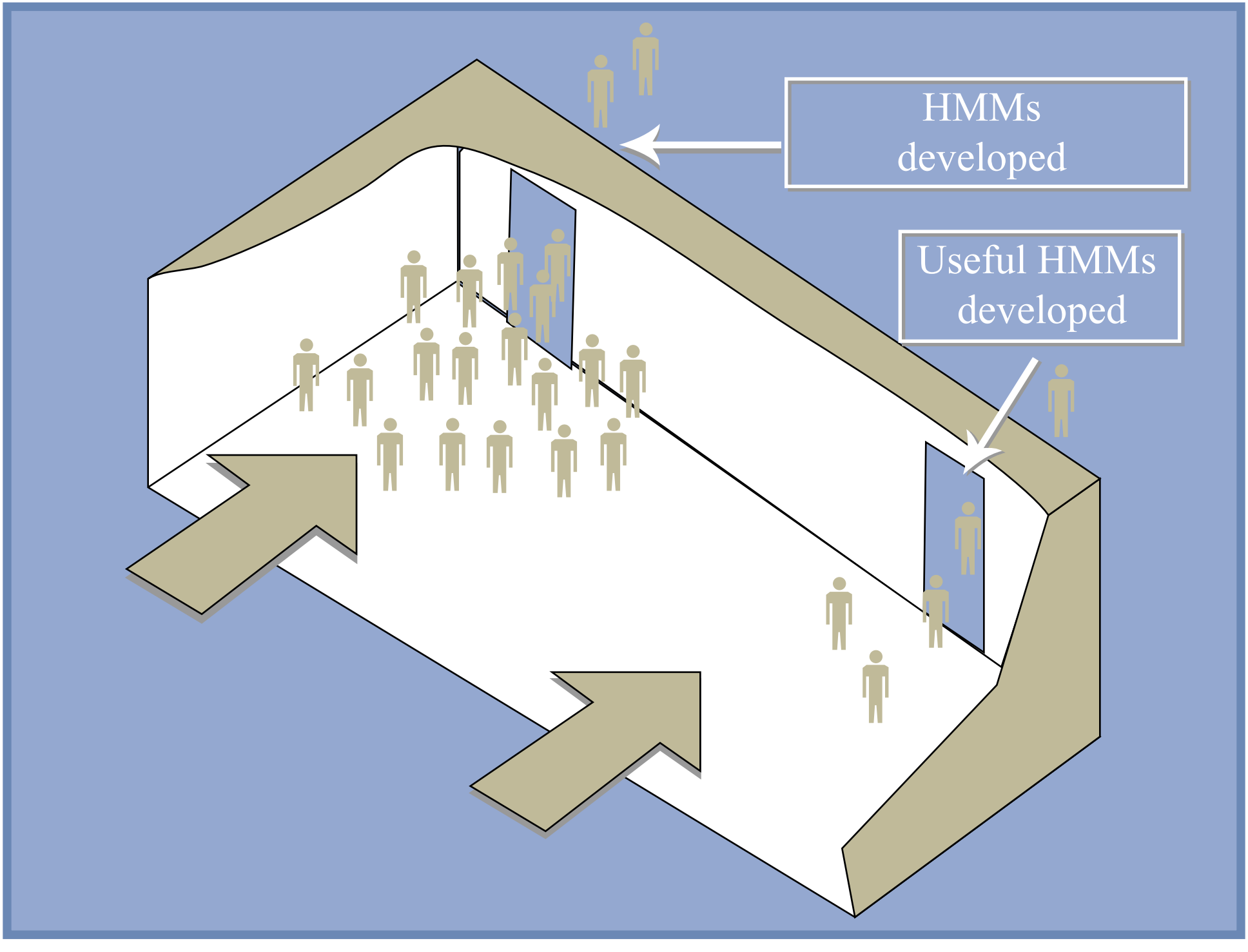
Solve recursively, i.e. determine $R_2^{(h)}$ in terms of $R_1^{(h)}$, etc.

A. Viterbi, an MIT BS/Meng student in E.E. - founder of Qualcomm

“Trellis” Diagram for Viterbi Algorithm



Run time for k-state HMM on sequence of length L?


HMMs
developed
Useful HMMs
developed