

# Probabilistic Sequence Models

Partly based on Chapter 3 of  
*Biological Sequence Analysis*

R. Durbin, S. Eddy, A. Krogh, G. Mitchinson  
[DEKM]

# Sequence questions

- *Discrimination*: decide whether a sequence belongs to a certain class of sequences
  - Is this DNA fragment coding?
  - Is this a CpG island? (promoters and start regions)
- *Segmentation*: parse a sequence into regions belonging to different classes
  - Parse this genomic sequence into coding and non-coding regions

# Sequence statistics

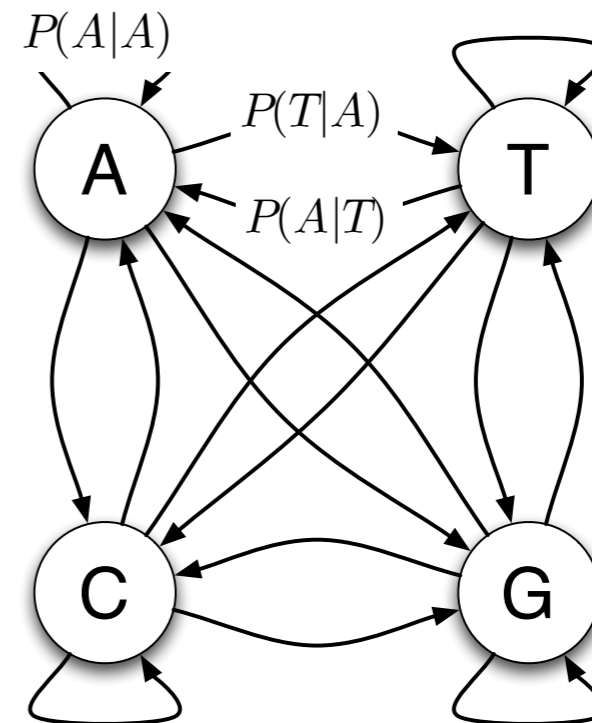
- Shorth-range statistics:  $k$ th-order Markov approximation

$$\begin{aligned} P(x_i | x_1 \cdots x_{i-1}) &\approx P(x_i | x_{i-k} \cdots x_{i-1}) \\ P(\mathbf{x}) &= \prod_{i=1}^n P(x_i | x_1 \cdots x_{i-1}) \\ &\approx \prod_{i=1}^n P(x_i | x_{i-k} \cdots x_{i-1}) \end{aligned}$$

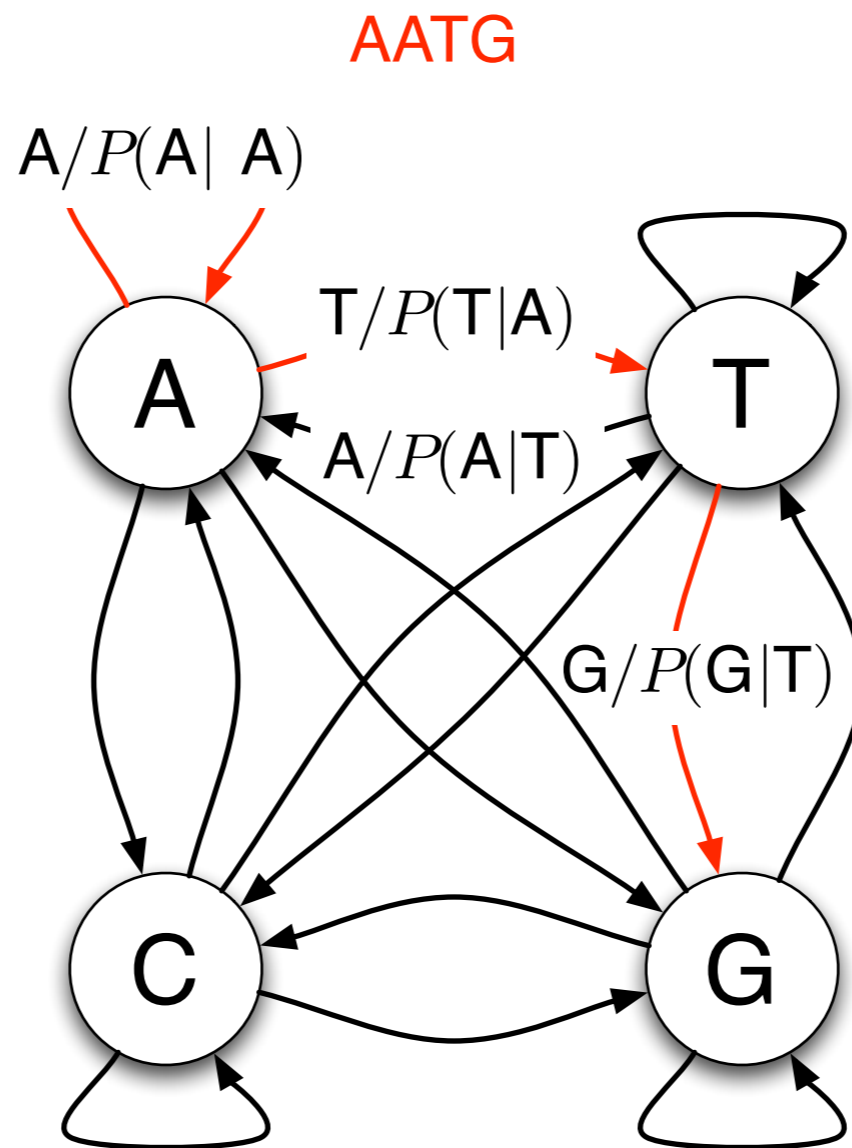
- First-order DNA model:

- Notation:

$$a_{st} = P(t|s)$$



# Markov chain as weighted automaton



# Boundary states

- *Sequence start*: assume special “invisible” sequence start markers to provide enough context before the actual start
- *Explicit sequence end*: use a special “invisible” end marker

$$P(\mathbf{x}) = P(x_1|\text{start}) \prod_{i=2}^n P(x_i|x_{i-1})P(\text{end}|x_n)$$

# Discrimination with Markov models

- Construct two models: *positive* and *negative*

$$\begin{aligned} a_{st}^+ &= \frac{C^+(s,t)}{\sum_{t'} C^+(s,t')} \\ a_{st}^- &= \frac{C^-(s,t)}{\sum_{t'} C^-(s,t')} \\ C^M(s,t) &= \text{count of } st \text{ words in } M \text{ data} \end{aligned}$$

- Score (log likelihood ratio):

$$S(\mathbf{x}) = \log \frac{P(\mathbf{x}|+)}{P(\mathbf{x}|-)} = \sum_{i=1}^n \log \frac{a_{x_{i-1}x_i}^+}{a_{x_{i-1}x_i}^-}$$

# Estimating transition probabilities

- The maximum likelihood estimator (empirical relative frequency) fails if there are very rare words
  - unseen transitions get zero probability
- Count smoothing techniques:
  - pseudo counts
  - Bayesian interpretation
  - discounting and Good-Turing estimates

# Segmentation

- Assume two first-order models (+ and -)
- What's the most likely segmentation of a sequence into alternating portions generated by the two models?



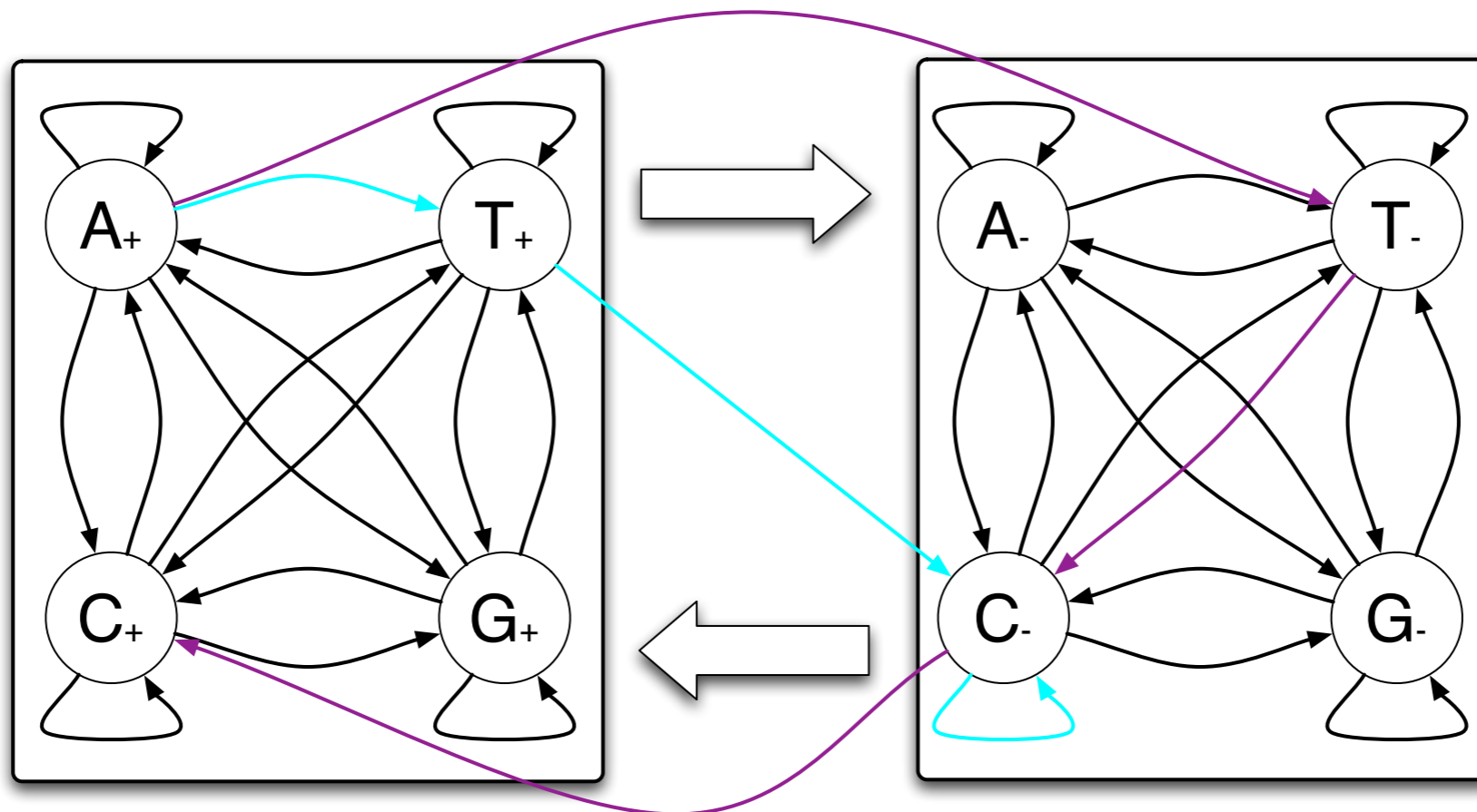
- How do we switch between models?

# Hidden states

- Transitions between all pairs of states
- Multiple state sequences compatible with given letter sequence

ATCC

+- -+  
++ --



# Hidden Markov model

- State transition probabilities

$$a_{st} = P(s_i = t | s_{i-1} = s)$$

- Emission probabilities

$$e_s(x) = P(x_i = x | s_i = s)$$

- Previous example:

$$e_{N_+}(N') = e_{N_-}(N') = \delta(N, N')$$

# State sequences

- Joint probability of a state sequence and an input sequence

$$P(\mathbf{x}, \mathbf{s}) = \prod_i a_{s_{i-1} s_i} e_{s_i}(x_i) \quad \text{where } s_0 = \text{start}$$

- Most probable state sequence

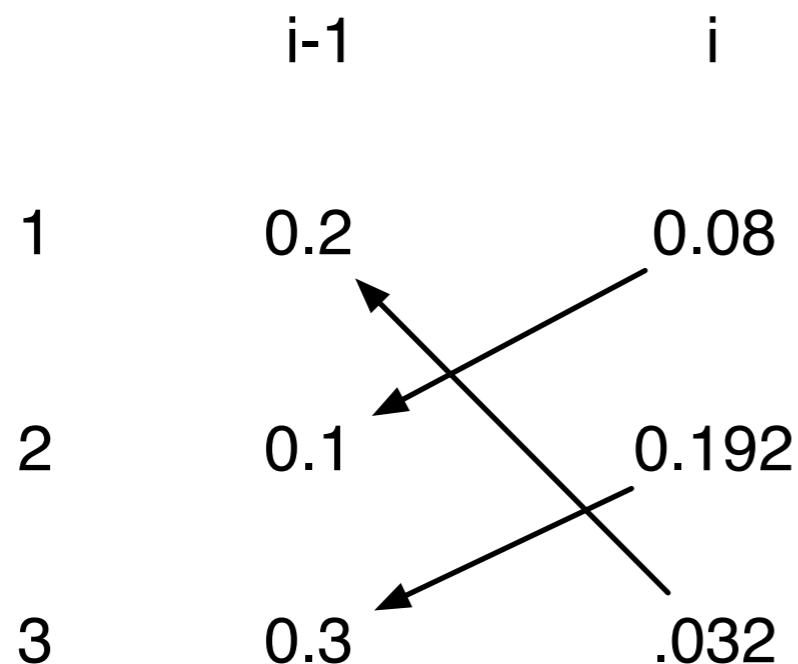
$$\arg \max_{\mathbf{s}} P(\mathbf{s}|\mathbf{x}) = \arg \max_{\mathbf{s}} \frac{P(\mathbf{x}, \mathbf{s})}{P(\mathbf{x})} = \arg \max_{\mathbf{s}} P(\mathbf{x}, \mathbf{s})$$

# Viterbi algorithm

- Most probable path ending in a state  $t$  at input position  $i$

$$v_i(t) = e_t(x_i) \max_s v_s(i-1) a_{st}$$

- Viterbi algorithm



$a_{st}$	1	2	3
1	0.1	0.1	0.8
2	0.8	0.1	0.1
3	0.1	0.8	0.1

$s$	$e_s(x_i)$
1	0.1
2	0.8
3	0.2