Today’s Lecture

- Dinucleotides in human genome
- Hidden Markov Models
  - Intro & Definitions
  - Examples

Dinucleotide Freqs – *H. sapiens* Chr.21

Nucleotide Freqs:

A	10032226	0.297;	**T**	9962530	0.295
G	6908202	0.204;	**C**	6921020	0.205

Entropy: 1.976 bits

Observed Dinuc Freqs

	A	**C**	**G**	**T**
A	0.099	0.051	0.069	0.078
C	0.073	0.052	0.011	0.069
G	0.059	0.043	0.052	0.050
T	0.066	0.059	0.072	0.098

Expected (*under independence*)

A	**C**	**G**	**T**
0.088	0.061	0.061	0.087
0.061	0.042	0.042	0.060
0.061	0.042	0.042	0.060
0.087	0.060	0.060	0.087

Observed / Expected

	A	**C**	**G**	**T**
A	1.124	0.839	1.139	0.891
C	1.204	1.243	0.260	1.139
G	0.974	1.025	1.245	0.839
T	0.752	0.976	1.204	1.125

Dinucleotide Freqs – *H. sapiens* Chr.22

```
Nucleotide Freqs:
   A  8745910  0.261; T  8720493  0.261
   G  7999585  0.239; C  7997931  0.239
 Entropy: 1.999 bits
```

Observed Dinuc Freqs

	A	C	G	T
A	0.077	0.051	0.075	0.058
C	0.077	0.071	0.016	0.075
G	0.061	0.057	0.071	0.051
T	0.047	0.061	0.077	0.076

Expected (*under independence*)

A	C	G	T
0.068	0.062	0.062	0.068
0.062	0.057	0.057	0.062
0.062	0.057	0.057	0.062
0.068	0.062	0.062	0.068

Observed / Expected

	A	C	G	T
A	1.125	0.817	1.205	0.855
C	1.233	1.236	0.285	1.206
G	0.975	0.989	1.237	0.818
T	0.684	0.977	1.233	1.124

Failure of independence for ‘background’

Nucleotide Freqs (*C. elegans* chr. 1):
A 4575132 (.321) ; C 2559048 (.179) ; G 2555862 (.179); T 4582688 (.321)

dinucleotide frequencies (5′ nuc to left, 3′ nuc at top – e.g. obs freq of *ApC* is .047): (Note “symmetry”!)

Observed

	A	C	G	T
A	0.135	0.047	0.051	0.088
C	0.061	0.035	0.033	0.051
G	0.063	0.034	0.034	0.047
T	0.061	0.064	0.061	0.135

Expected (*under independence*)

	A	C	G	T
A	0.103	0.057	0.057	0.103
C	0.057	0.032	0.032	0.058
G	0.057	0.032	0.032	0.057
T	0.103	0.058	0.057	0.103

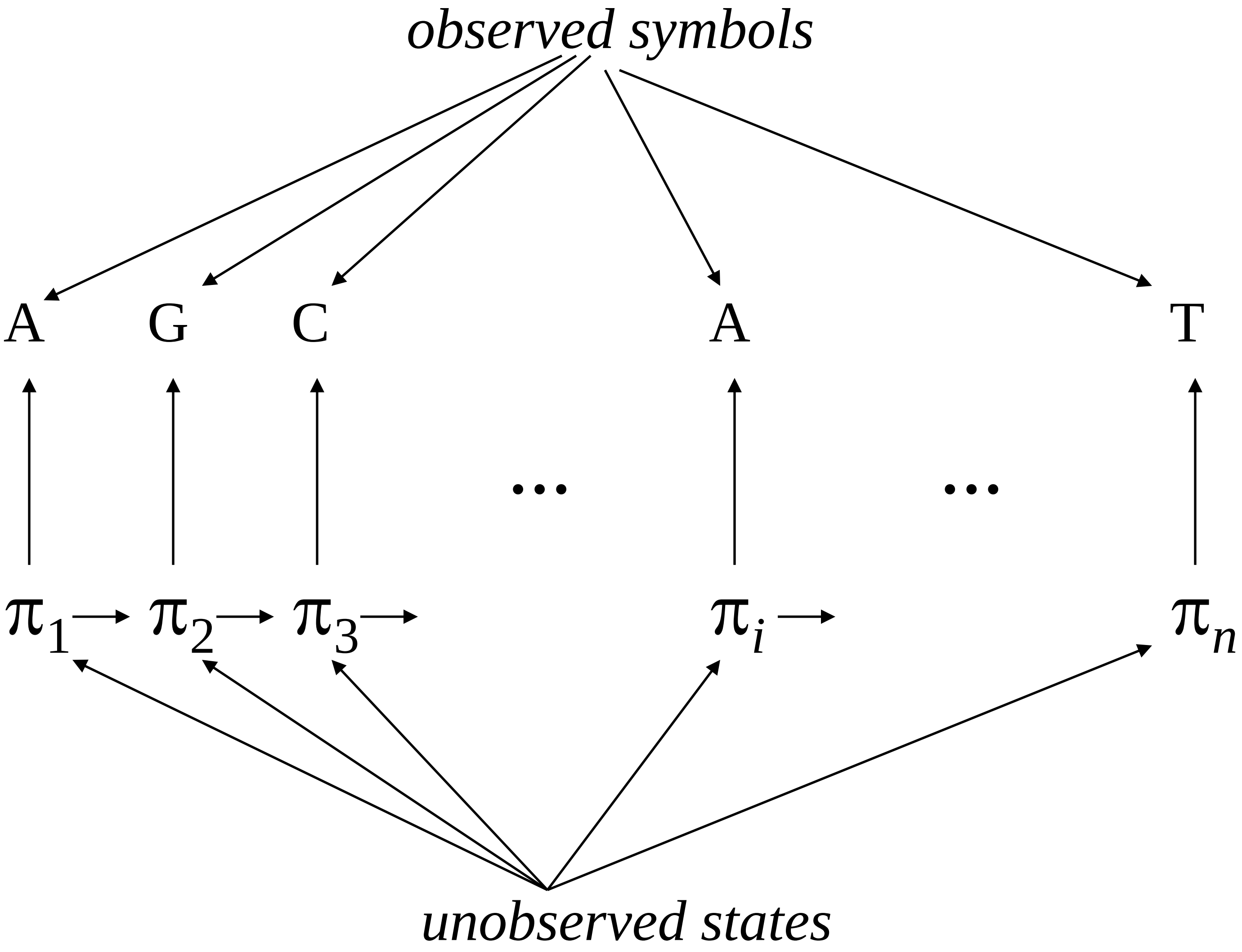
Observed / Expected

	A	C	G	T
A	1.314	0.818	0.885	0.853
C	1.055	1.075	1.031	0.886
G	1.106	1.062	1.074	0.818
T	0.597	1.105	1.056	1.313

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) site,
    - coding region of gene
  - rate of evolution
  - protein structural element
  - speech phoneme

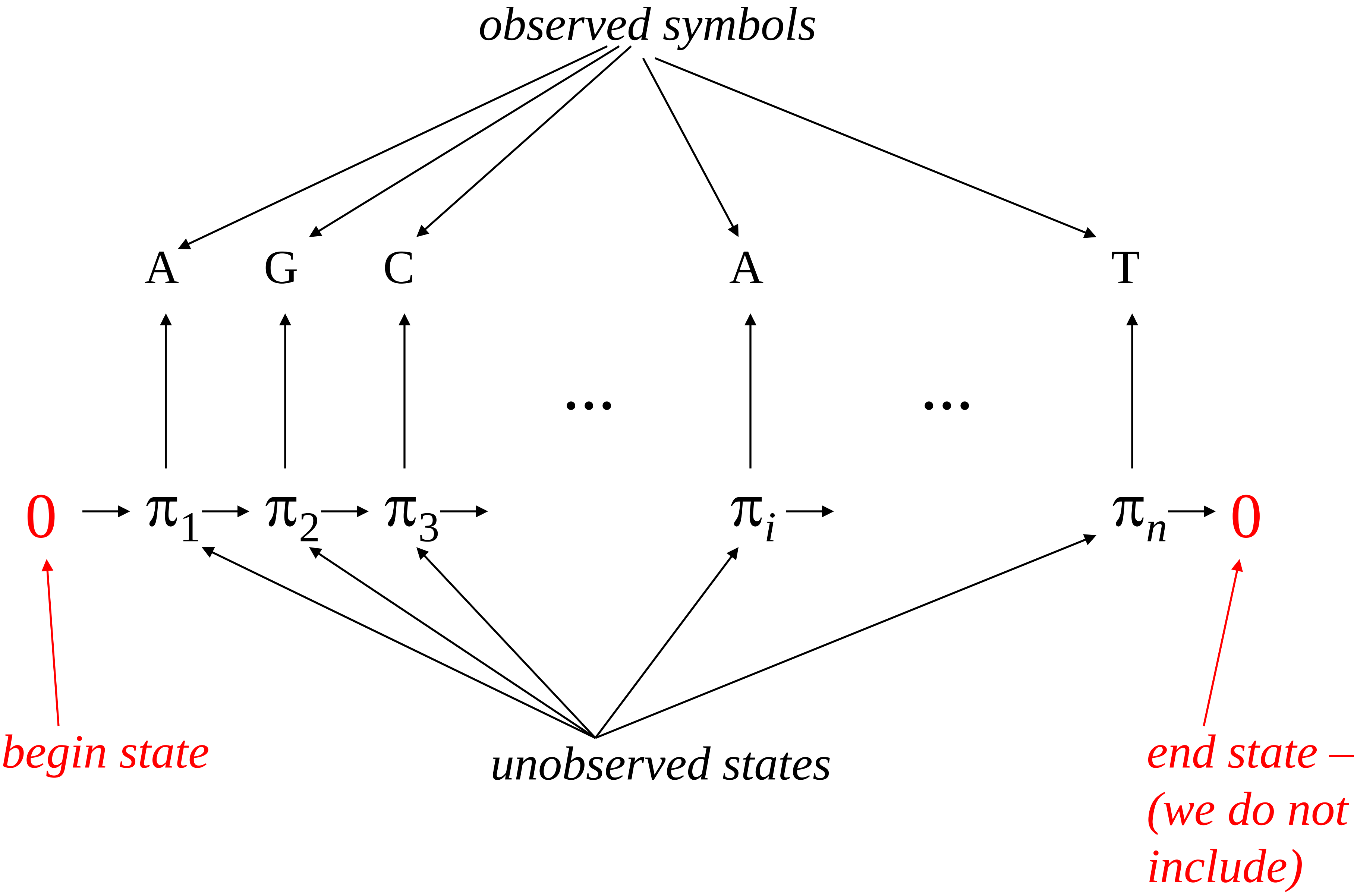


observed symbols

A G C … A … T

$\pi_1 \rightarrow \pi_2 \rightarrow \pi_3 \rightarrow \quad \pi_i \rightarrow \quad \pi_n$

unobserved states



observed symbols

unobserved states

begin state

end state –
(we do not
include)

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 B = {b} of *observed symbols*
- Set S = {k} of *hidden states* (usually k = 0,1, 2 ...,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

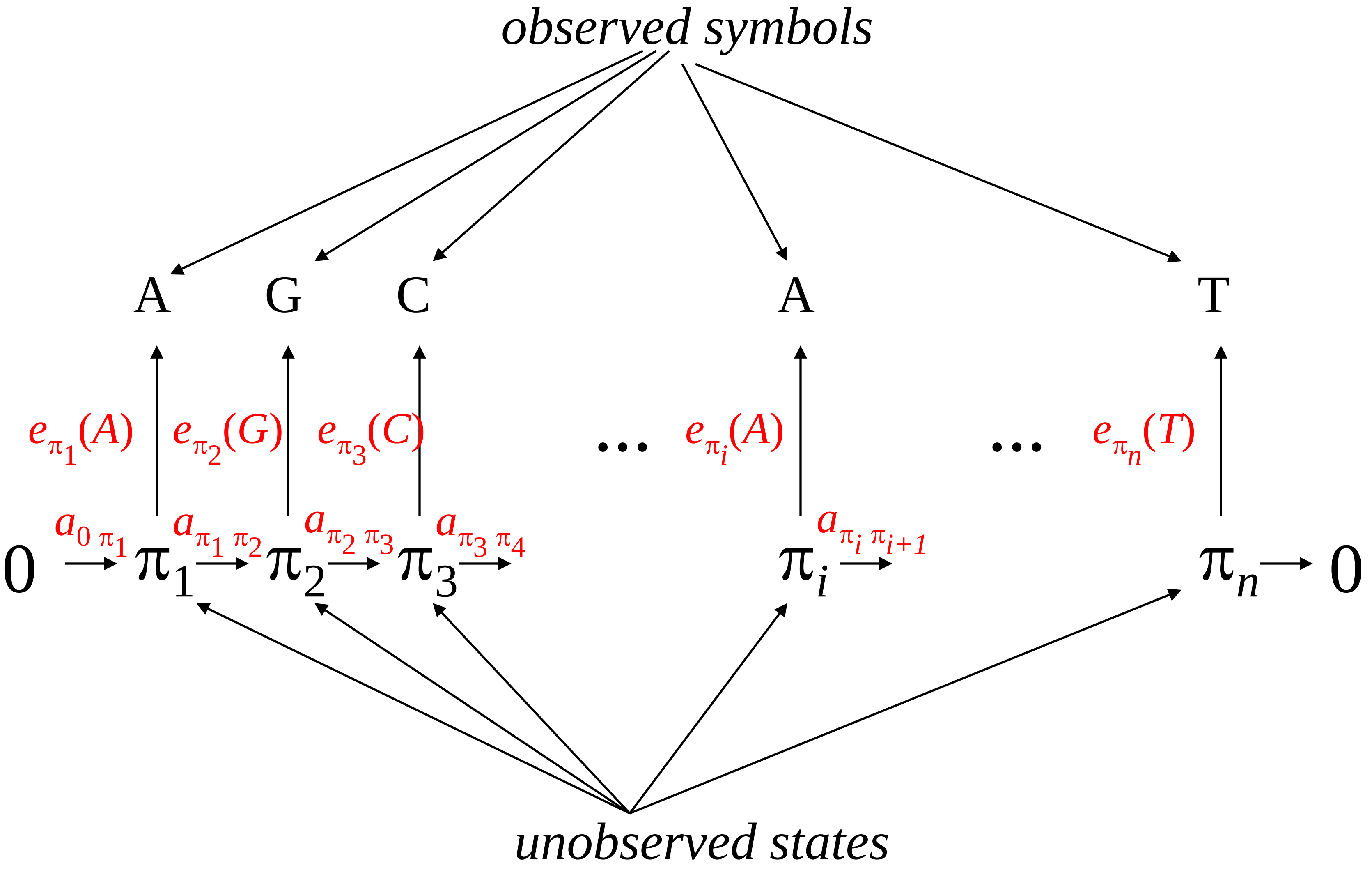
  transition probabilities (a_{kl}): a_{kl} = Prob(next state is l | curr state is k)

  $\sum_l a_{kl} = 1$, 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



observed symbols

A G C A T

$e_{\pi_1}(A)$ $e_{\pi_2}(G)$ $e_{\pi_3}(C)$ … $e_{\pi_i}(A)$ … $e_{\pi_n}(T)$

$0 \xrightarrow{a_{0\,\pi_1}} \pi_1 \xrightarrow{a_{\pi_1\,\pi_2}} \pi_2 \xrightarrow{a_{\pi_2\,\pi_3}} \pi_3 \xrightarrow{a_{\pi_3\,\pi_4}} \quad \pi_i \xrightarrow{a_{\pi_i\,\pi_{i+1}}} \quad \pi_n \rightarrow 0$

unobserved states

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

  emission probabilities:

  $e_k(b)$ = Prob(observed symbol is b | curr state is k)

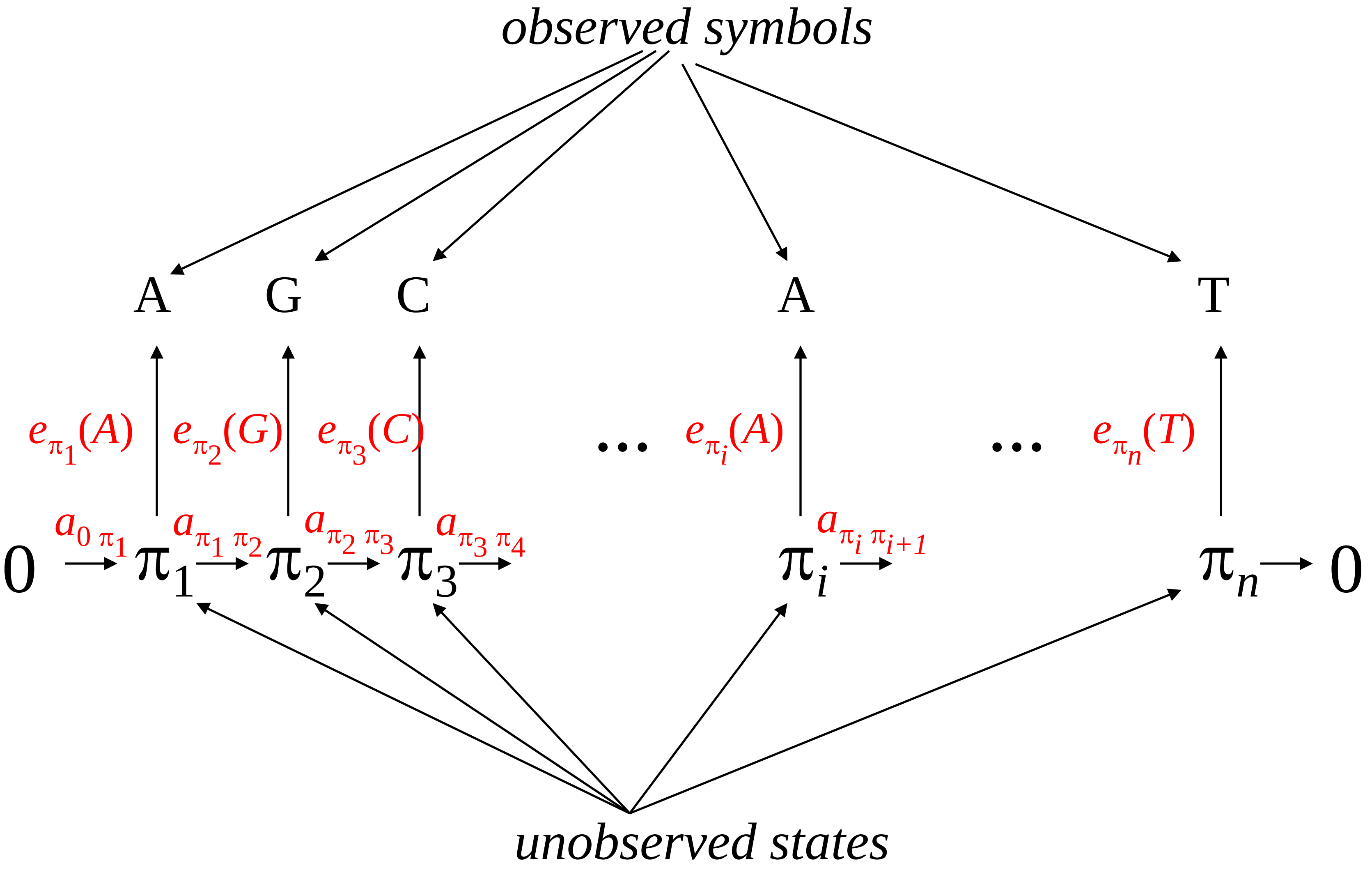
  (begin and end states do not emit symbols)

- 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



observed symbols

A G C A T

$e_{\pi_1}(A)$ $e_{\pi_2}(G)$ $e_{\pi_3}(C)$ … $e_{\pi_i}(A)$ … $e_{\pi_n}(T)$

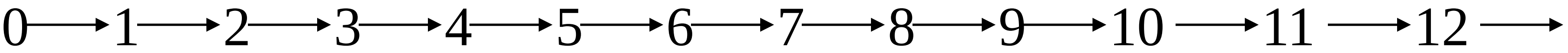
$0 \xrightarrow{a_{0\,\pi_1}} \pi_1 \xrightarrow{a_{\pi_1\,\pi_2}} \pi_2 \xrightarrow{a_{\pi_2\,\pi_3}} \pi_3 \xrightarrow{a_{\pi_3\,\pi_4}}$ $\pi_i \xrightarrow{a_{\pi_i\,\pi_{i+1}}}$ $\pi_n \rightarrow 0$

unobserved states

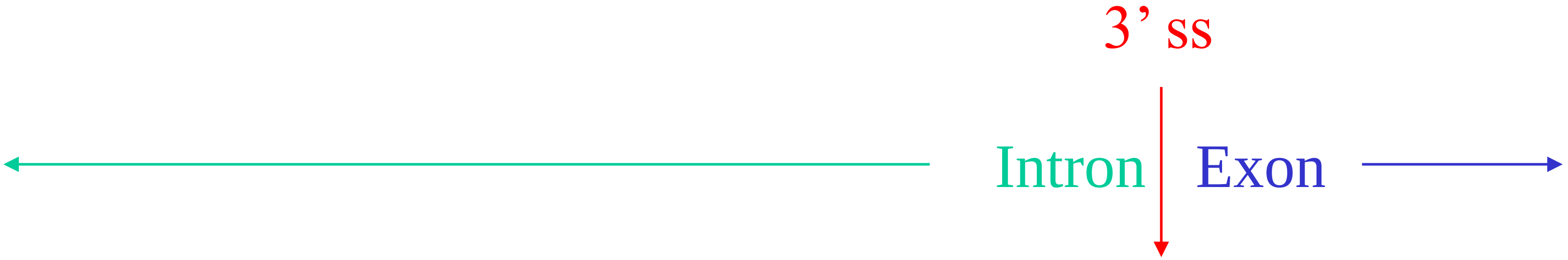
HMM Examples

- 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 (transits with non-zero prob – all are 1)



HMM for *C. elegans* 3’ Splice Sites


3’ ss

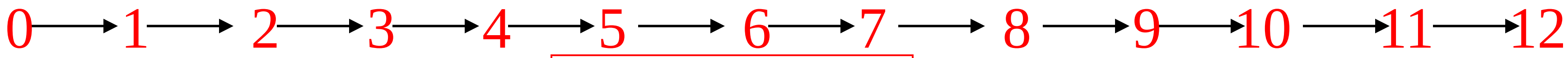
Intron

Exon

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

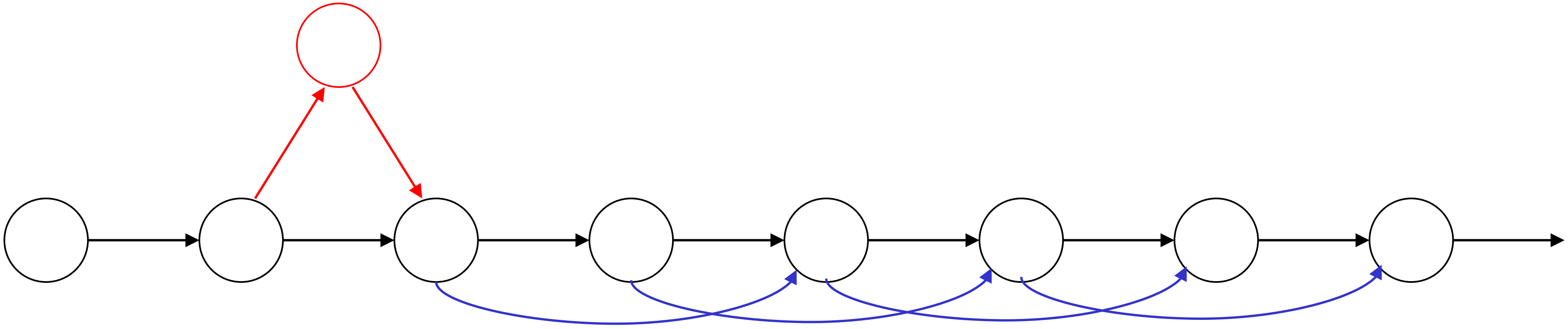

0 → 1 → 2 → 3 → 4 → 5 → 6 → 7 → 8 → 9 → 10 → 11 → 12

‘hidden’ states

- Can expand model 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 no longer necessarily 1 or 0

Insertions & Deletions in Site Model

insertion state



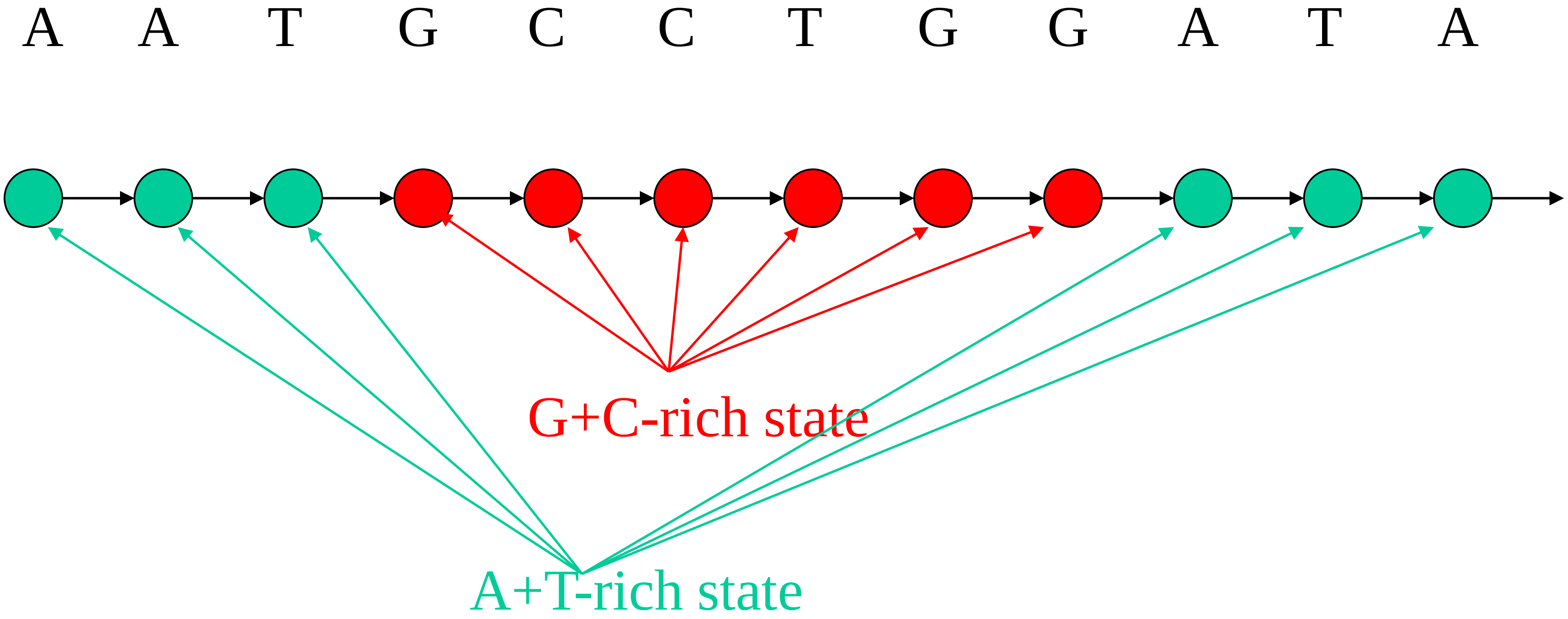
other transitions correspond
to deletions

Examples (cont’d) – 1-state HMMs

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

Examples (cont'd) – 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).



A A T G C C T G G A T A

G+C-rich state

A+T-rich state

2-state HMMs

- Can find most probable state decomposition (‘Viterbi path’) consistent with observed sequence
- Advantages over linked-list dynamic programming method (lecture 4) for finding high-scoring segments:
  - That method assumes you *know* appropriate parameters to find targeted regions; HMM method can *estimate* parameters.
  - HMM (easily) finds multiple segments
  - HMM can attach *probabilities* to alternative decompositions
  - HMM generalization to > 2 *types* of segments is easy – just allow more states!
- Disadvantage:
  - Markov assumption on state transitions implies geometric distribution for lengths of regions -- may not be appropriate