

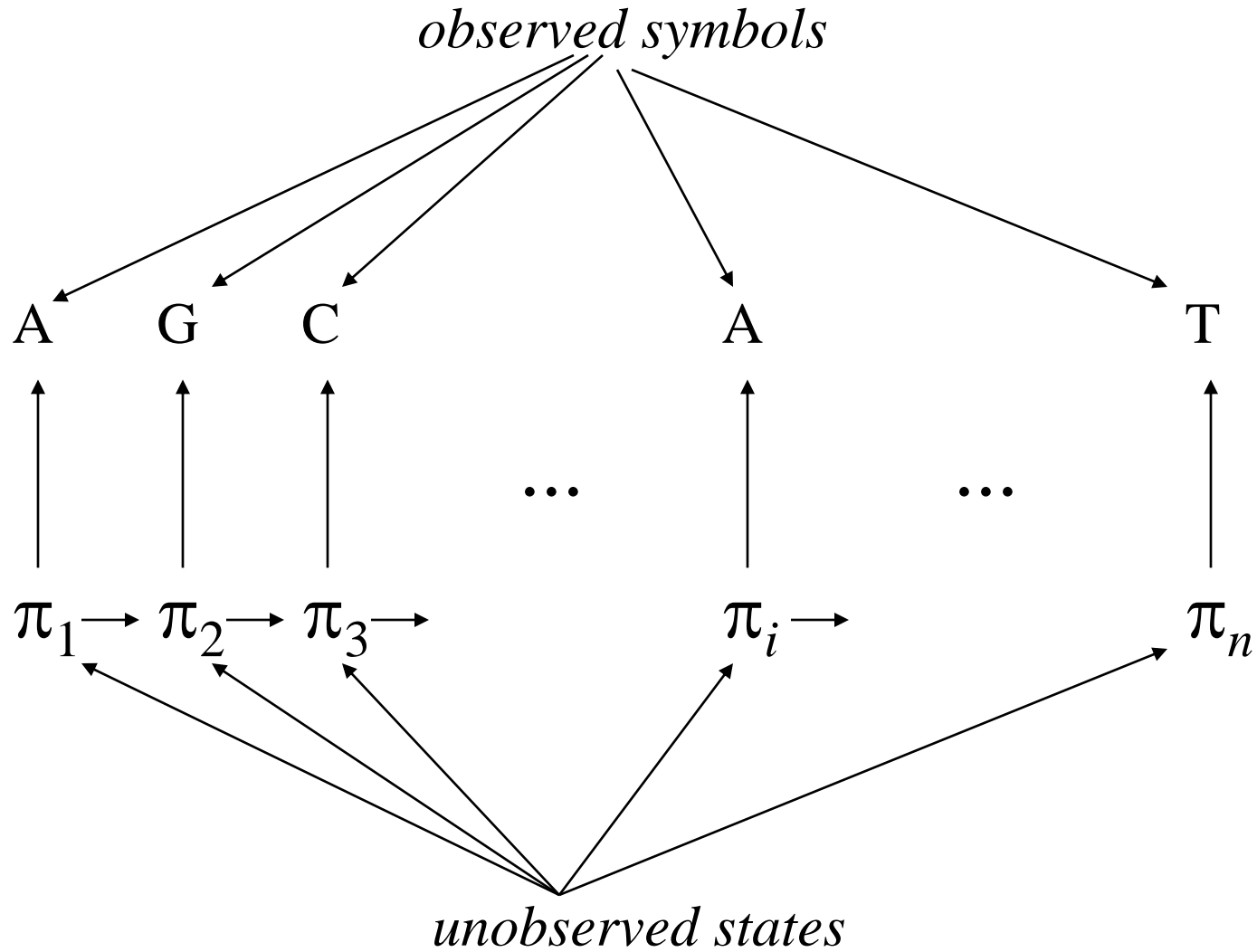
# Lecture 13

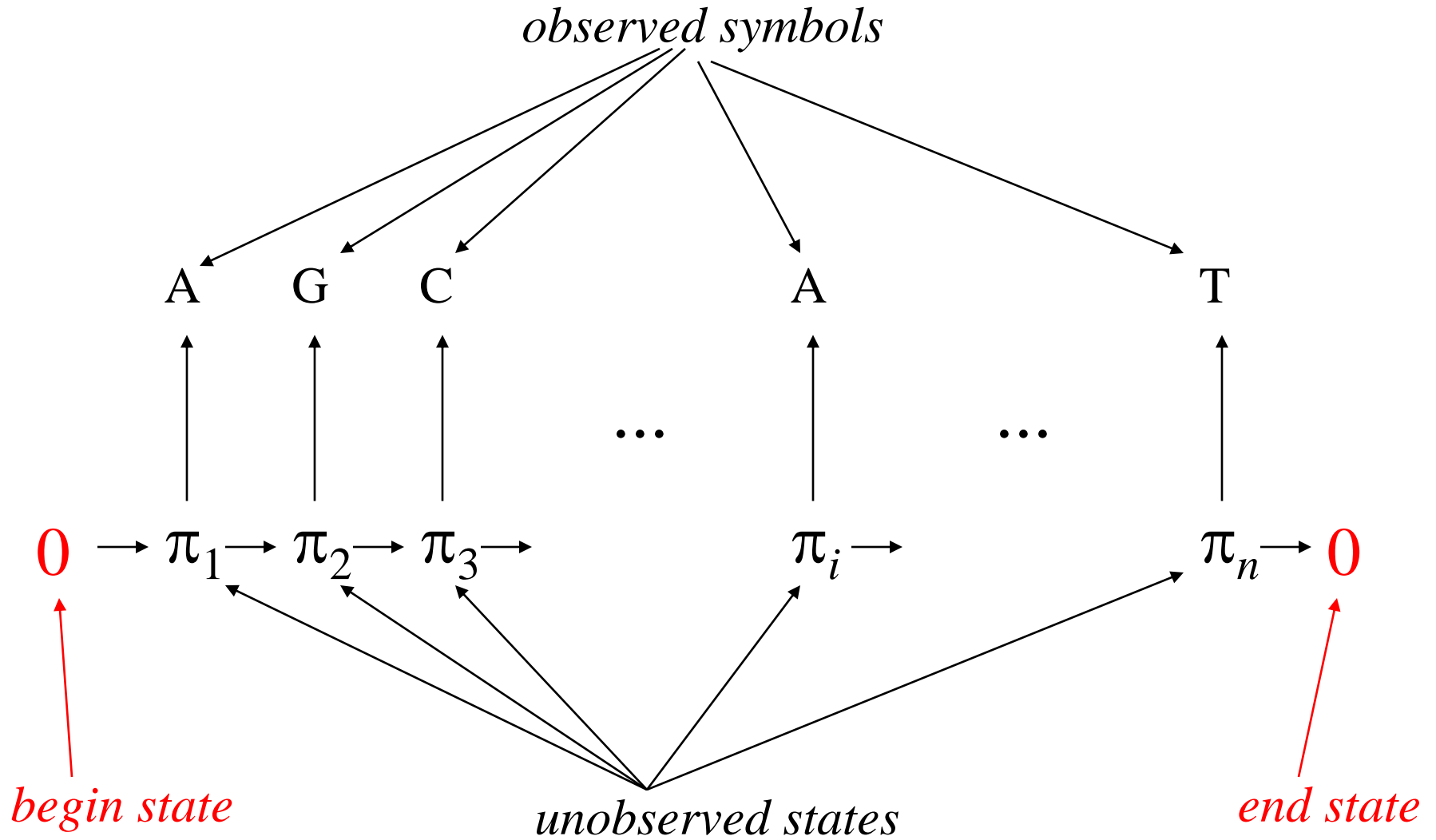
- Hidden Markov Models
  - Intro & Definitions
  - Examples

# Hidden Markov Models

- Probability models for sequences of *observed symbols*, e.g.
  - nucleotide or amino acid residues
  - aligned pairs of residues
  - aligned set of residues corresponding to leaves of an underlying evolutionary tree
  - angles in protein chain (structure modelling)
  - sounds (speech recognition)

- Assume a sequence of “*hidden*” (unobserved) *states* underlies each observed symbol sequence
- Each state “*emits*” symbols (one symbol at a time)
- States may correspond to underlying “reality” we are trying to infer, e.g.
  - unobserved biological feature:
    - (positions within) a site
  - rate of evolution
  - protein structural element
  - speech phoneme





# Advantages of HMMs

- Flexible – gives reasonably good models in wide variety of situations
- Computationally efficient
- Often interpretable:
  - hidden states can correspond to biological features.
  - can find most probable sequence of hidden states  
= biological “parsing” of residue sequence.

# HMMs: Formal Definition

- Alphabet  $\mathcal{B} = \{b\}$  of *observed symbols*
- Set  $\mathcal{S} = \{k\}$  of *hidden states* (usually  $k = 0, 1, 2 \dots, m$ ; 0 is reserved for “begin” state, and sometimes also an “end” state)
- (Markov chain property): prob of state occurring at given position depends only on immediately preceding state, and is given by
  - transition probabilities* ( $a_{kl}$ ):  $a_{kl} = \text{Prob}(\text{next state is } l \mid \text{curr state is } k)$   
 $\sum_l a_{kl} = 1$ , for each  $k$ .
  - Usually, many transition probabilities are set to 0.
  - Model *topology* is the # of states, and *allowed* (i.e.  $a_{kl} \neq 0$ ) transitions.

Sometimes omit begin state, in which case need *initiation probabilities* ( $p_k$ ) for sequence starting in a given state

## *from lecture 3:*

- Conditional probabilities (as on the previous slide) can be used to define a ***first-order Markov model*** (or ***Markov chain model***) for sequence probabilities:

$$\begin{aligned} P(s_1 s_2 s_3 \cdots s_n) \\ \equiv P(s_1) P(s_2 / s_1) P(s_3 / s_2) \cdots P(s_n / s_{n-1}) \end{aligned}$$



*observed symbols*

A

G

C

A

T

$e_{\pi_1}(A)$

$e_{\pi_2}(G)$

$e_{\pi_3}(C)$

...

$e_{\pi_i}(A)$

...

$e_{\pi_n}(T)$

$0 \xrightarrow{a_{0\pi_1}} \pi_1 \xrightarrow{a_{\pi_1\pi_2}} \pi_2 \xrightarrow{a_{\pi_2\pi_3}} \pi_3 \xrightarrow{a_{\pi_3\pi_4}}$

$\pi_i \xrightarrow{a_{\pi_i\pi_{i+1}}}$

$\pi_n \rightarrow 0$

*unobserved states*

- Prob that symbol occurs at given sequence position depends only on hidden state at that position, and is given by

*emission probabilities:*

$$e_k(b) = \text{Prob}(\text{observed symbol is } b \mid \text{curr state is } k)$$

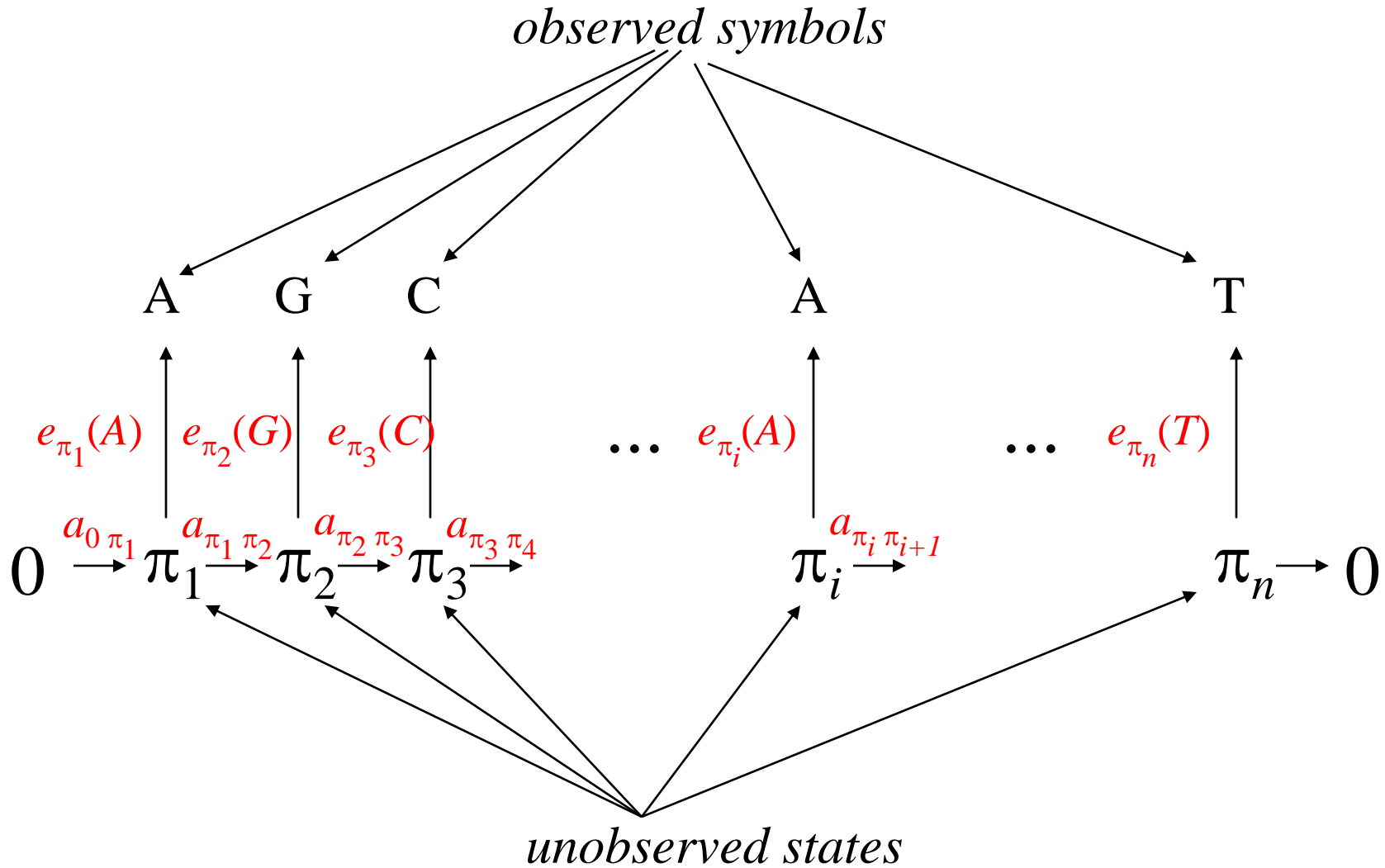
(begin and end states do not emit symbols)

- Note that
  - there are no *direct* dependencies between observed symbols in the sequence, however
  - there are *indirect* dependencies implied by state dependencies

# Where do the parameters come from?

- Can either
  - *define* parameter values *a priori*, or
  - *estimate* them from training data (observed sequences of the type to be modelled).
- Usually one does a mixture of both –
  - model topology is defined (some transitions set to 0), but
  - remaining parameters estimated

# Hidden Markov Model



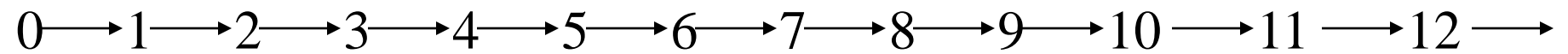
# HMM examples: 1-state HMMs

- single state, emitting residues with specified freqs:  
= ‘background’ model

# HMM examples: site models

- “states” correspond to positions (columns in the tables). state  $i$  transitions only to state  $i+1$ :
  - $a_{i,i+1} = 1$  for all  $i$ ;
  - all other  $a_{ij}$  are 0
- emission probabilities are position-specific frequencies: values in frequency table columns

# Topology for Site HMM: 'allowed' transitions



# HMM for *C. elegans* 3' Splice Sites



A	3276	3516	2313	476	67	757	240	8192	0	3359	2401	2514
C	970	648	664	236	129	1109	6830	0	0	1277	1533	1847
G	593	575	516	144	39	595	12	0	8192	2539	1301	1567
T	3353	3453	4699	7336	7957	5731	1110	0	0	1017	2957	2264

**CONSENSUS** W W W T T t C A G r w w

Emission probabilities	A	0.400	0.429	0.282	0.058	0.008	0.092	0.029	1.000	0.000	0.410	0.293	0.307
	C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
	G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
	T	0.409	0.422	0.574	0.896	0.971	0.700	0.135	0.000	0.000	0.124	0.361	0.276

0 → 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 11 → 12

'hidden' states

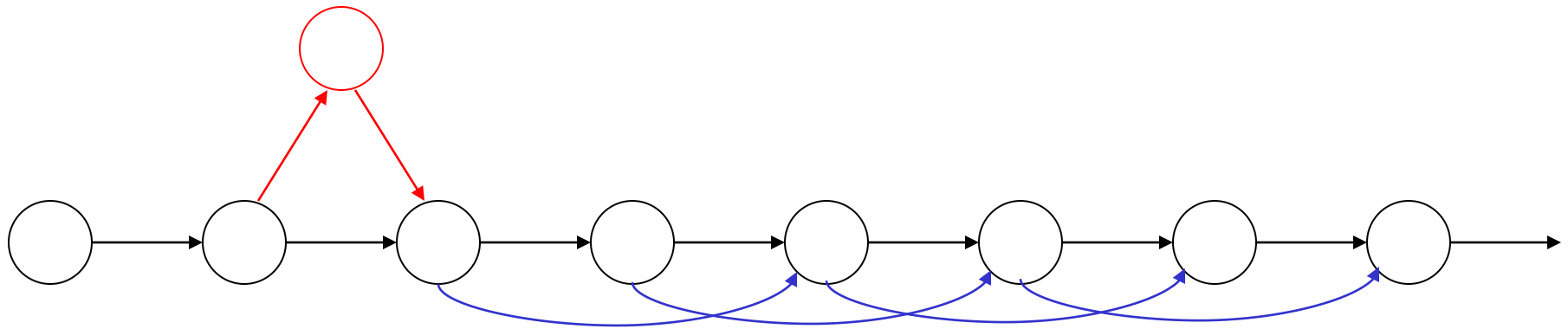


# Expanded site models

- Can expand site models to allow omission of nuc at some positions by including other (downstream) transitions (or via “silent states”)
- Can allow insertions by including additional states.
- transition probabilities then no longer necessarily 1 or 0

# Insertions & Deletions in Site Model

insertion state



other transitions correspond  
to deletions

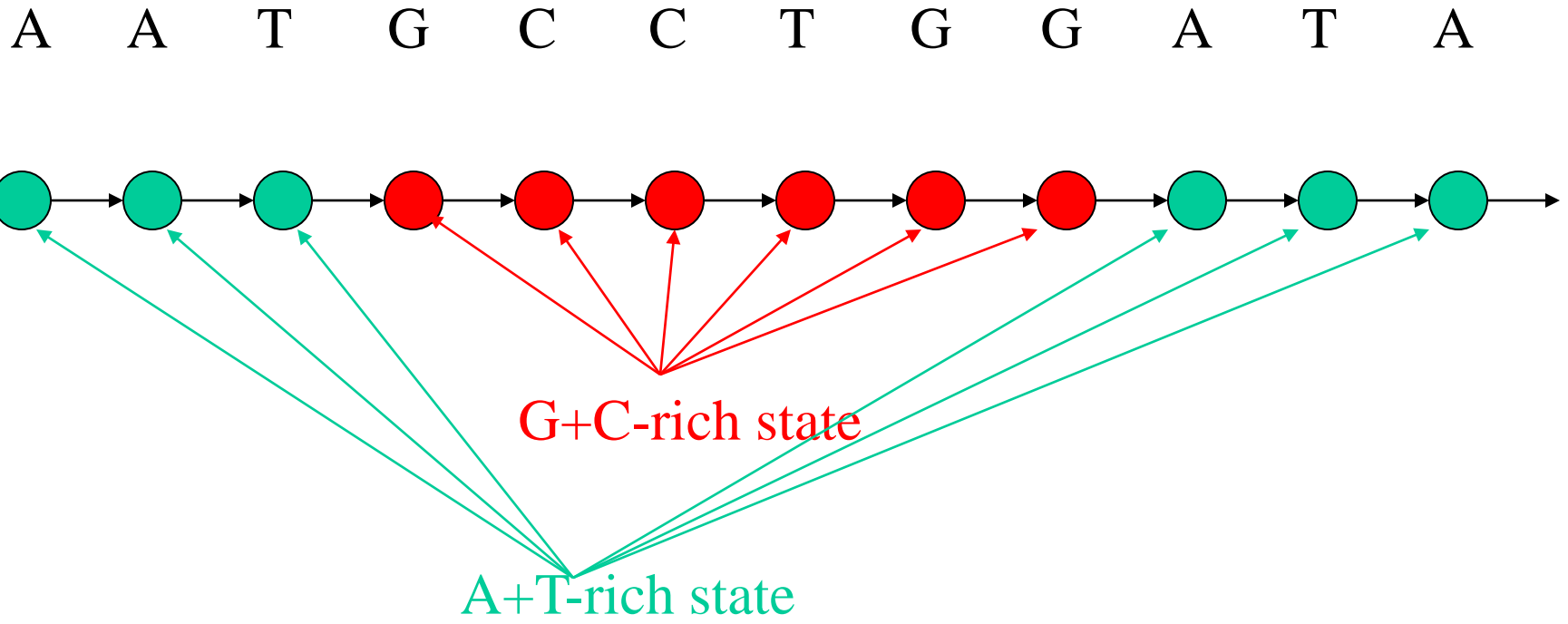
# HMM examples (in Durbin *et al.*)

- protein families (like site models – but important to allow insertions & deletions)
- Pair HMMs
- protein structure (symbols emitted are structural elements)

# HMM examples: 2-state HMMs

- if  $a_{11}$  and  $a_{22}$  are small (close to 0), and  $a_{12}$  and  $a_{21}$  are large (close to 1), then get (nearly) periodic model with period 2; e.g.
  - dinucleotide repeat in DNA, or
  - (some) beta strands in proteins.
- if  $a_{11}$  and  $a_{22}$  large, and  $a_{12}$  and  $a_{21}$  small, then get models of alternating regions of different compositions (specified by emission probabilities), e.g.
  - higher vs. lower G+C content regions (RNA genes in thermophilic bacteria); or
  - hydrophobic vs. hydrophilic regions of proteins (e.g. transmembrane domains).

Closely related to D-segment method (lecture 12)!



# HMM examples: Markov models

- Ordinary Markov chain model:
  - states = observed symbols
  - emission probs = 1 or 0
  - transition probs = prob of observing a symbol, given the preceding one.
- Order  $k$  Markov model
  - states = length  $k$  words (e.g.  $b_1b_2 \dots b_k$ )
  - (unique) symbol emitted by  $b_1b_2 \dots b_k$  is  $b_k$
  - transition prob from  $b_1b_2 \dots b_k$  to  $c_1c_2 \dots c_k$  is non-zero only if
    - $c_1c_2 \dots c_{k-1} = b_2b_3 \dots b_k$ , in which case it is  $P(b_{k+1}|b_1b_2 \dots b_k)$  where  $b_{k+1} = c_k$

## *from lecture 3:*

- Similarly, one can define an a ***order-k Markov model*** in which the probability of  $s_i$  is conditional on  $s_{i-k} \dots s_{i-2} s_{i-1}$  (i.e. the  $k$  preceding residues)
- Note that the required number of parameters is exponential in  $k$
- The ***independence model*** (which is usually good enough for us!) = the ***order-0 Markov model***