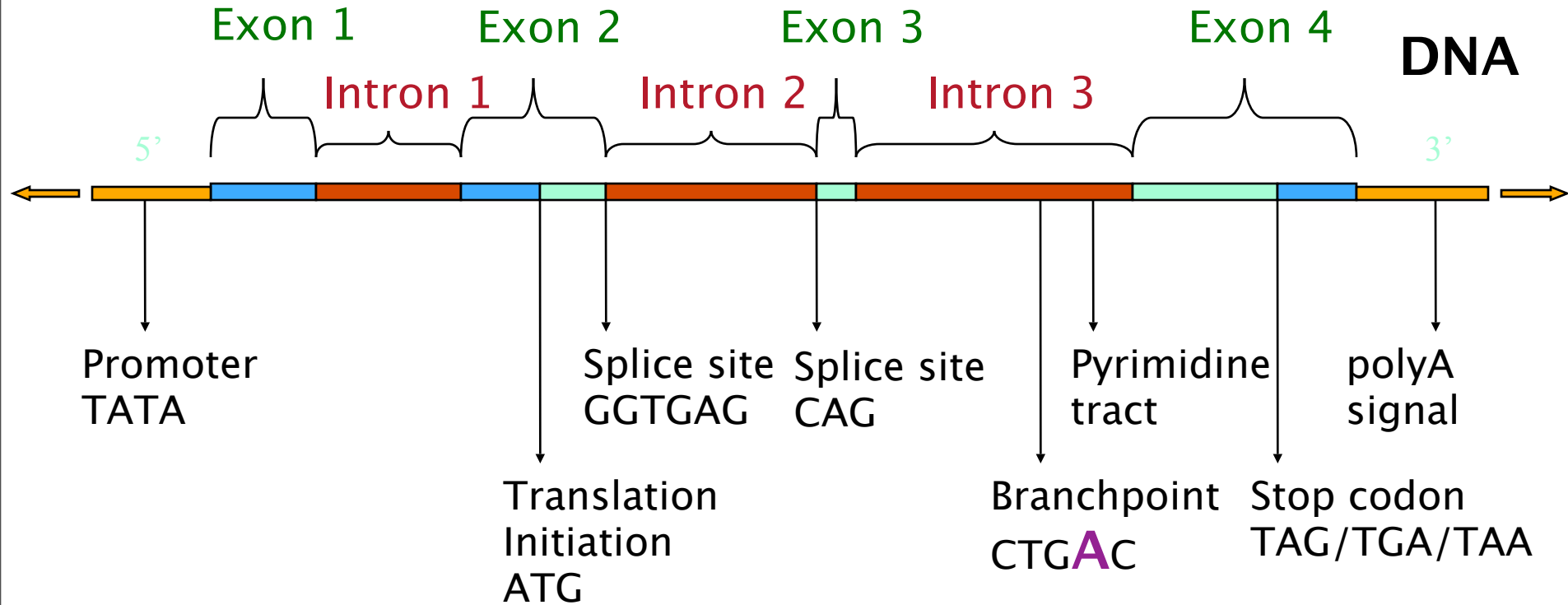


# Gene finding with Hidden Markov models

Rachel Brem, Mike Eisen, Lior Pachter

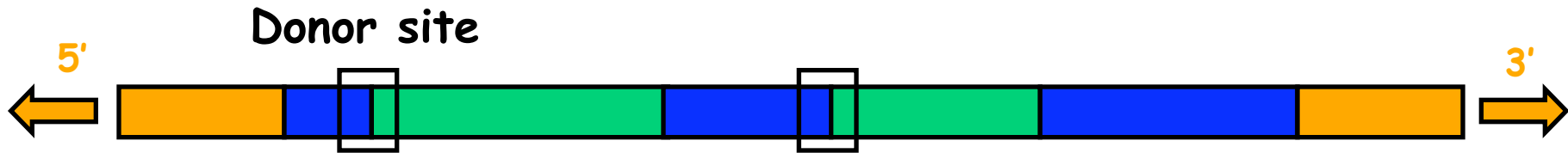
# Gene Structure



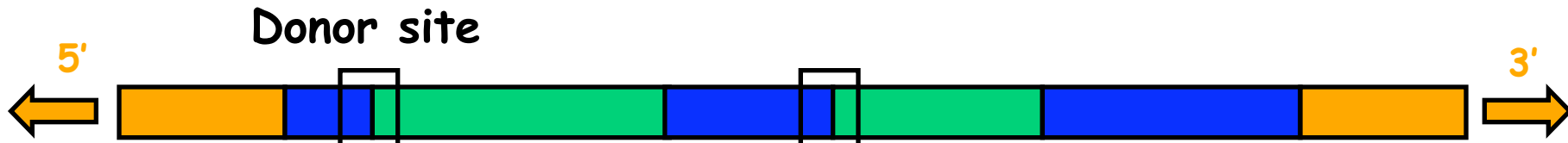
# Splice site modeling



# Splice site modeling



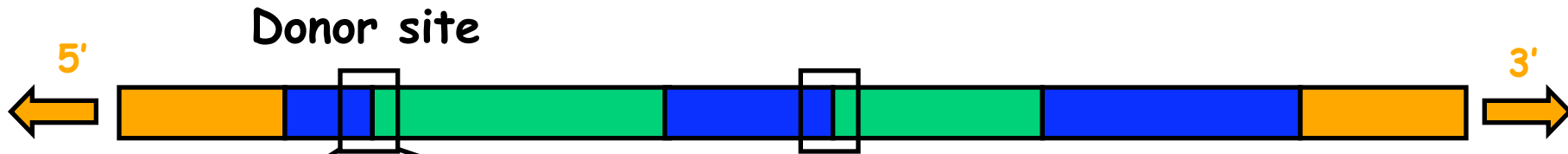
# Splice site modeling



**Position**

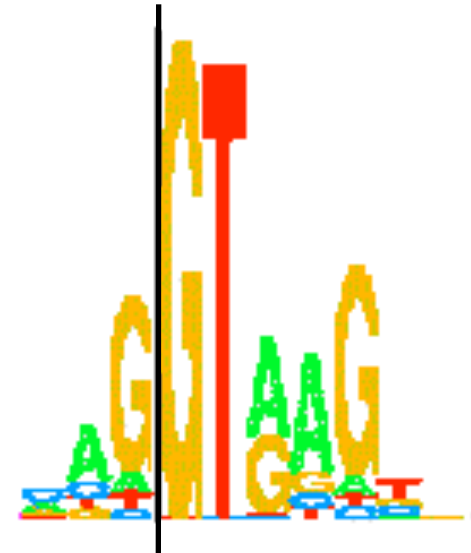
%	-8	...	-2	-1	0	1	2	...	17
A	26	...	60	9	0	1	54	...	21
C	26	...	15	5	0	1	2	...	27
G	25	...	12	78	99	0	41	...	27
T	23	...	13	8	1	98	3	...	25

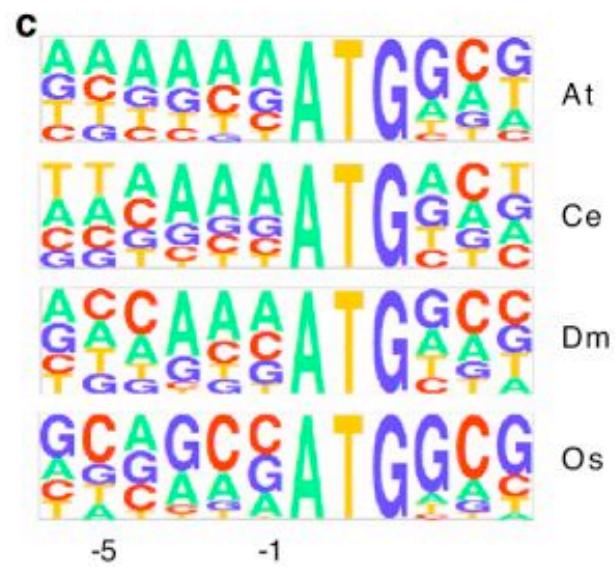
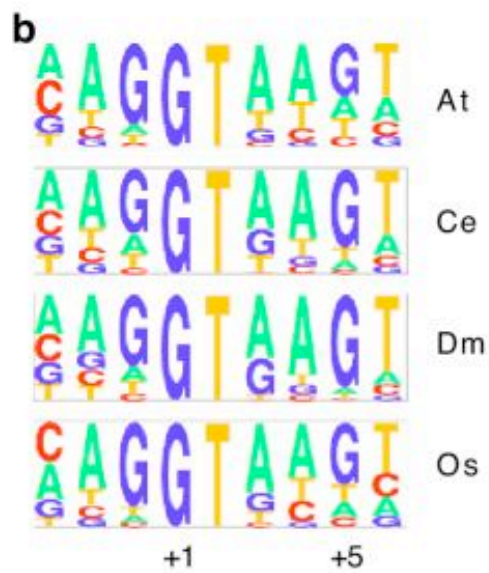
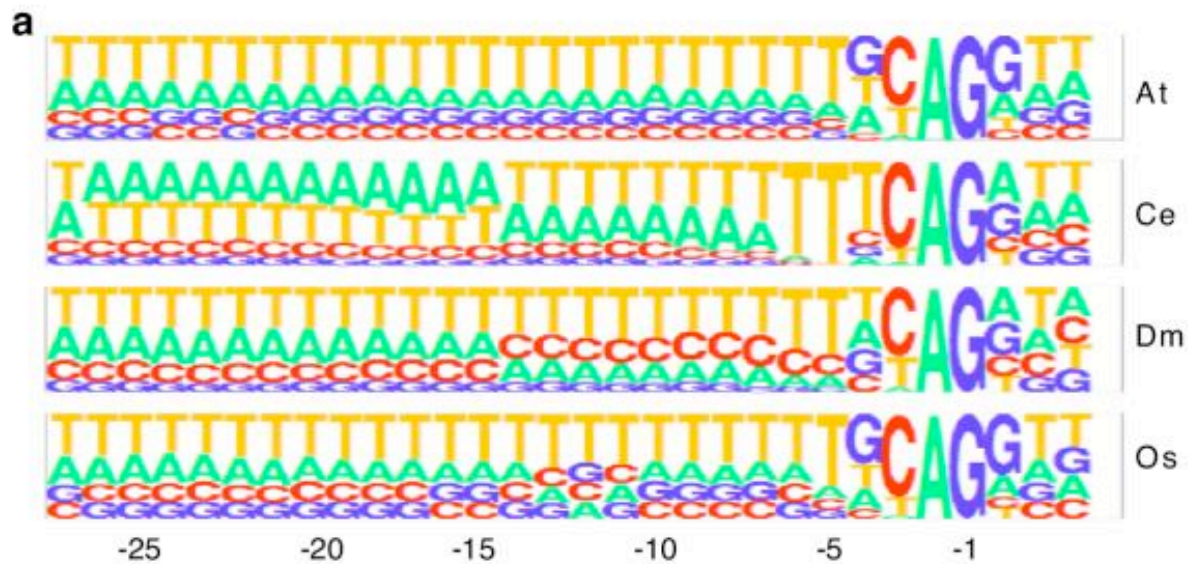
# Splice site modeling



Position

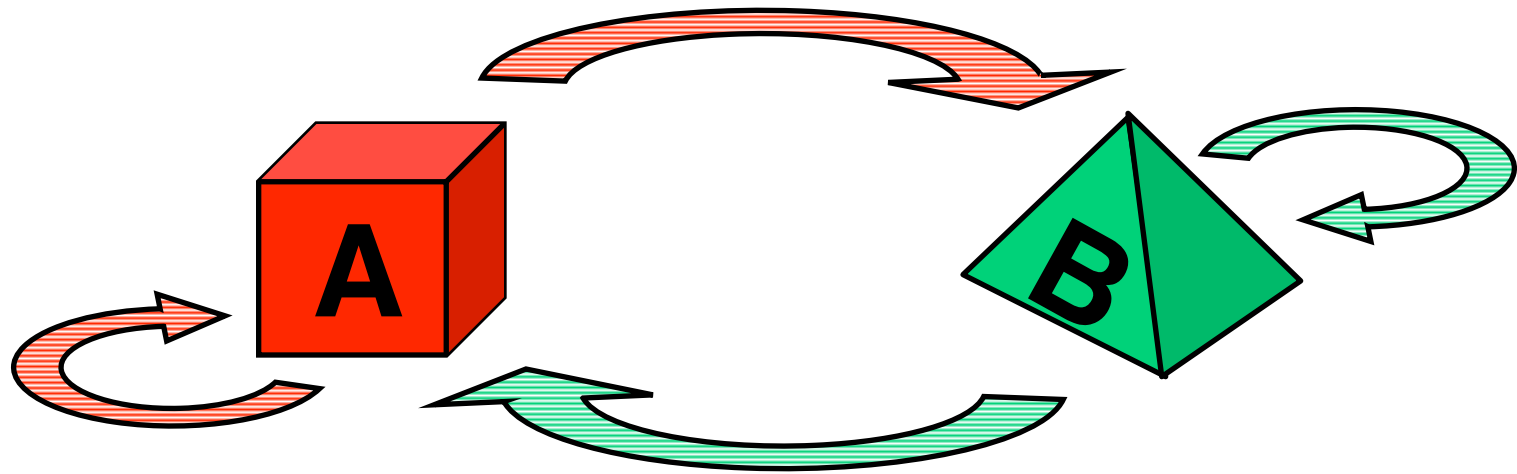
%	-8	...	-2	-1	0	1	2	...	17
A	26	...	60	9	0	1	54	...	21
C	26	...	15	5	0	1	2	...	27
G	25	...	12	78	99	0	41	...	27
T	23	...	13	8	1	98	3	...	25



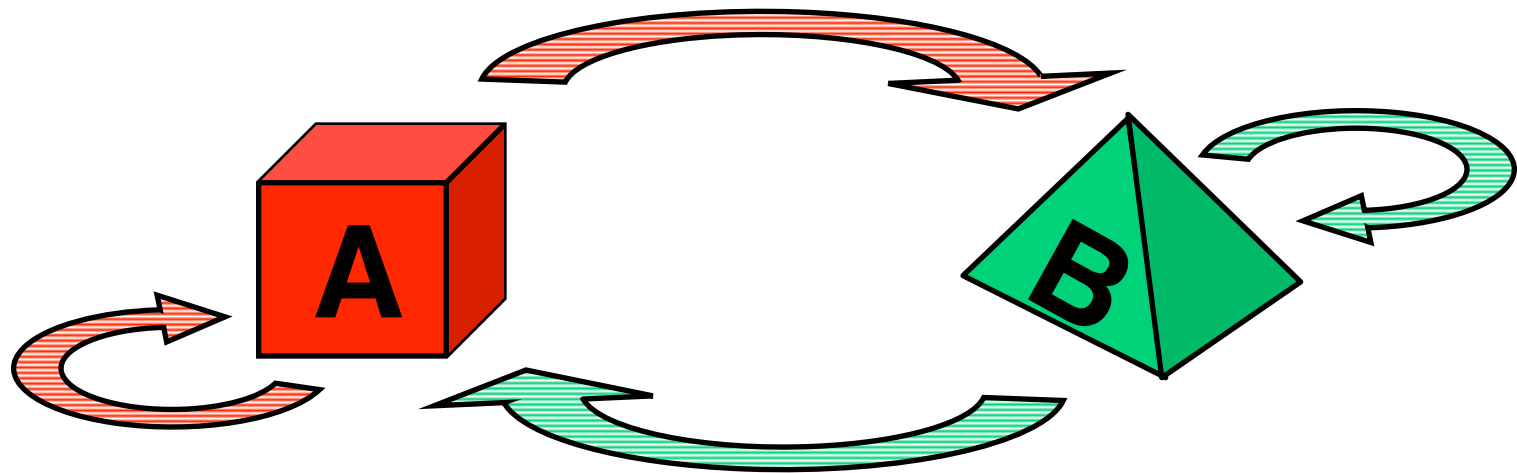




# A simple HMM



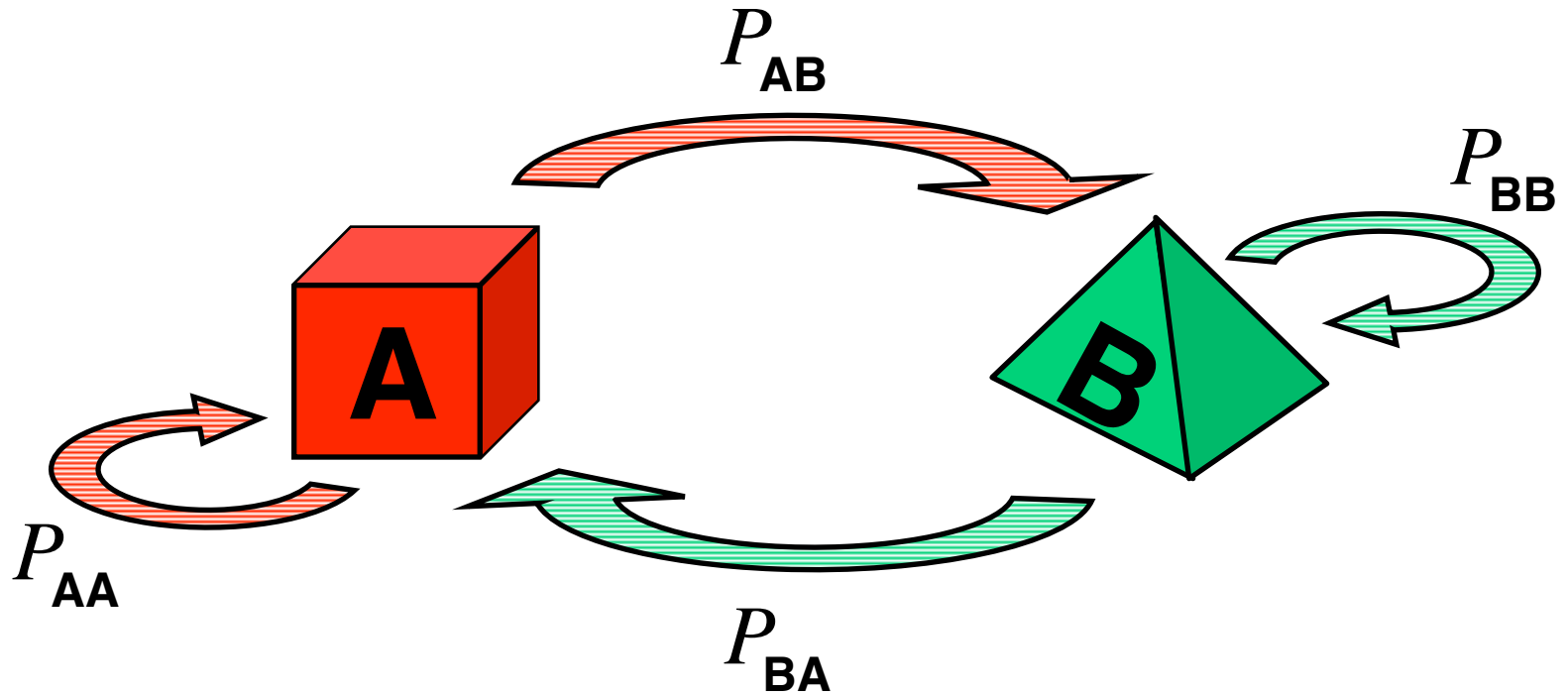
# A simple HMM



Initial distribution:

$$\pi = (\pi_A, \pi_B)$$

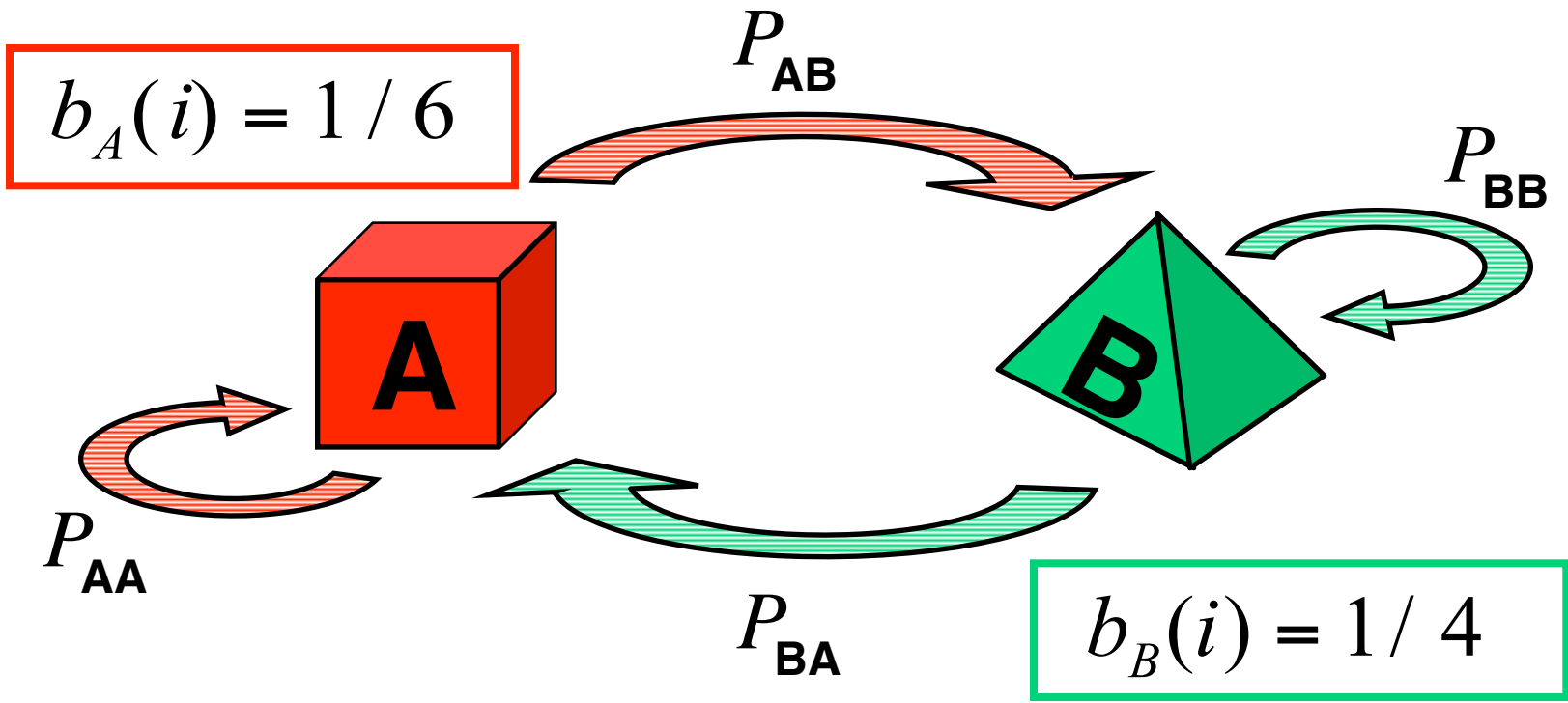
# A simple HMM



Initial distribution:

$$\pi = (\pi_A, \pi_B)$$

# A simple HMM



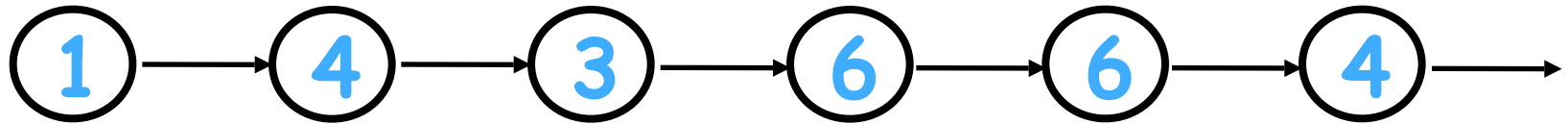
Initial distribution:

$$\pi = (\pi_A, \pi_B)$$

A lattice view

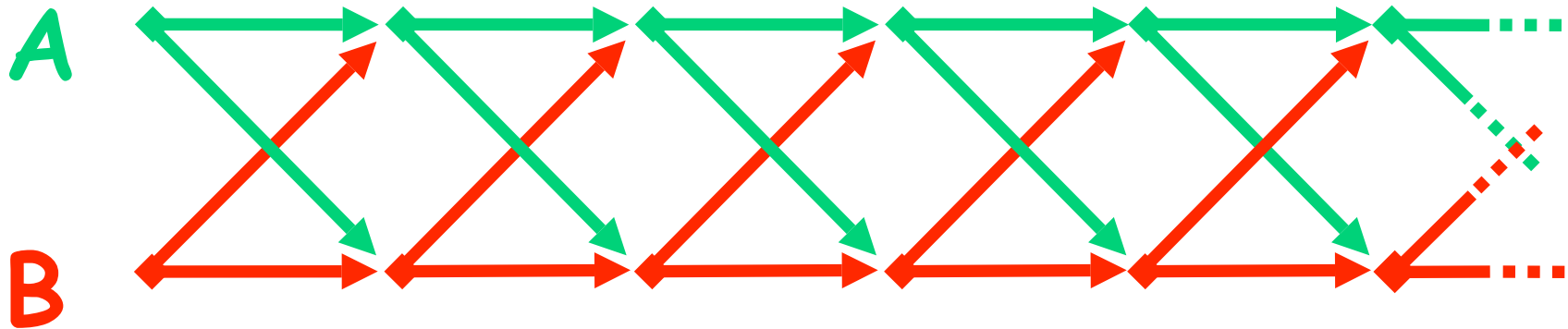
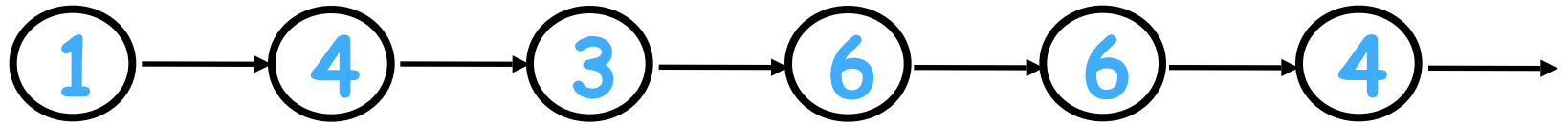
# A lattice view

Observed sequence:



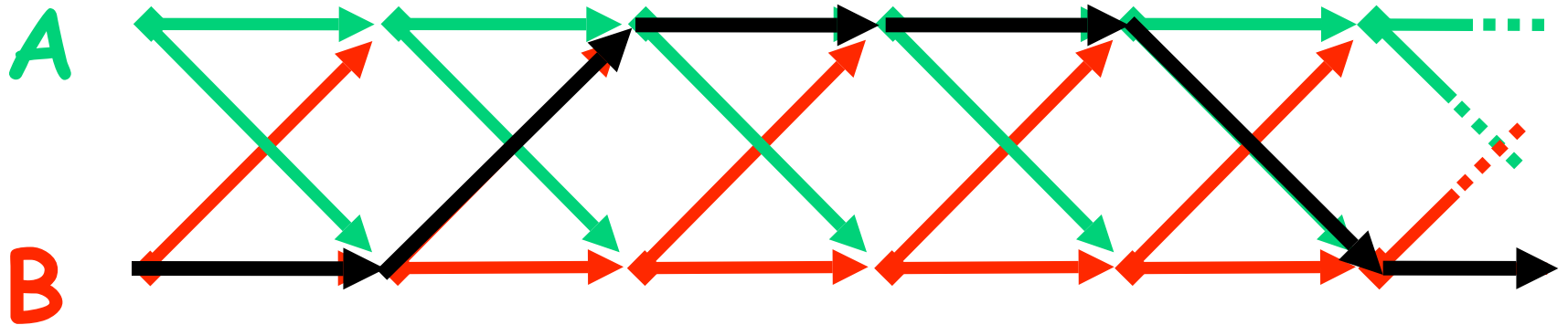
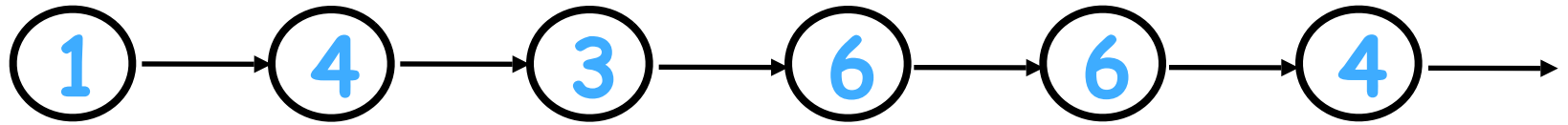
# A lattice view

Observed sequence:



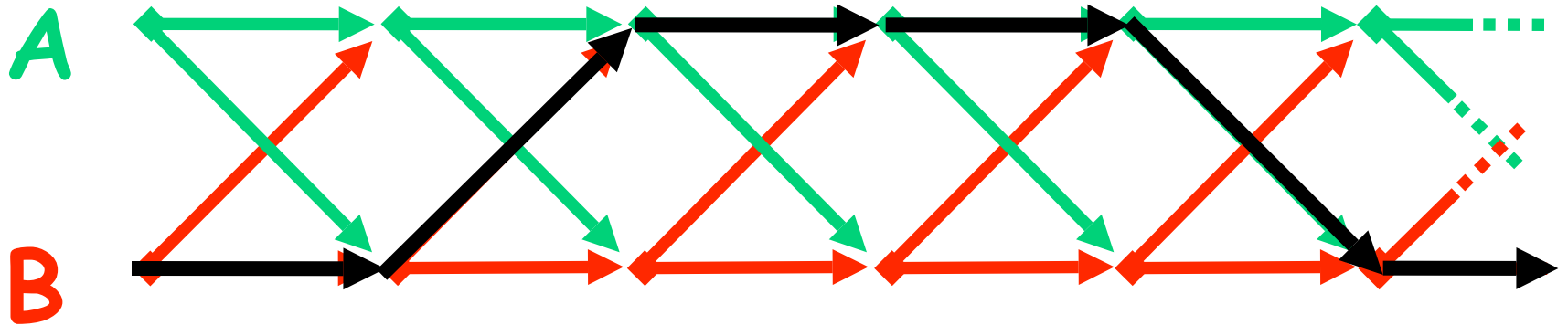
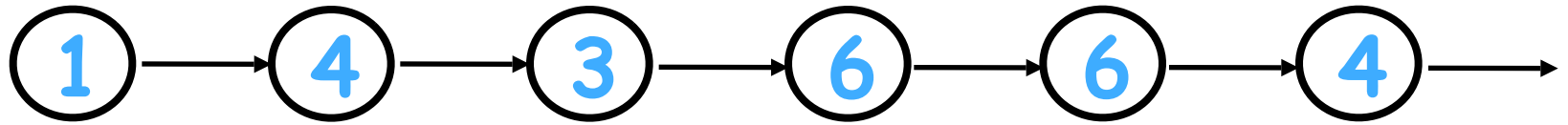
# A lattice view

Observed sequence:

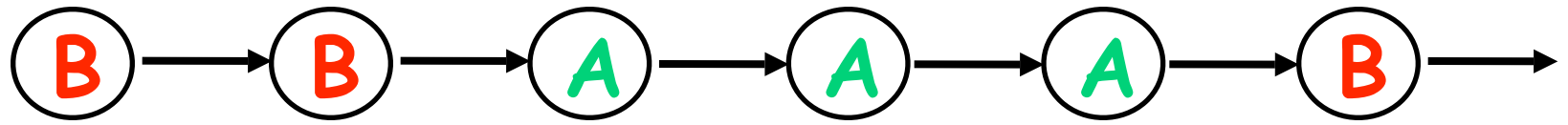


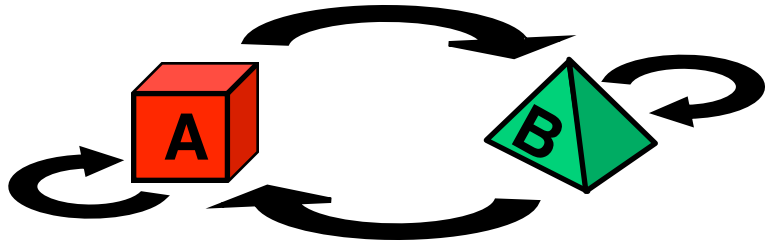
# A lattice view

Observed sequence:



Hidden sequence:





Observed:  
1,4,3,6,6,4...

Questions:

1. What is the most likely die sequence?
2. What is the probability of the observed sequence?
3. What is the probability that the 3<sup>rd</sup> state is B, given the observed sequence?

# The HMM algorithms

**Forward:**

$$\alpha_t(i) = P(\text{observed sequence, ending in state } i \text{ at base } t)$$

**Backward:**

$$\beta^t(i) = P(\text{obs. after } t \mid \text{ending in state } i \text{ at base } t)$$

**Viterbi:**

$$\delta^t(i) = \max P(\text{obs.}, \text{ending in state } i \text{ at base } t)$$

**Questions:**

1. What is the most likely die sequence?
2. What is the probability of the observed sequence?
3. What is the probability that the 3<sup>rd</sup> state is B, given the observed sequence?

# The HMM algorithms

Forward:

$$\alpha_t(i) = P(\text{observed sequence, ending in state } i \text{ at base } t)$$

Backward:

$$\beta^t(i) = P(\text{obs. after } t \mid \text{ending in state } i \text{ at base } t)$$

Viterbi:

$$\delta^t(i) = \max P(\text{obs.}, \text{ending in state } i \text{ at base } t)$$

Questions:

1. What is the most likely die sequence? **Viterbi**
2. What is the probability of the observed sequence?
3. What is the probability that the 3<sup>rd</sup> state is B, given the observed sequence?

# The HMM algorithms

Forward:

$$\alpha_t(i) = P(\text{observed sequence, ending in state } i \text{ at base } t)$$

Backward:

$$\beta^t(i) = P(\text{obs. after } t \mid \text{ending in state } i \text{ at base } t)$$

Viterbi:

$$\delta^t(i) = \max P(\text{obs.}, \text{ending in state } i \text{ at base } t)$$

Questions:

1. What is the most likely die sequence? **Viterbi**
2. What is the probability of the observed sequence? **Forward**
3. What is the probability that the 3<sup>rd</sup> state is B, given the observed sequence?

# The HMM algorithms

Forward:

$$\alpha_t(i) = P(\text{observed sequence, ending in state } i \text{ at base } t)$$

Backward:

$$\beta^t(i) = P(\text{obs. after } t \mid \text{ending in state } i \text{ at base } t)$$

Viterbi:

$$\delta^t(i) = \max P(\text{obs.}, \text{ending in state } i \text{ at base } t)$$

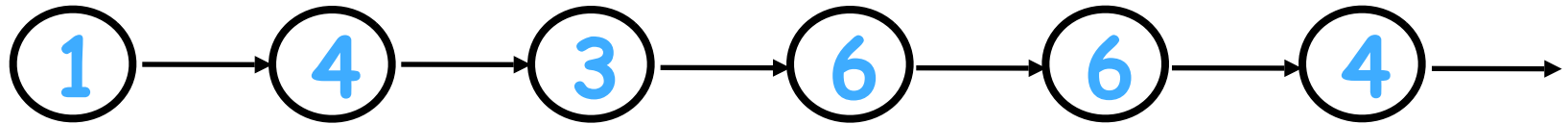
Questions:

1. What is the most likely die sequence? **Viterbi**
2. What is the probability of the observed sequence? **Forward**
3. What is the probability that the 3<sup>rd</sup> state is B, given the observed sequence? **Backward**

A lattice view

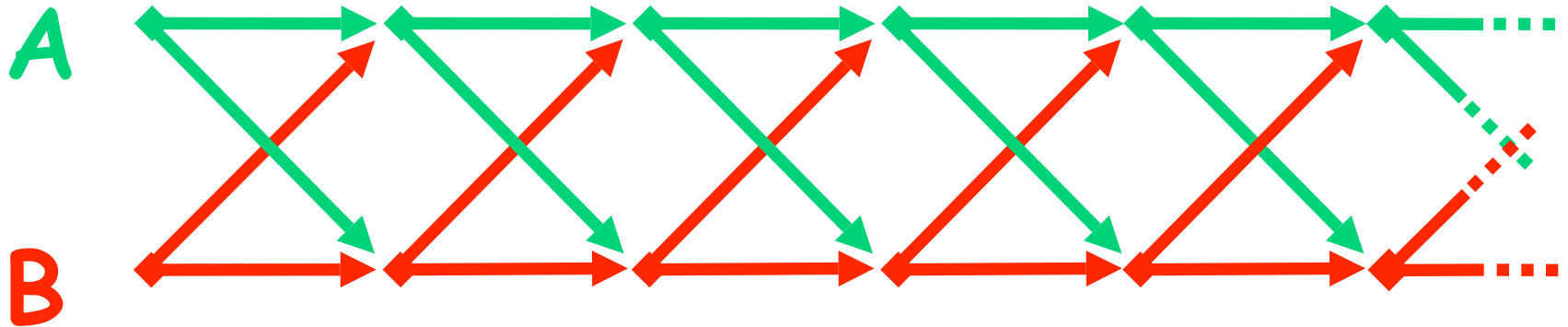
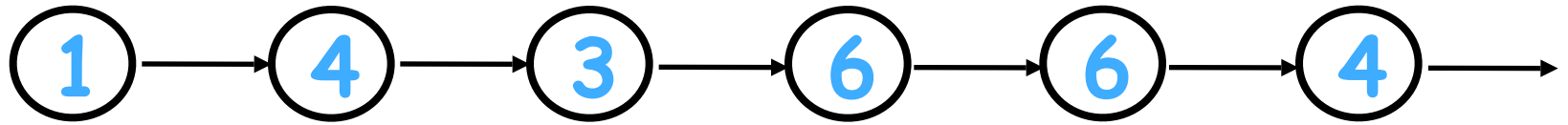
# A lattice view

Observed sequence:



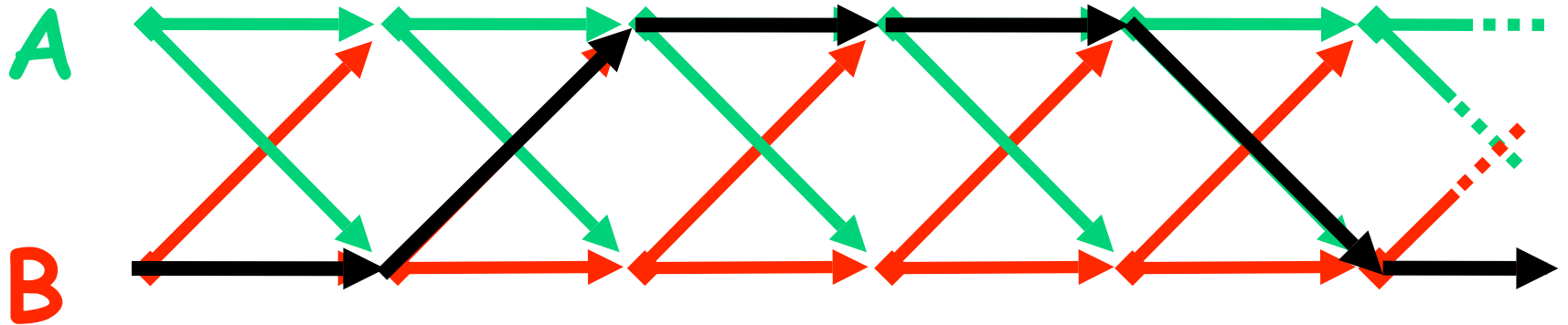
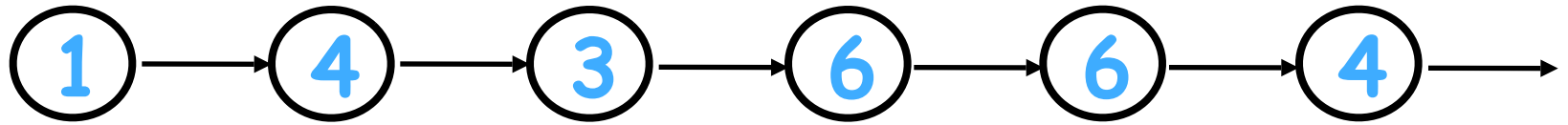
# A lattice view

Observed sequence:



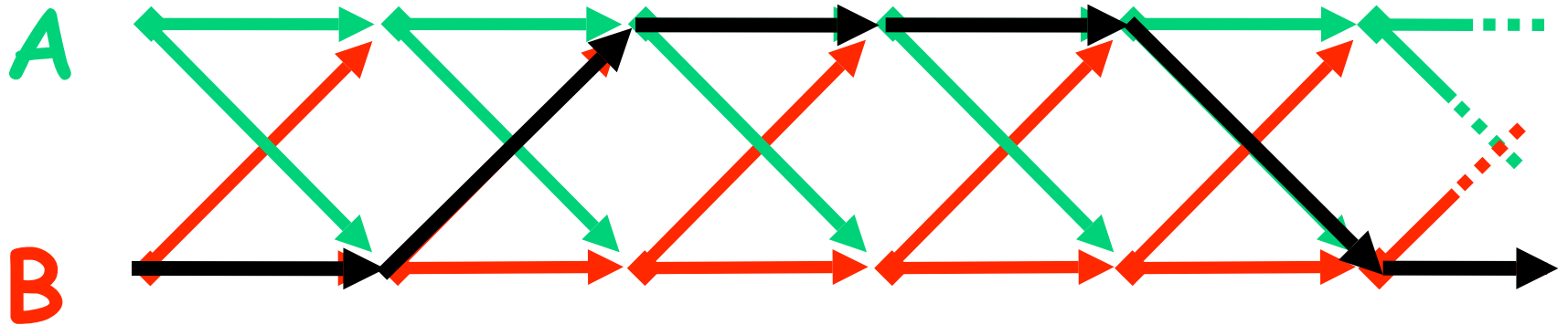
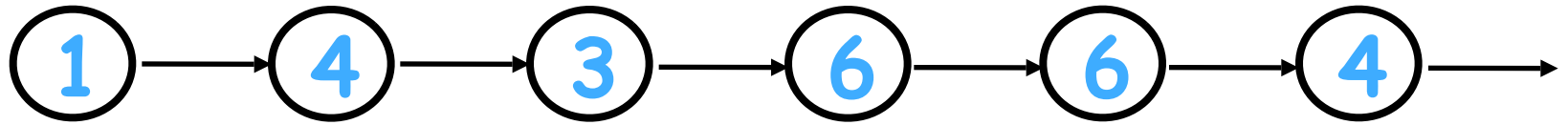
# A lattice view

Observed sequence:

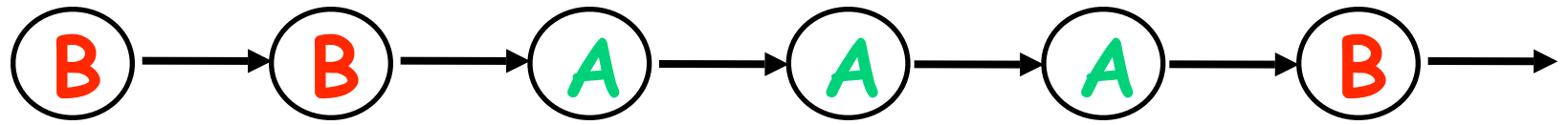


# A lattice view

Observed sequence:



Hidden sequence:



# Hidden Markov Models (HMMs)

- Underlying generates a sequence of states.

Markov chain = distribution of next state depends only on present

Hidden = the state sequence



Observed = outputs from the states

**GTCAGAGTAGCAAAGTAGACACTCCAGTAACGC**

# Approaches to Gene recognition

- Homology
  - BLAST, Procrustes, Exonerate
- De Novo
  - GRAIL, FGENESH, GENSCAN, Genie, Glimmer, SNAP
- Hybrids
  - GenomeScan, Genie
- Comparative
  - Rosetta, SLAM, Twinscan

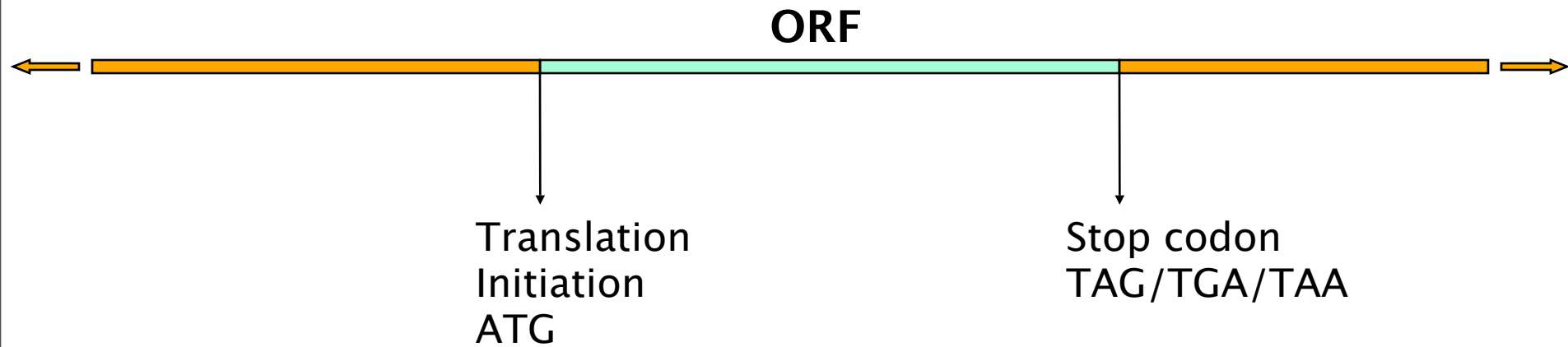
*Ab-initio gene finding:*

# Example: Glimmer

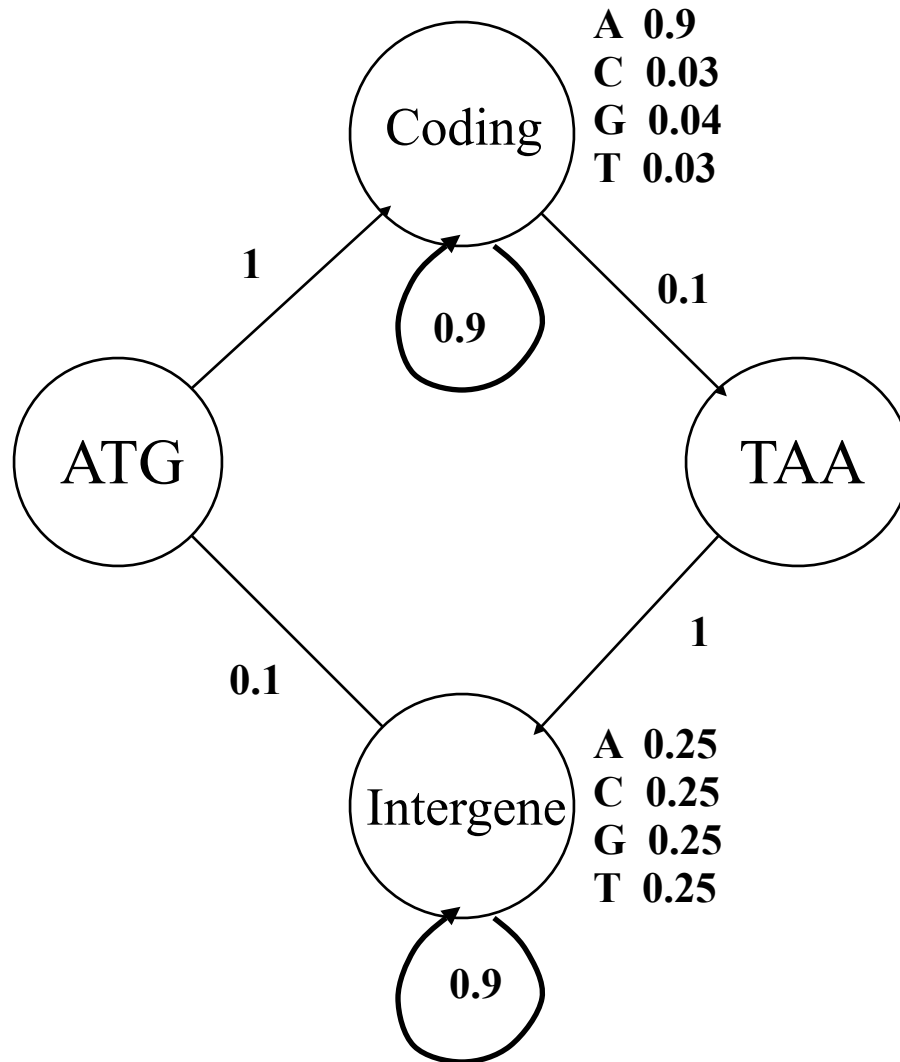
## Gene Finding in Microbial DNA

- No introns
- 90% coding
- Shorter genomes (less than 10 million bp)
- Lots of data

# Gene Structure in Prokaryotes

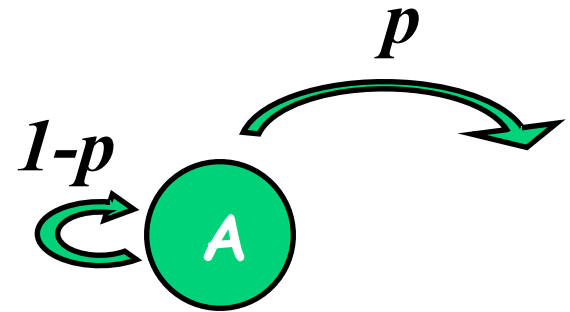


# Bacteriomaker



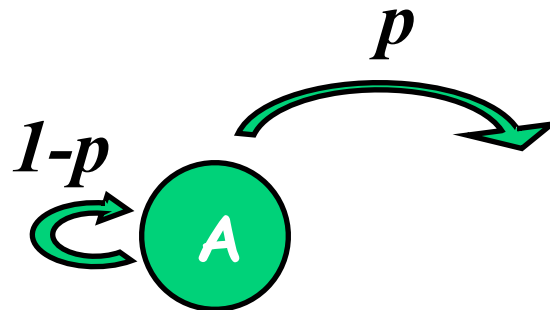
# HMM state duration times

- $\Pr(\text{leaving state}) = p$
  - $\Pr(\text{staying in state}) = 1 - p$
  - $\Pr(\text{output of exactly } r \text{ in state}) = (1-p)^r p$
- 
- Geometric distribution

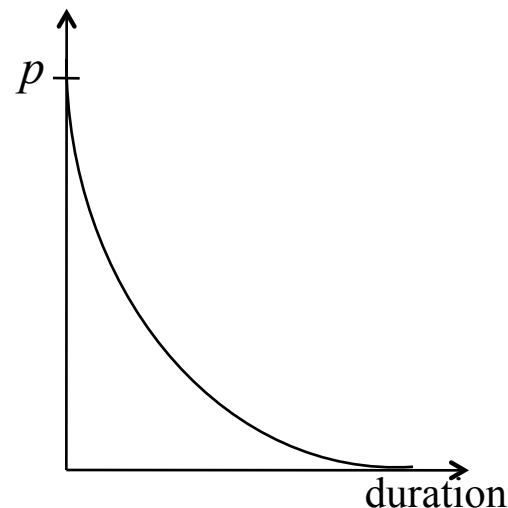


# HMM state duration times

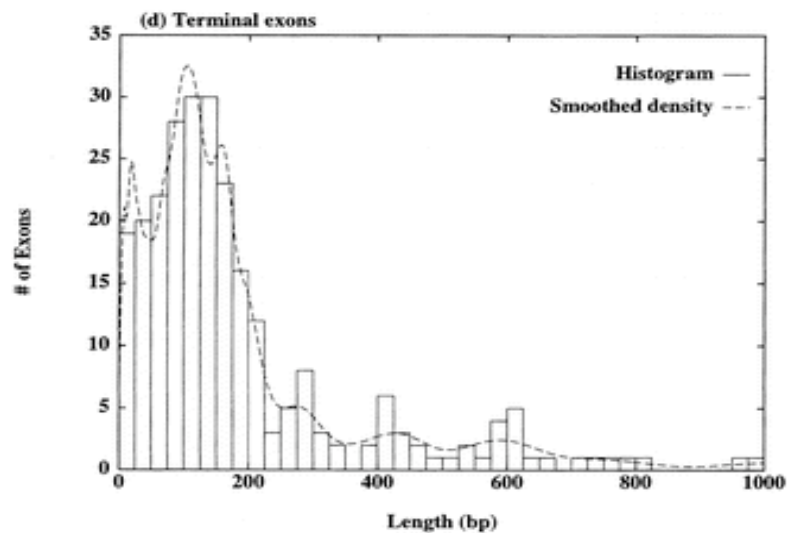
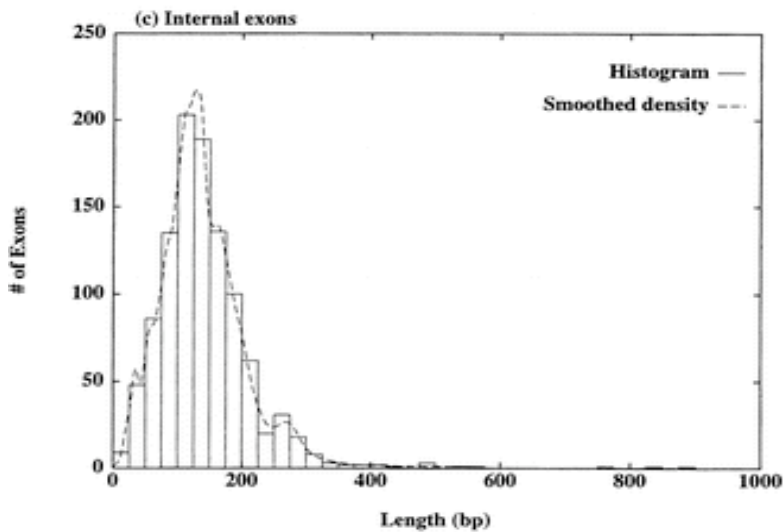
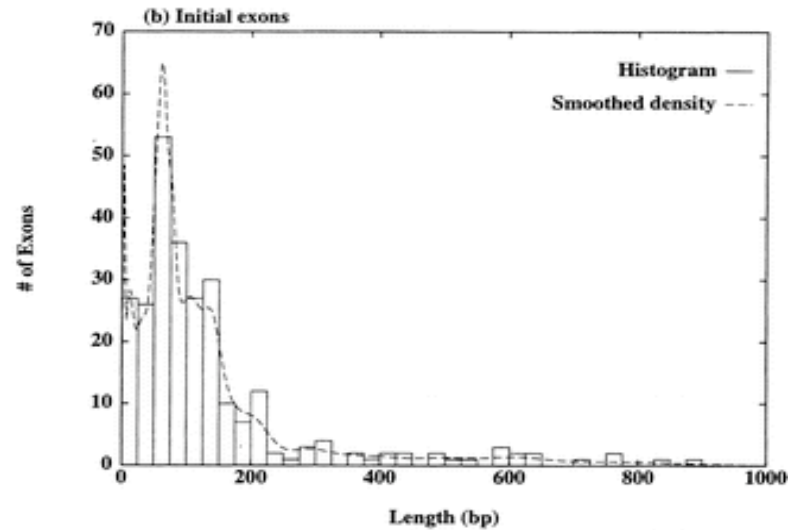
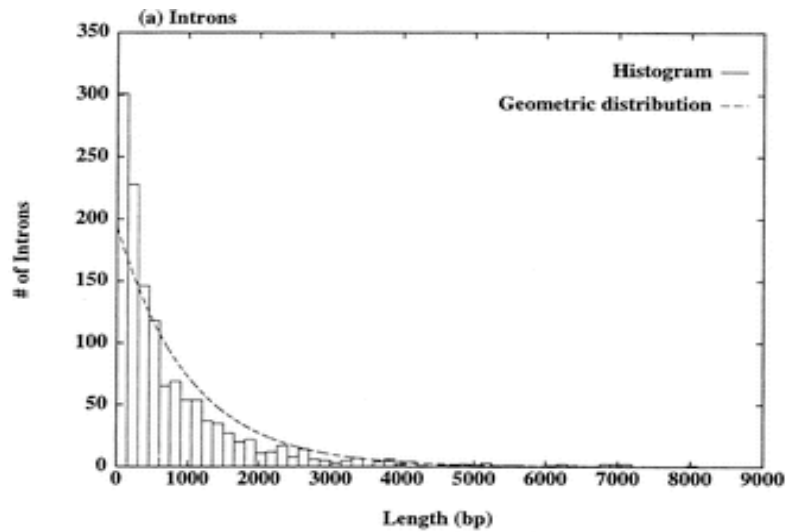
- $\Pr(\text{leaving state}) = p$
- $\Pr(\text{staying in state}) = 1 - p$
- $\Pr(\text{output of exactly } r \text{ in state}) = (1-p)^r p$



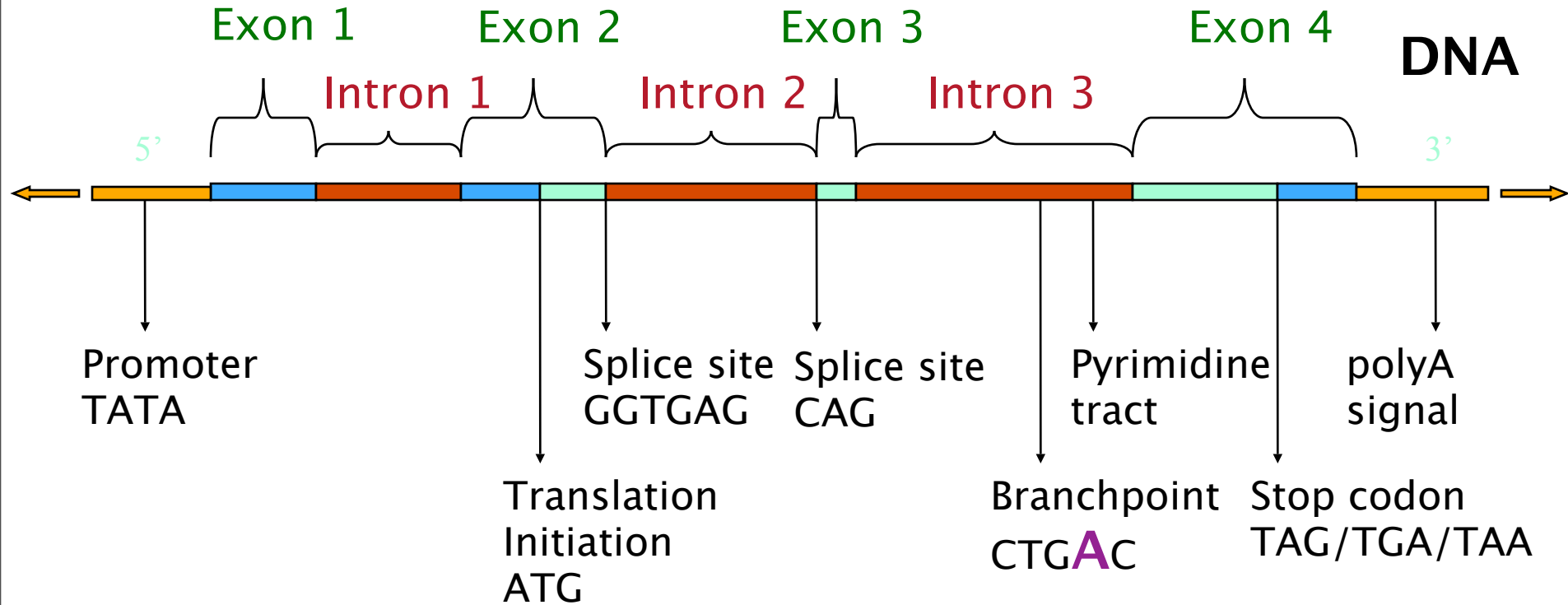
- Geometric distribution

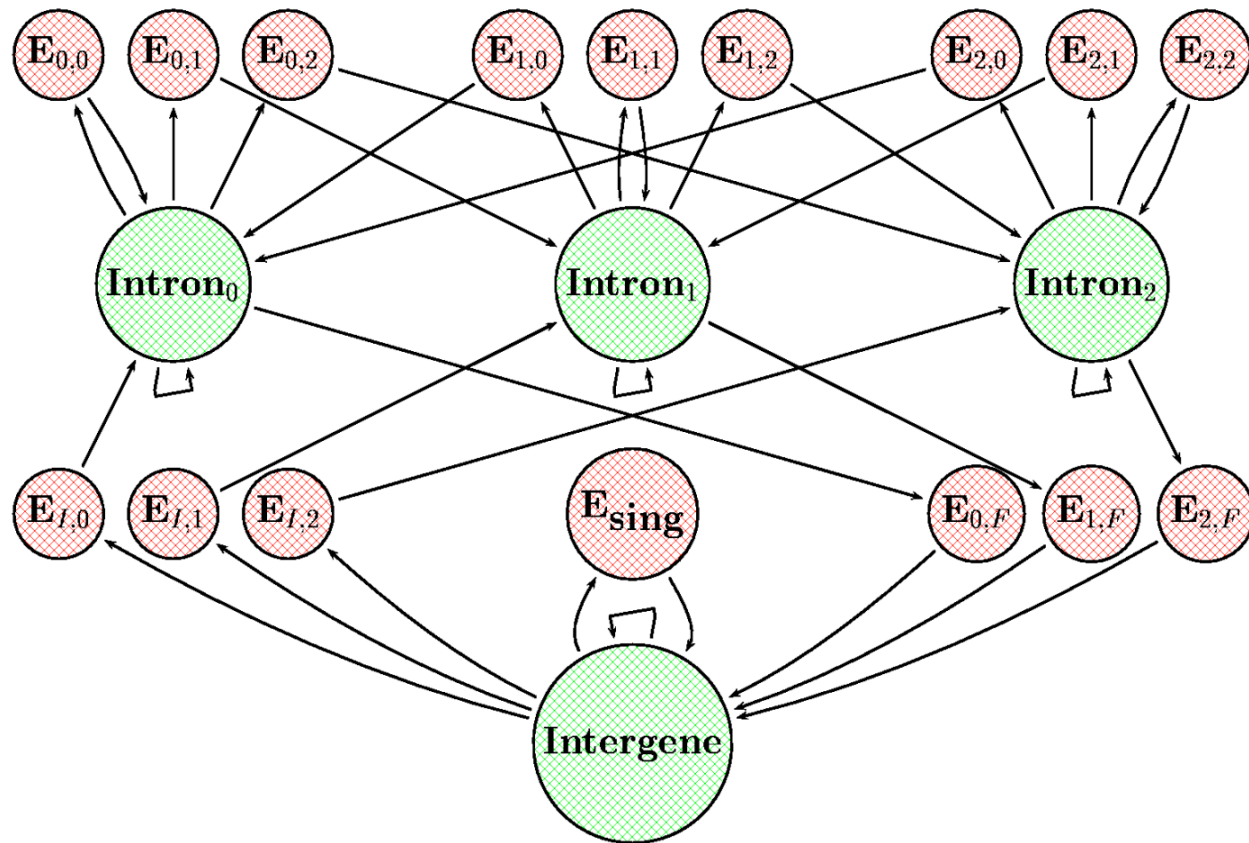


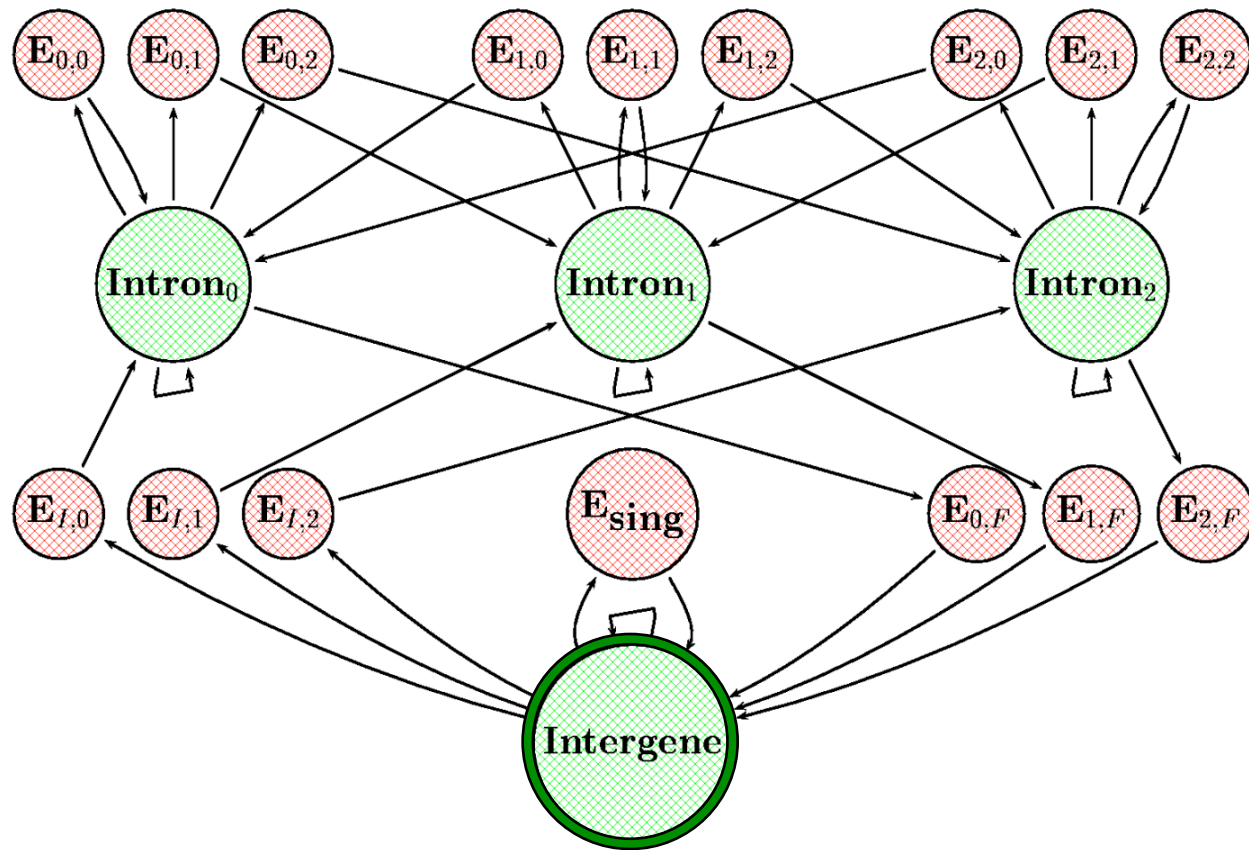
# Observed lengths

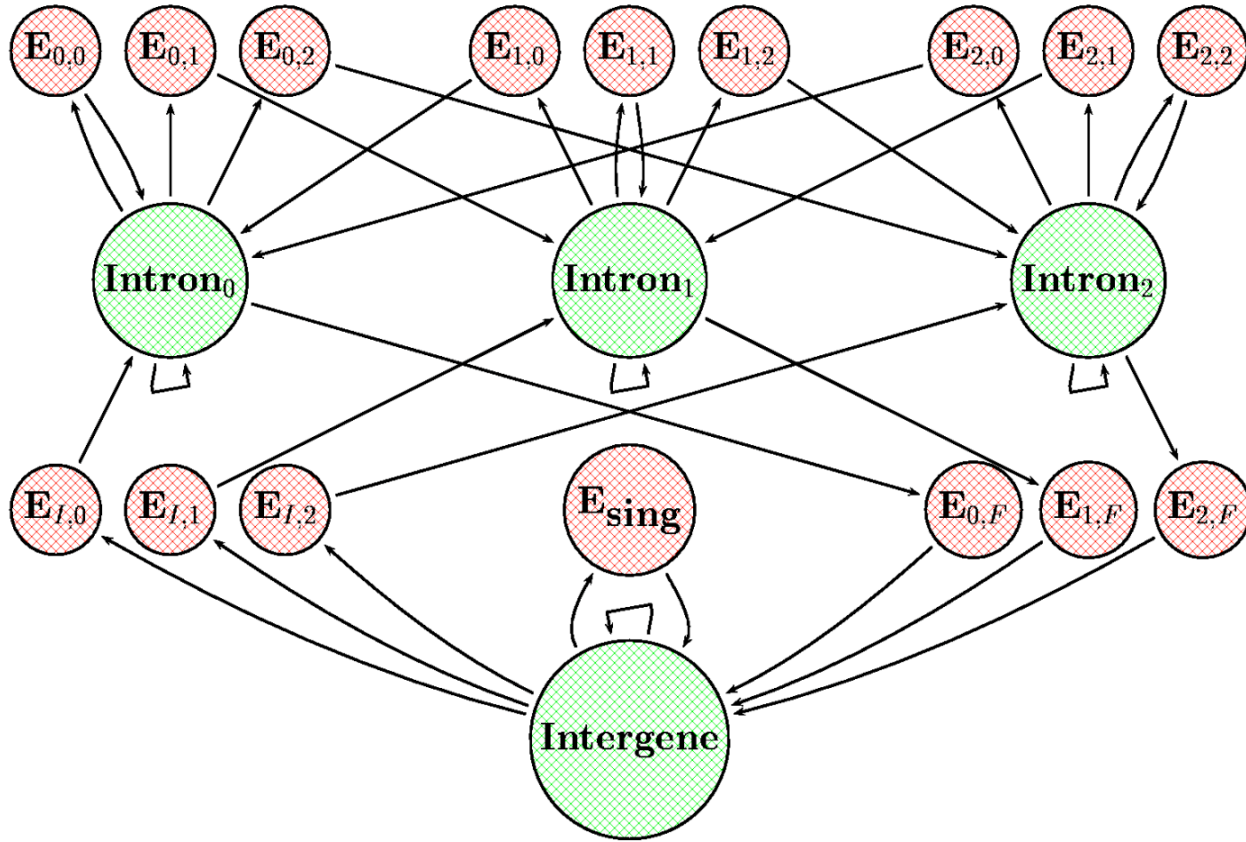


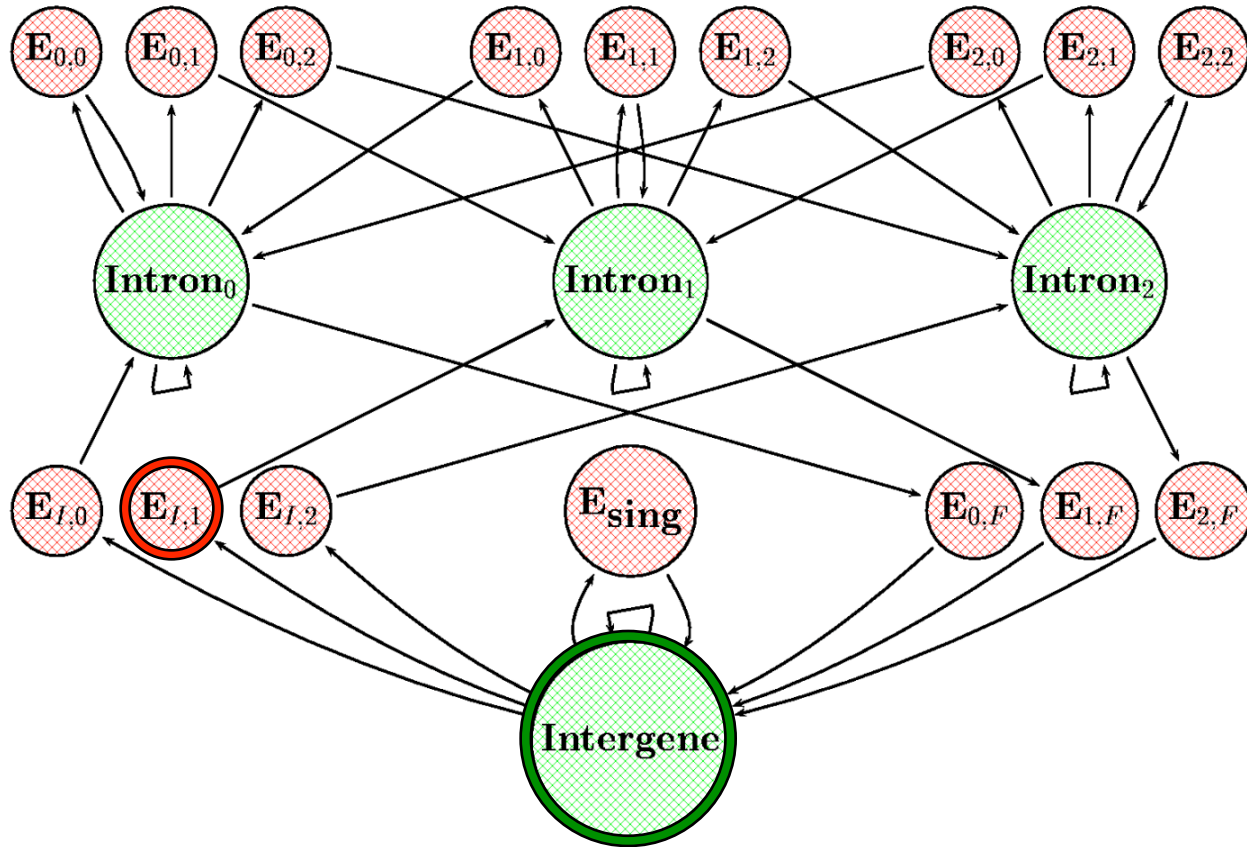
# The Gene Finding Problem



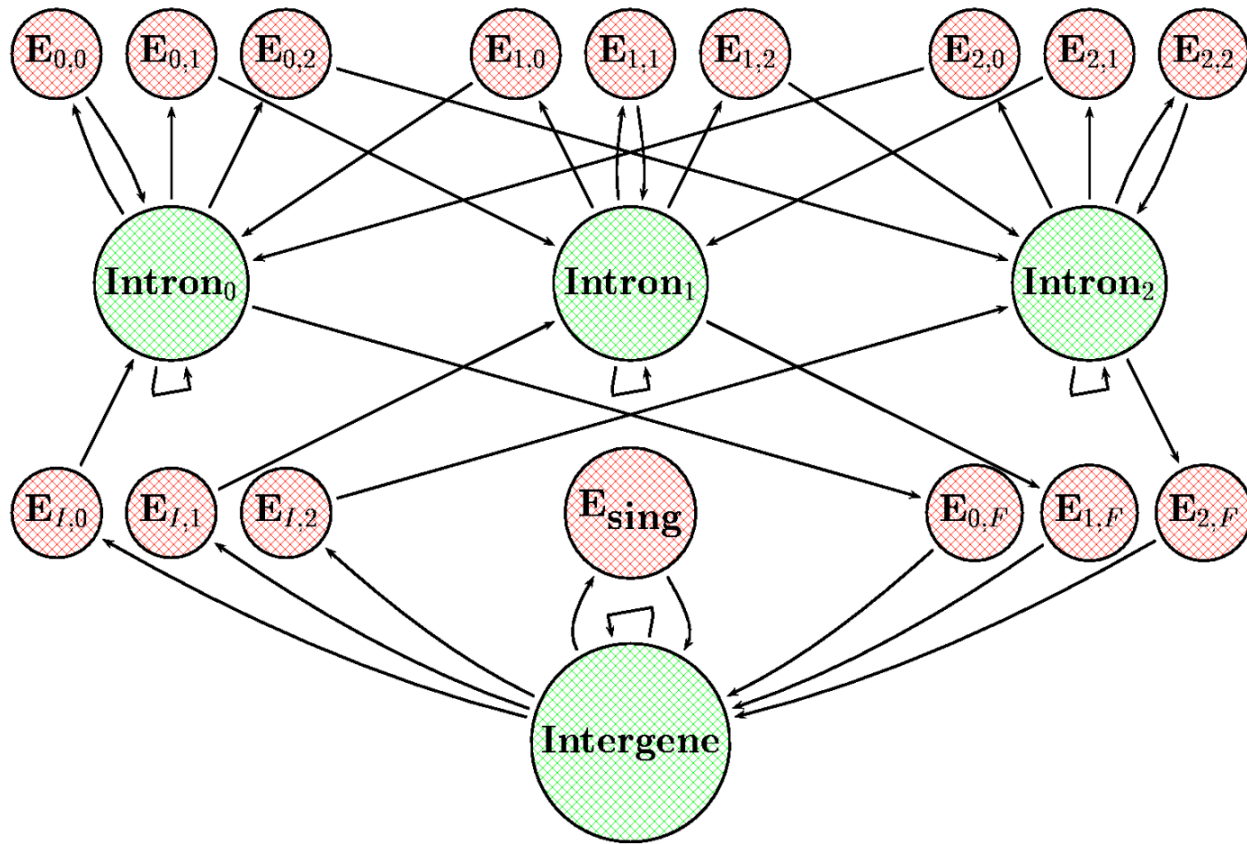




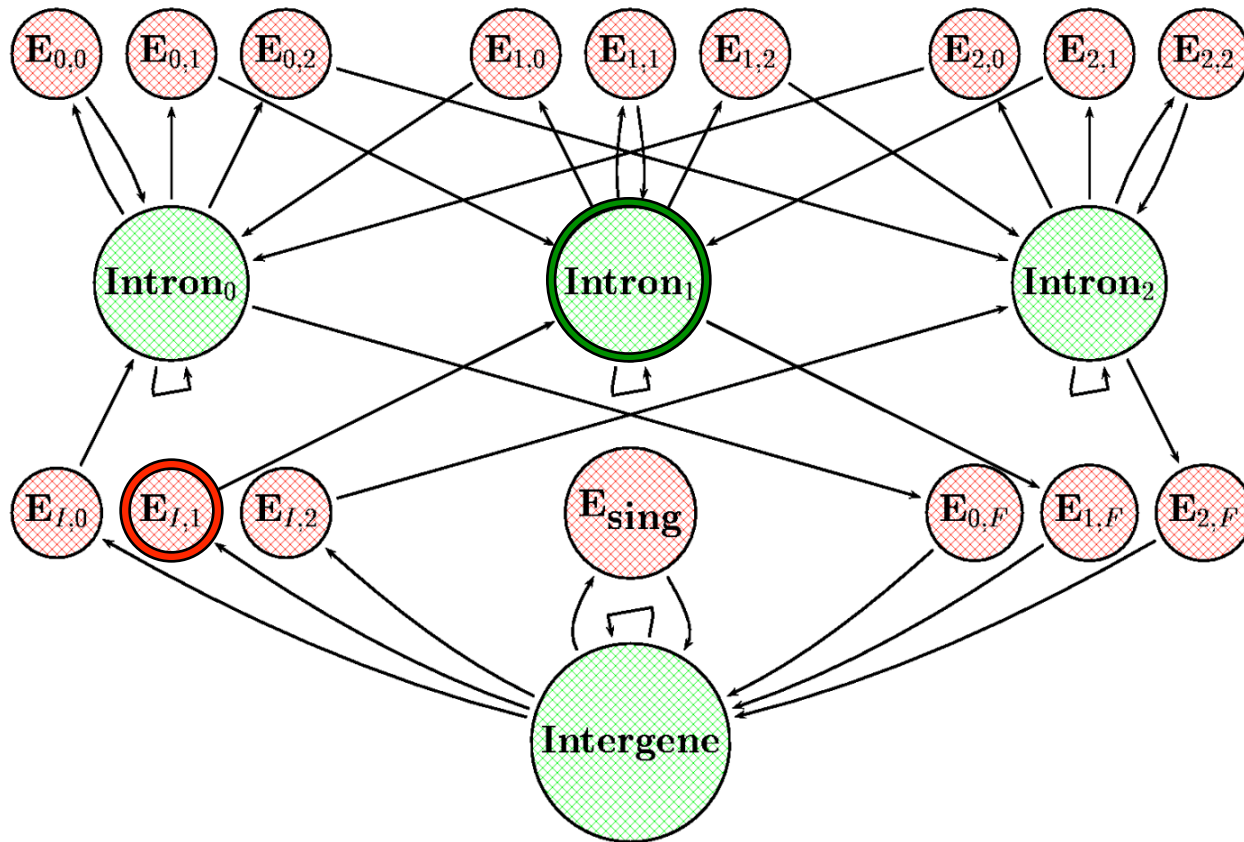




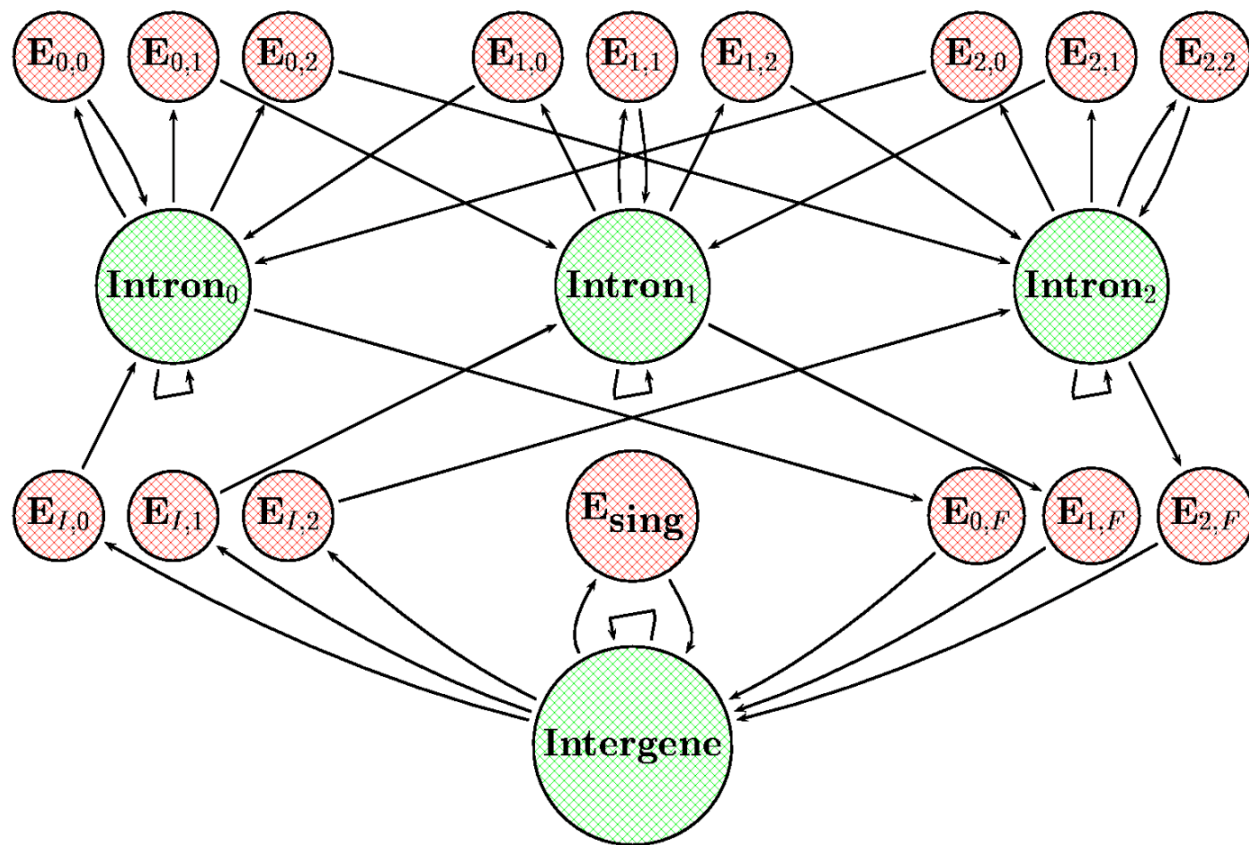
TAAT ATGTCACGG



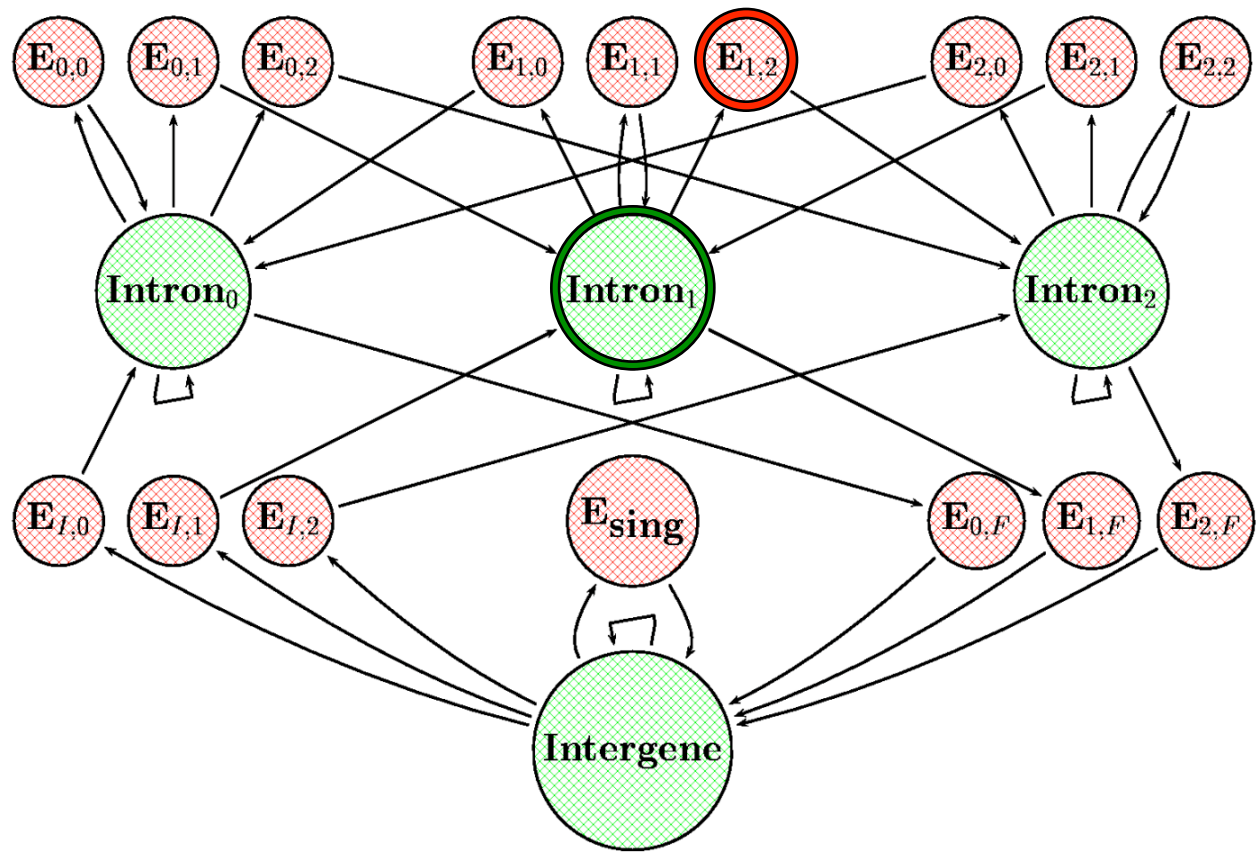
TAAT ATGTCACGG



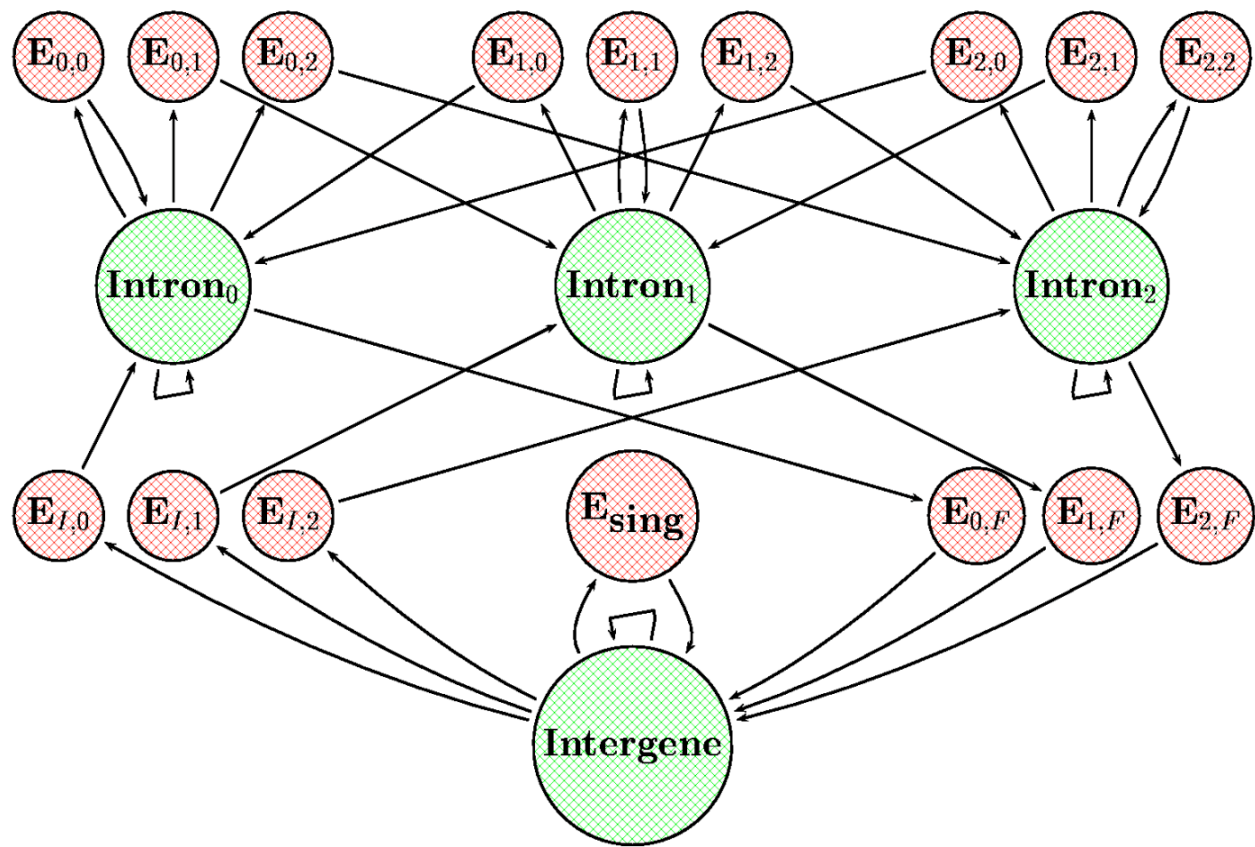
TAAT ATGTCACGG GTATTGAG



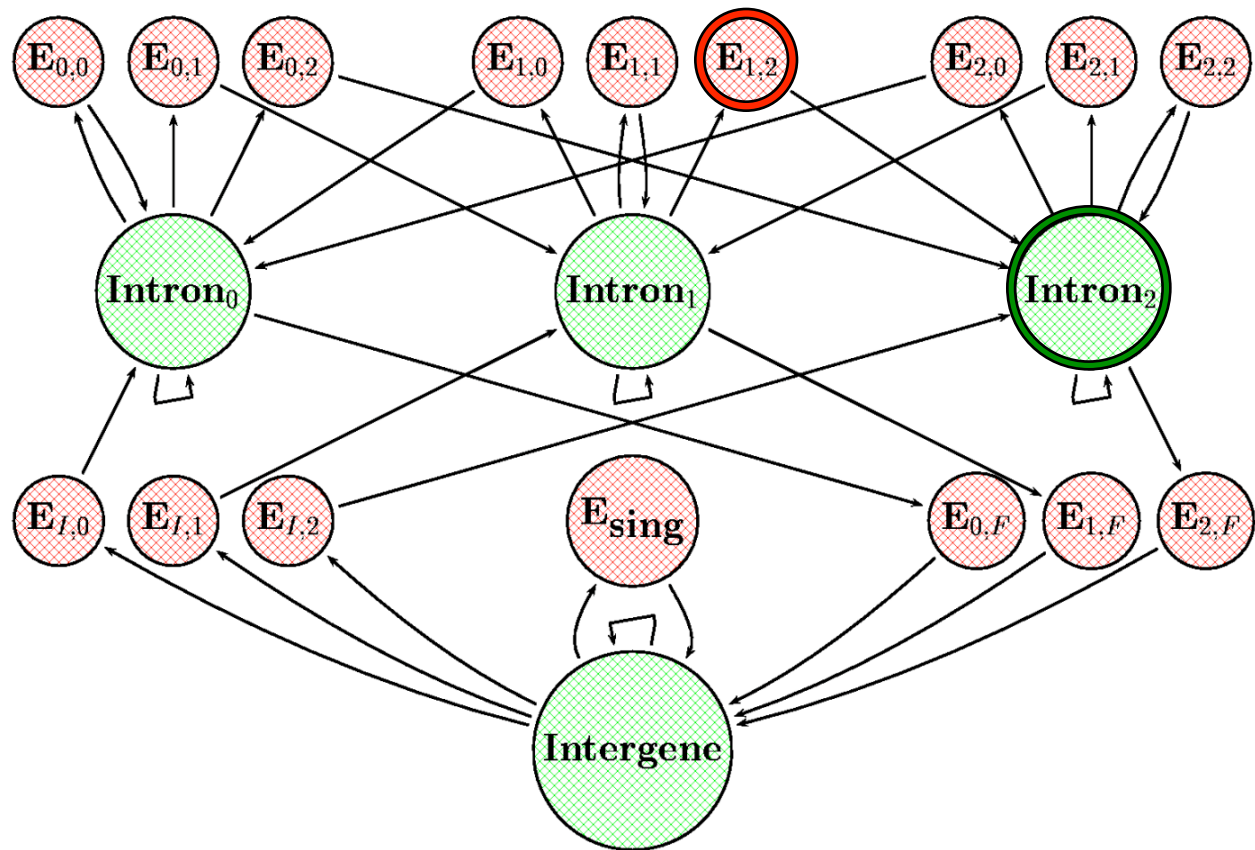
TAAT ATGTCACGG GTATTGAG



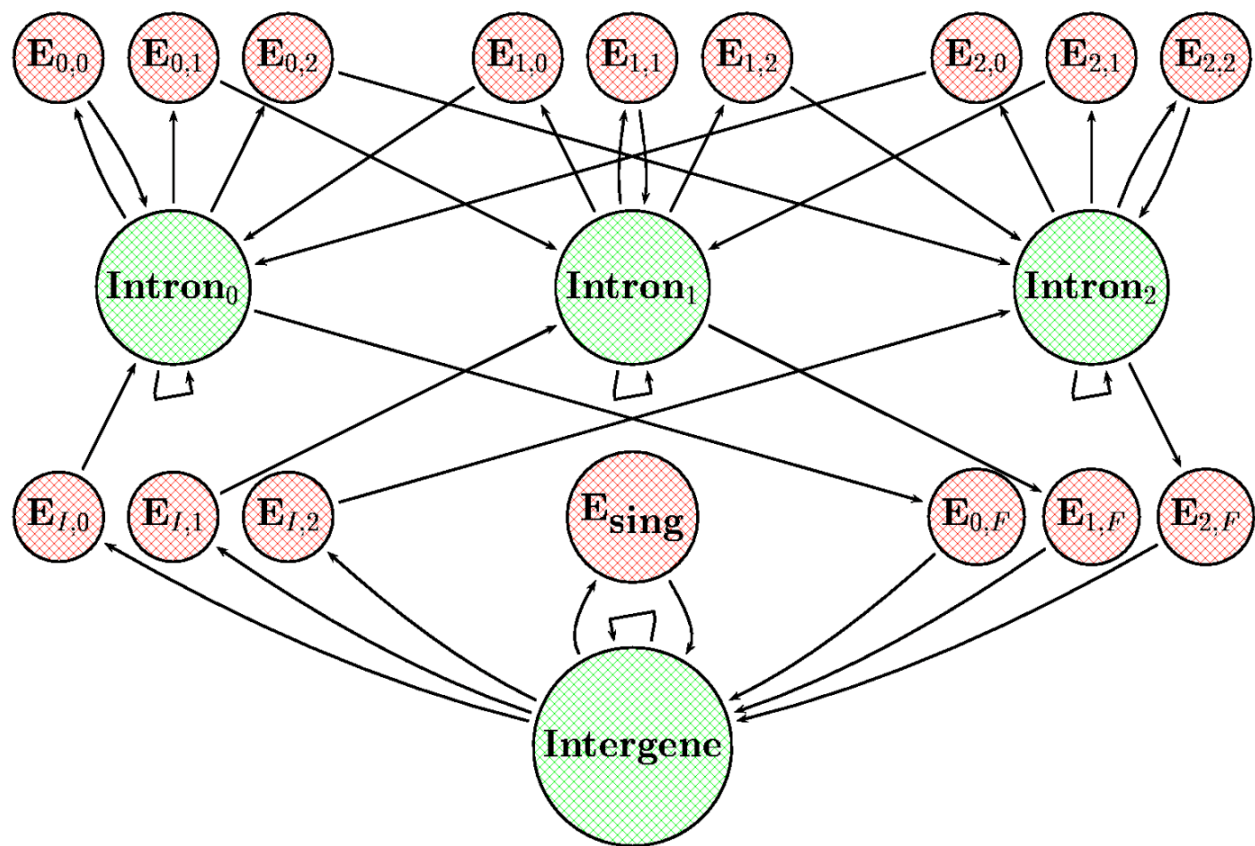
TAAT ATGCCACGG GTATTGAG CATTGTACACGGG



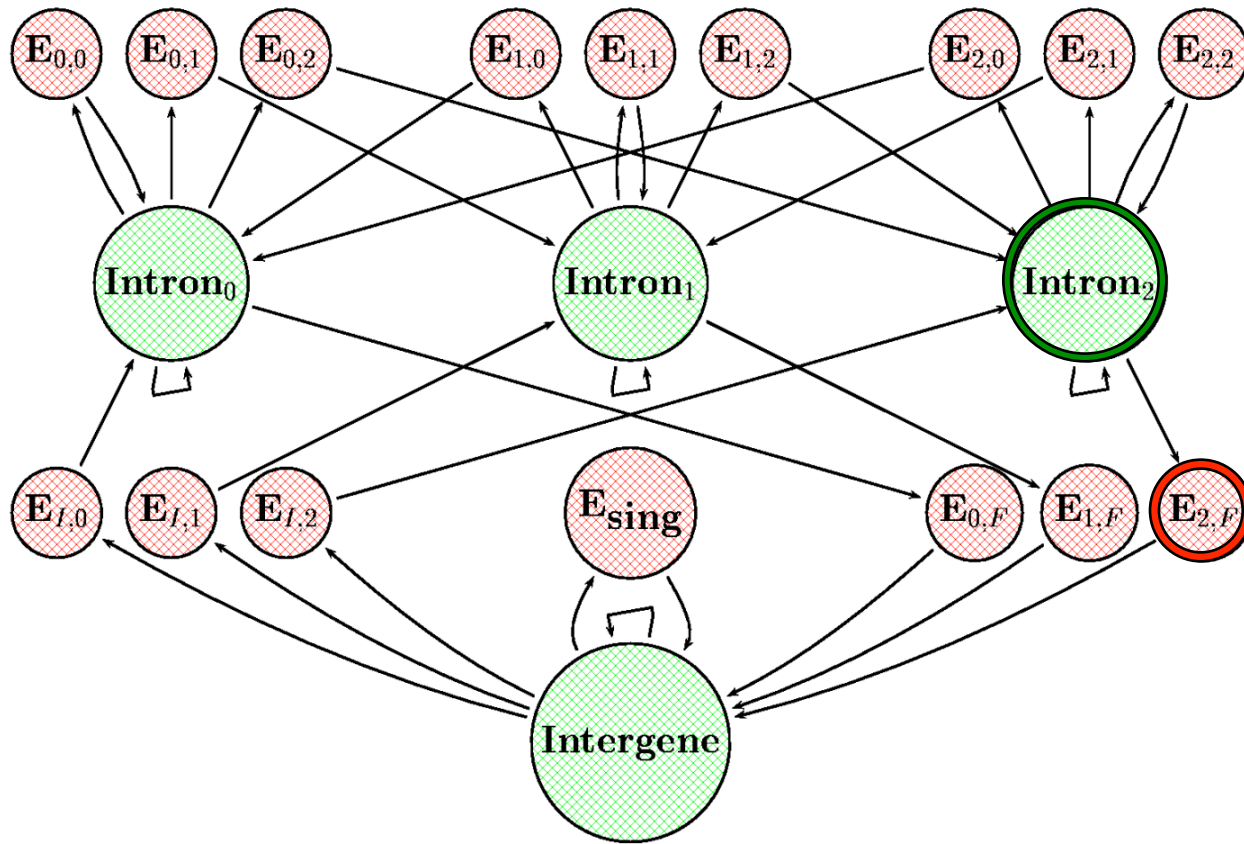
TAAT ATGTCACGG GTATTGAG CATTGTACACGGG



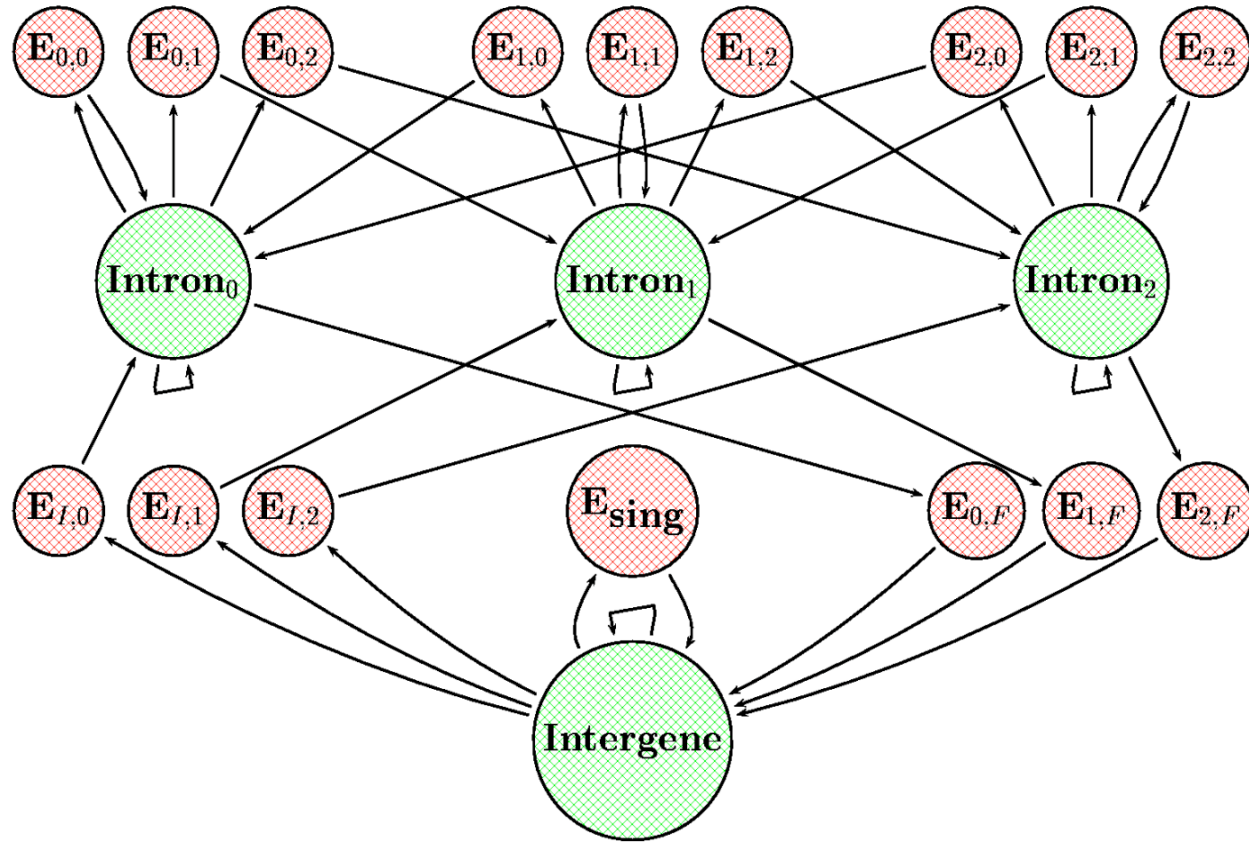
TAAT ATGCCACGG GTATTGAG CATTGTACACGGG GTATTGAG



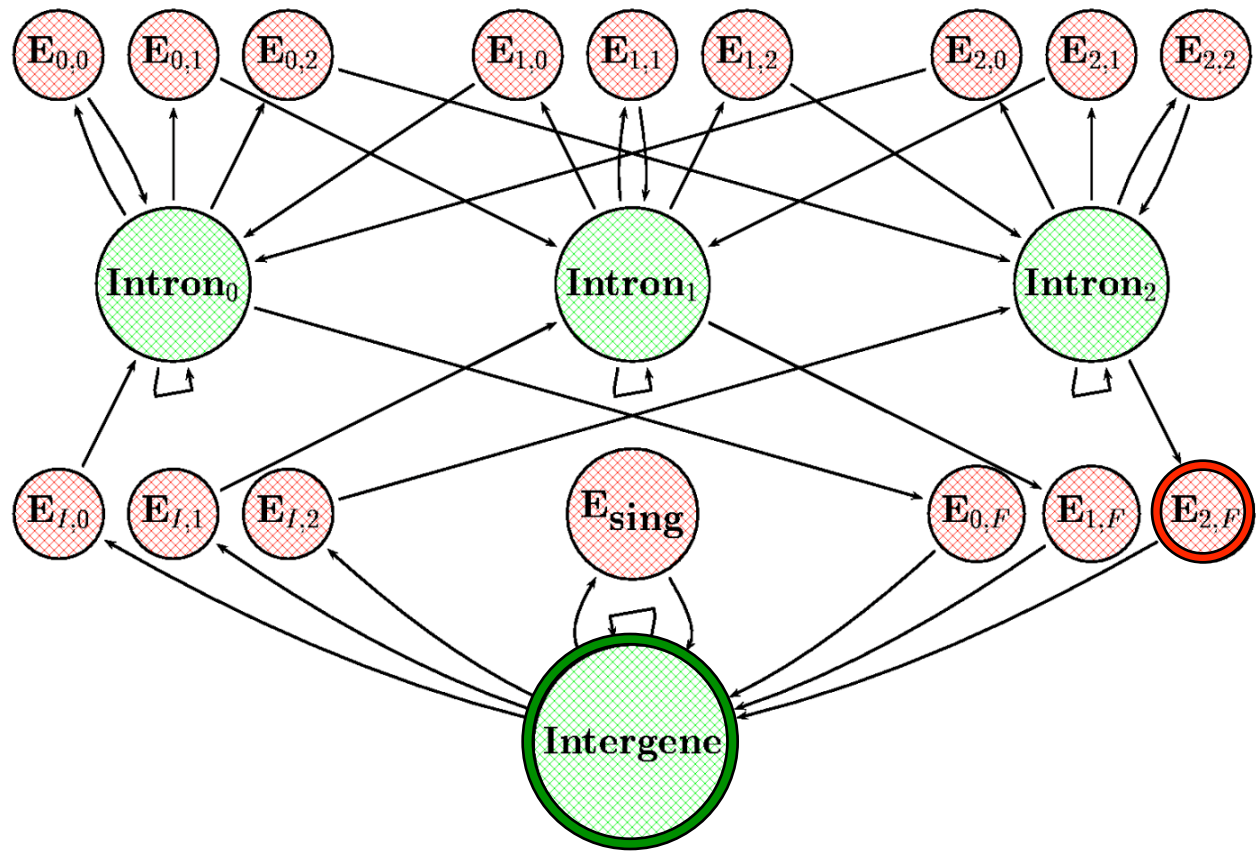
TAAT ATGCCACGG GTATTGAG CATTGTACACGGG GTATTGAG



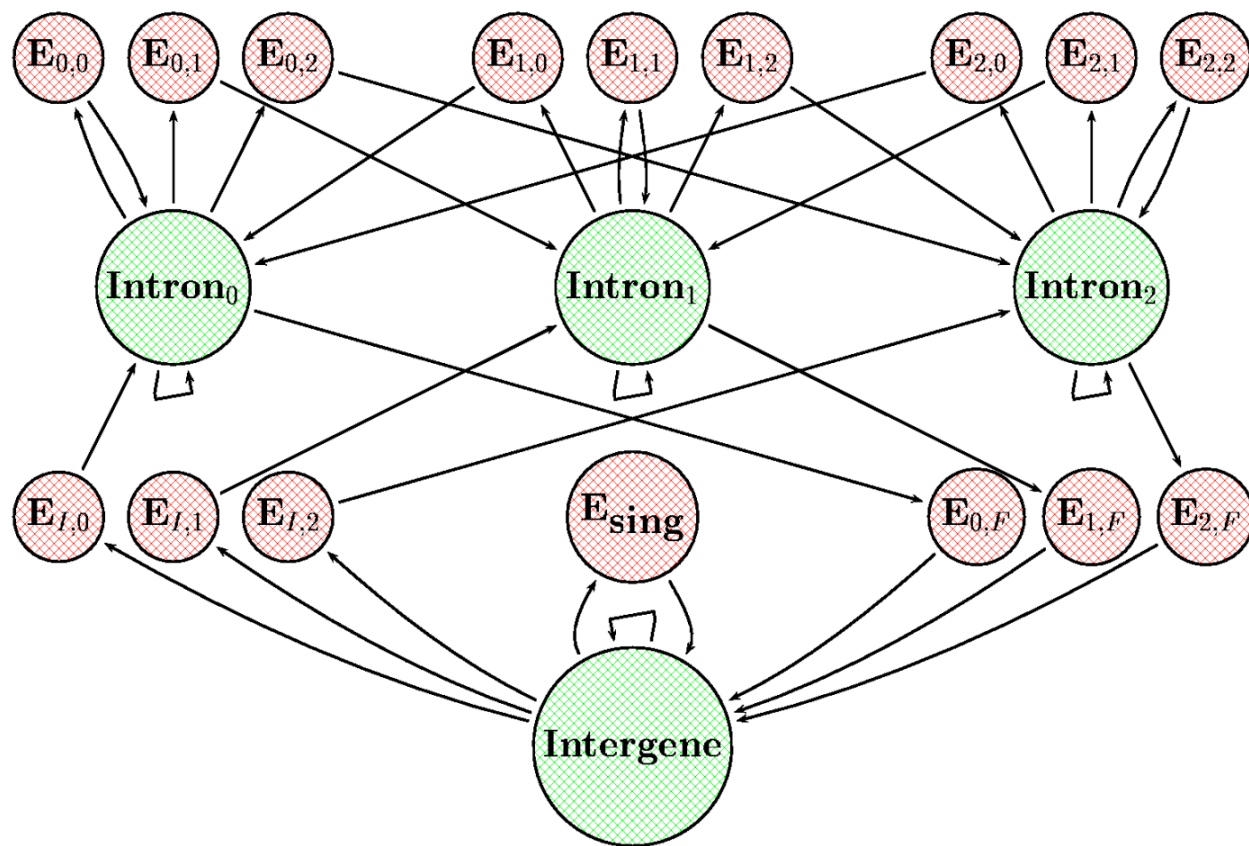
TAAT ATGCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA



TAAT ATGCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA



TAAT ATGCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



Using GHMMs for ab-initio gene

# Using GHMMs for ab-initio gene

In practice, have observed sequence

TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

# Using GHMMs for ab-initio gene

In practice, have observed sequence

TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

Predict genes by estimating hidden state sequence

TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



# Using GHMMs for ab-initio gene

In practice, have observed sequence

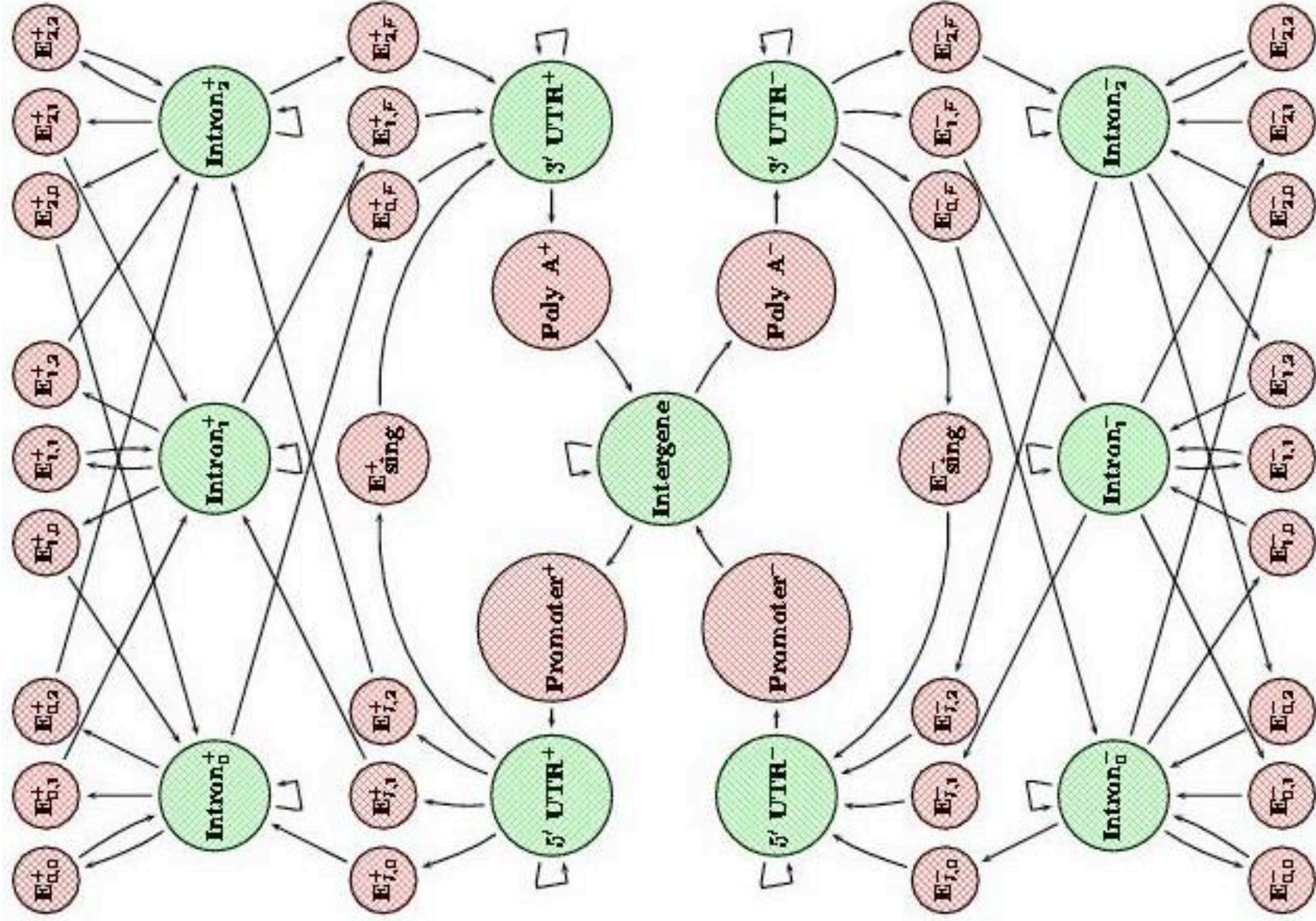
TAATATGTCCACGGGTATTGAGCATTGTACACGGGGTATTGAGCATGTAA TGAA

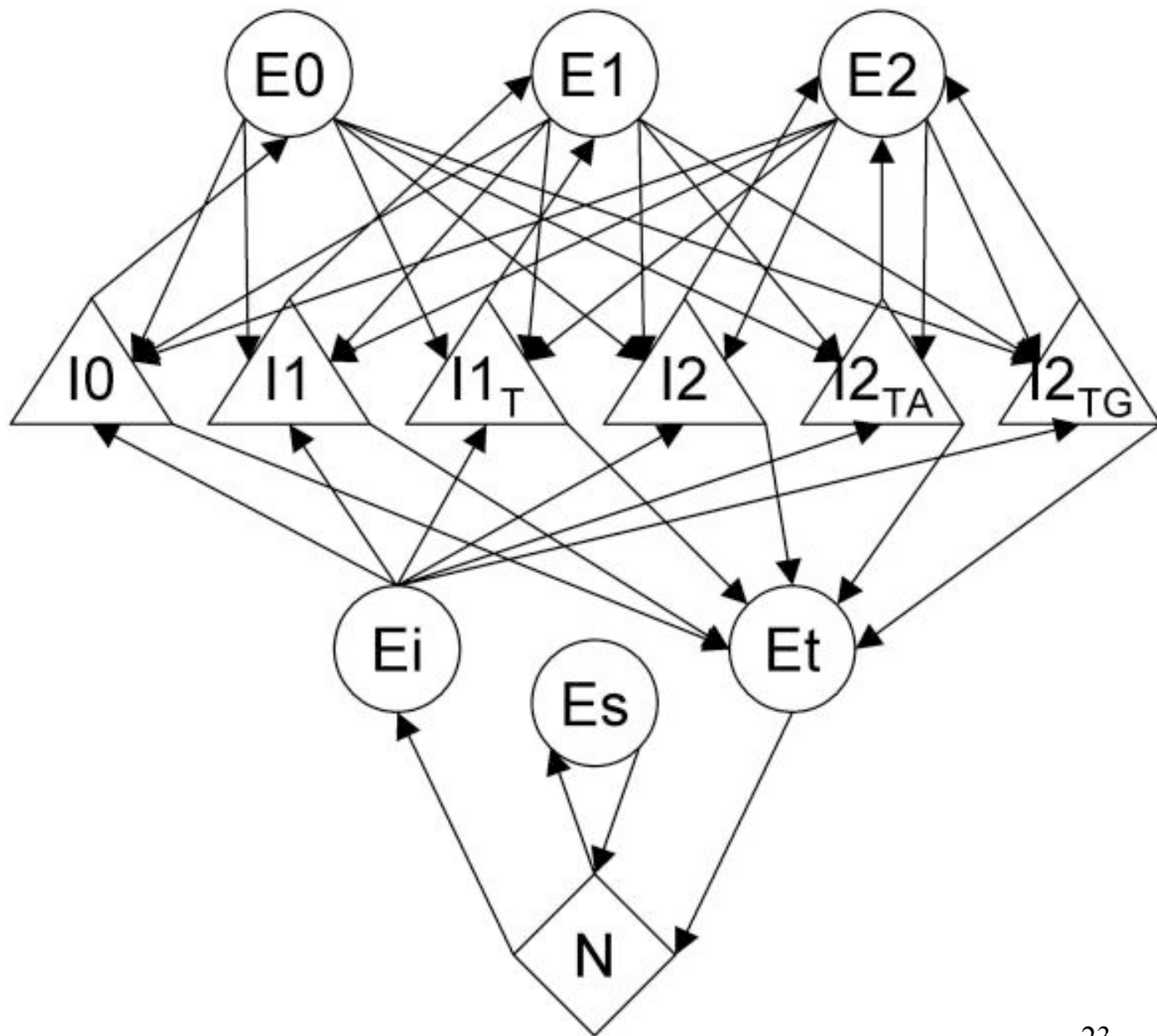
Predict genes by estimating hidden state sequence

TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



Usual solution: single most likely sequence of hidden states (Viterbi).





# Lattice view

