BIO-463
Genomics and bioinformatics

Lecture 3: Hidden Markov Models

Professors: Jacques Rougemont, Anne-Florence Bitbol, Raphaëlle Luisier

EPFL

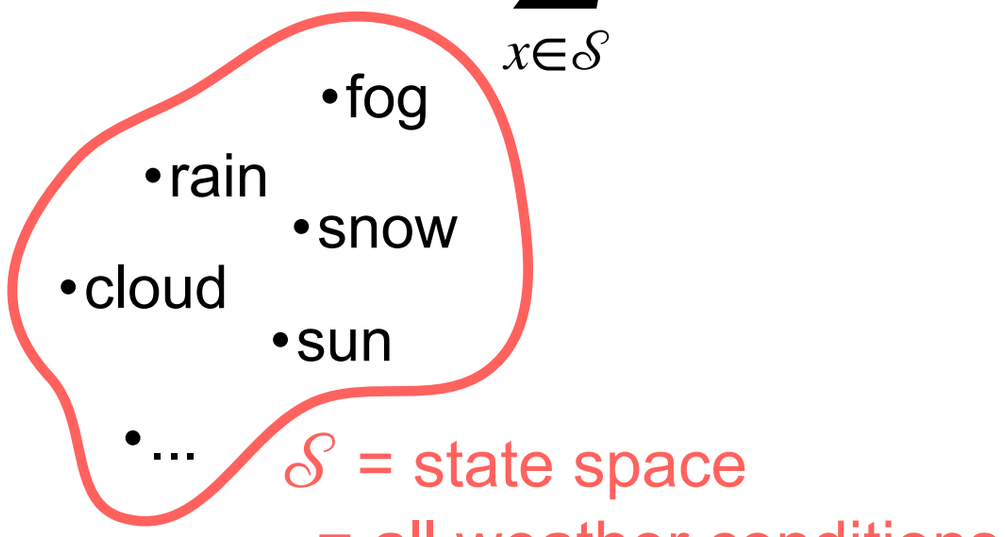
Probability: notation and formulas

$P(x)$ probability of event x
$P(x \mid y)$ probability of x knowing y (**conditioned on** y)
$P(x, y)$ probability of x and y

$$P(x, y) = P(x \mid y)P(y)$$

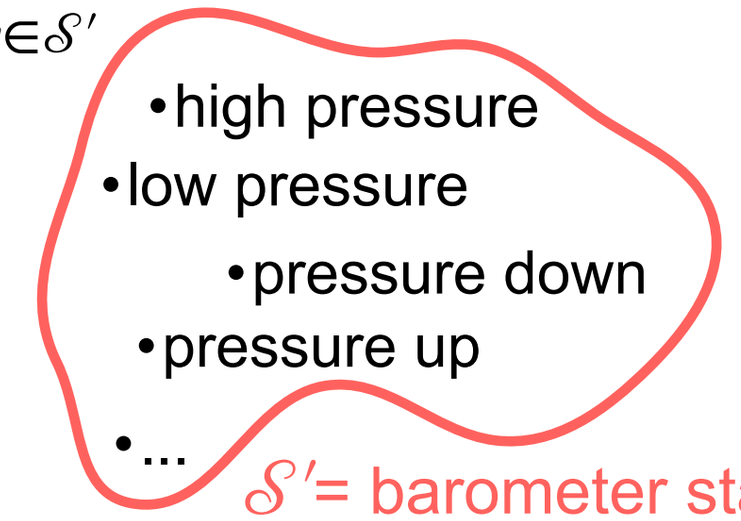
If x, y independent: $P(x, y) = P(x)P(y)$, therefore $P(x \mid y) = P(x)$

$$\sum_{x \in \mathcal{S}} P(x) = 1$$



$\mathcal{S}$ = state space
= all weather conditions

$$\sum_{y \in \mathcal{S}'} P(x \mid y)P(y) = P(x)$$



$\mathcal{S}'$ = barometer state

Bayes "Theorem"

D = "data"

θ = "parameters"

$$\underbrace{P(\theta \mid D)}_{\text{posterior}} = \frac{\overbrace{P(D \mid \theta)}^{\text{likelihood}} \overbrace{P(\theta)}^{\text{prior}}}{P(D)} = \frac{P(D \mid \theta)P(\theta)}{\sum_y P(D \mid y)P(y)}$$

Rare disease: 1/10'000

Diagnostic test:

FP (false positive) = 1/100

Sensitivity = 99.9% (FN = 1/1000)

test result \ disease state	d (yes)	h (no)	
+	10(-0.01)	1000	1010
–	0(+0.01)	98990	98990
	10	99990	100000

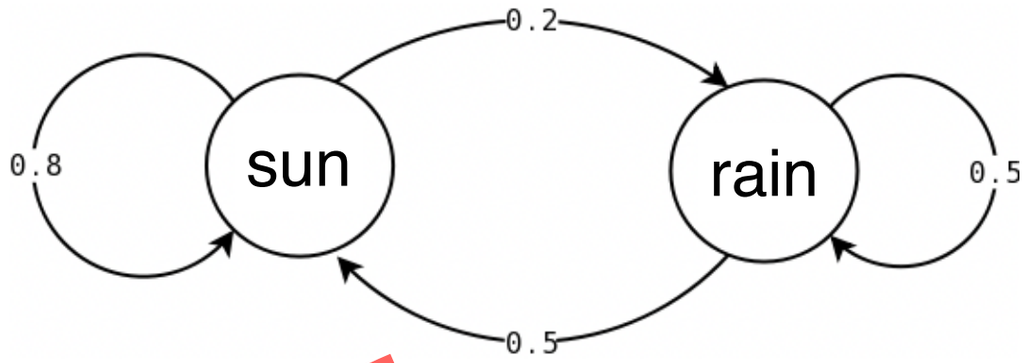
$$P(d|+) = \frac{P(+|d)P(d)}{P(+|d)P(d) + P(+|h)P(h)}$$

$$= \frac{0.999 \cdot 10^{-4}}{0.999 \cdot 10^{-4} + 0.01 \cdot (1 - 10^{-4})}$$

$$\approx \frac{10^{-4}}{10^{-4} + 10^{-2}} = \frac{1}{101}$$

Markov models

- discrete time evolution
- finite state space
- fixed set of transition probabilities



$$M = \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.8 & 0.5 \\ 0.2 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} \text{sun} \\ \text{rain} \end{array}$$

today → tomorrow

$$P(S_{n+1} = \mathbf{r} \,|\, S_n = \mathbf{s})$$

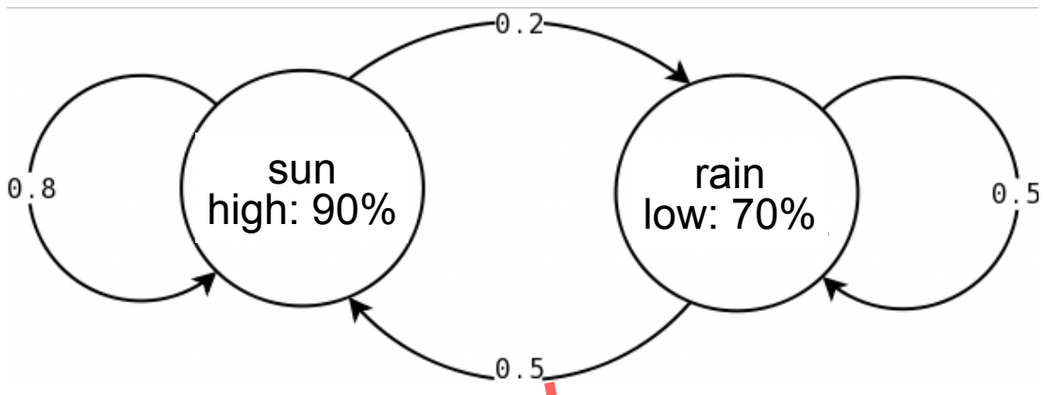
model generates plausible sequences

```
ssssrrsssssrsssrsrsssssssrrrsssssssrsssssssr...
```

$$P(S_{n+1} = \mathbf{s}) = \sum_{\sigma} P(S_{n+1} = \mathbf{s} \,|\, S_n = \sigma) P(S_n = \sigma)$$

$$\begin{pmatrix} p_s(t+1) \\ p_r(t+1) \end{pmatrix} = M \cdot \begin{pmatrix} p_s(t) \\ p_r(t) \end{pmatrix}$$

Hidden Markov models (HMM)



HMM generates

hidden states

transition matrix

$$M = \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.8 & 0.5 \\ 0.2 & 0.5 \end{pmatrix} \end{array} \begin{array}{l} \text{sun} \\ \text{rain} \end{array}$$

emission matrix

$$E = \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \end{array} \begin{array}{l} \text{high} \\ \text{low} \end{array}$$

emitted symbols

$$P(T_n = \mathbf{l} \,|\, S_n = \mathbf{s})$$

time

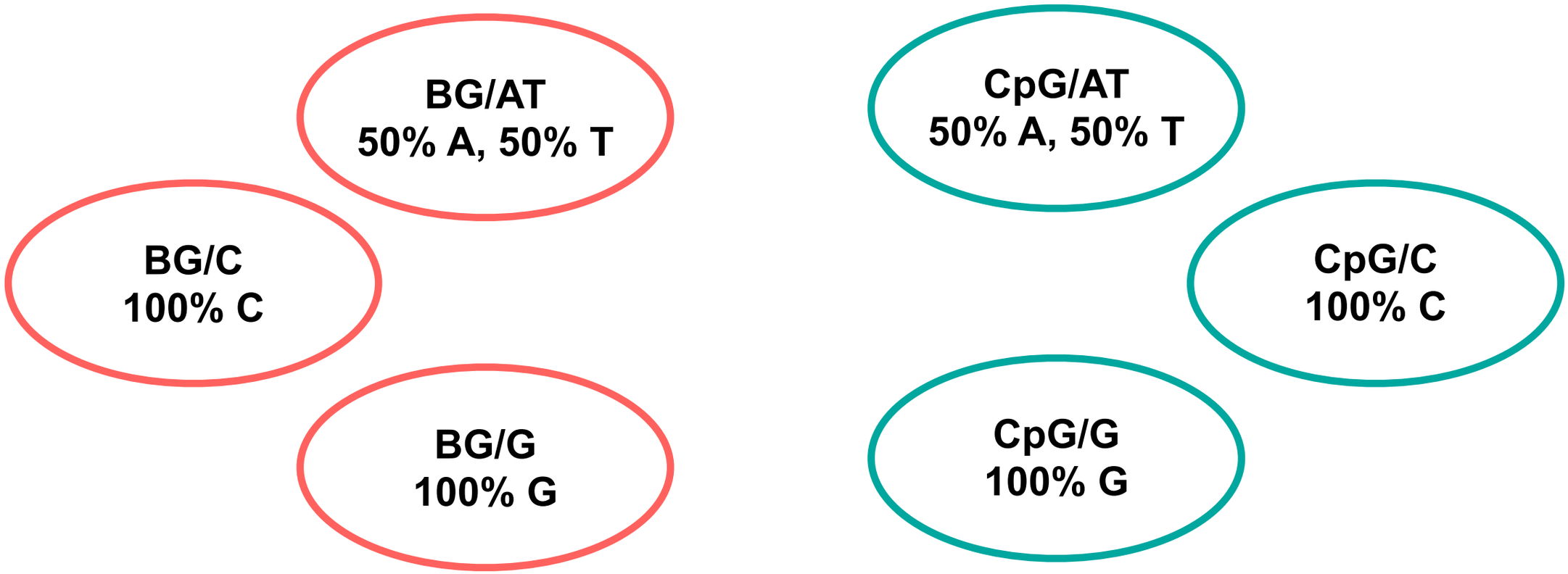
Observed sequence $\mathbf{T} =$ `hhlhhhhlllhhlhhlhhlhlllhhlllllhlhlllhllh`

Hidden sequence $\mathbf{S} =$ `ssssrrssssrsssrsrsssssssrrrsssssssrsssssssr`

Example: CpG islands

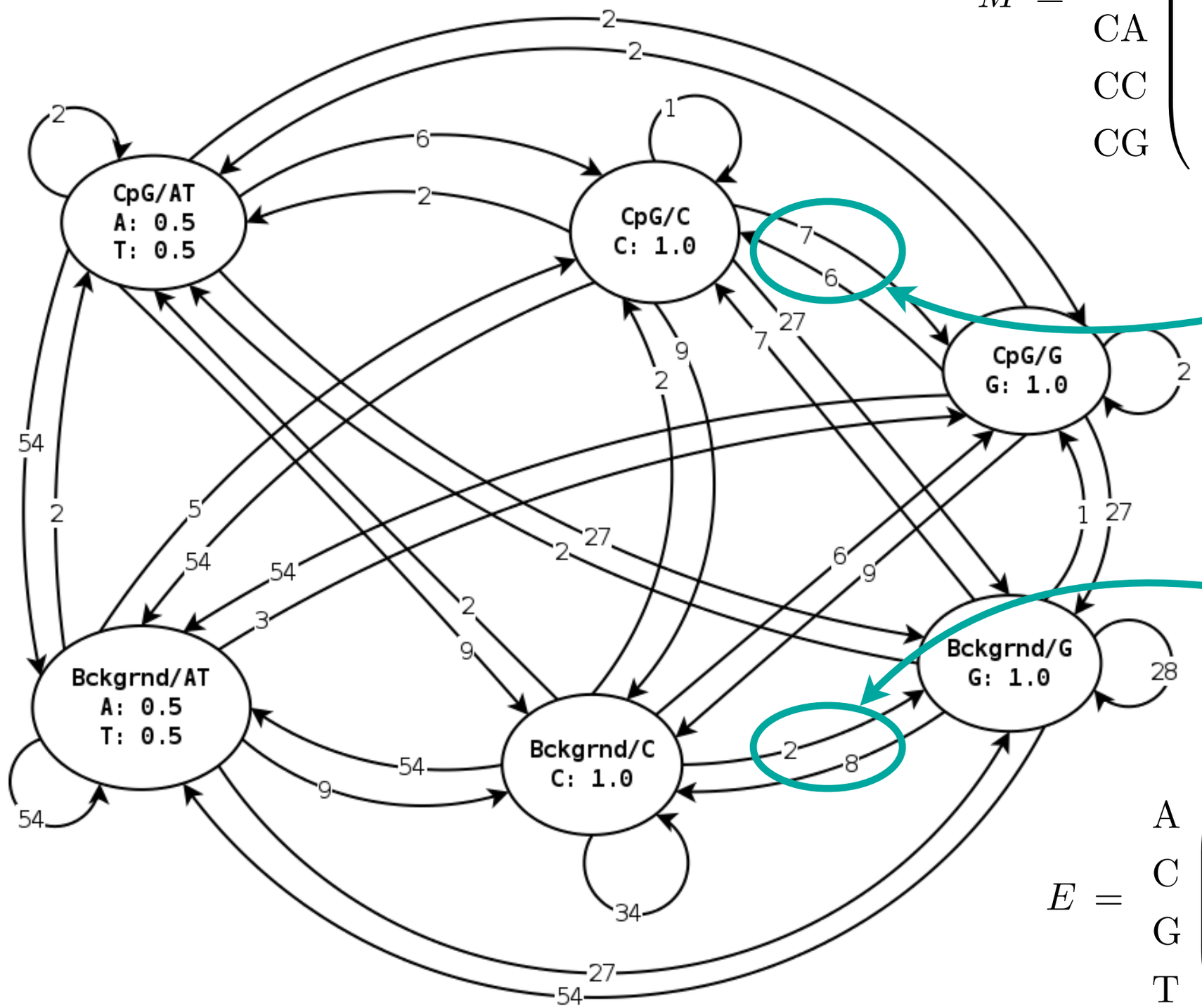
- Human genome: 30% A/T, 20% C/G
- Frequency of GC ~ 4%, CC and GG ~ 5%, but CG < 1% !
- However there are regions called CpG islands containing many consecutive CG

- HMM states: Background (not a CpG island) or CpG
- States are further divided into A/T, C, G



Example: CpG islands

$$M = \begin{array}{c} \\ \text{BA} \\ \text{BC} \\ \text{BG} \\ \text{CA} \\ \text{CC} \\ \text{CG} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{BA} & \text{BC} & \text{BG} & \text{CA} & \text{CC} & \text{CG} \end{array} \\ \begin{pmatrix} .54 & .54 & .54 & .54 & .54 & .54 \\ .09 & .34 & .08 & .09 & .09 & .09 \\ .27 & .02 & .28 & .27 & .27 & .27 \\ .02 & .02 & .02 & .06 & .02 & .02 \\ .05 & .02 & .07 & .02 & .01 & .06 \\ .03 & .06 & .01 & .02 & .07 & .02 \end{pmatrix} \end{array}$$



CG, GC > CC, GG

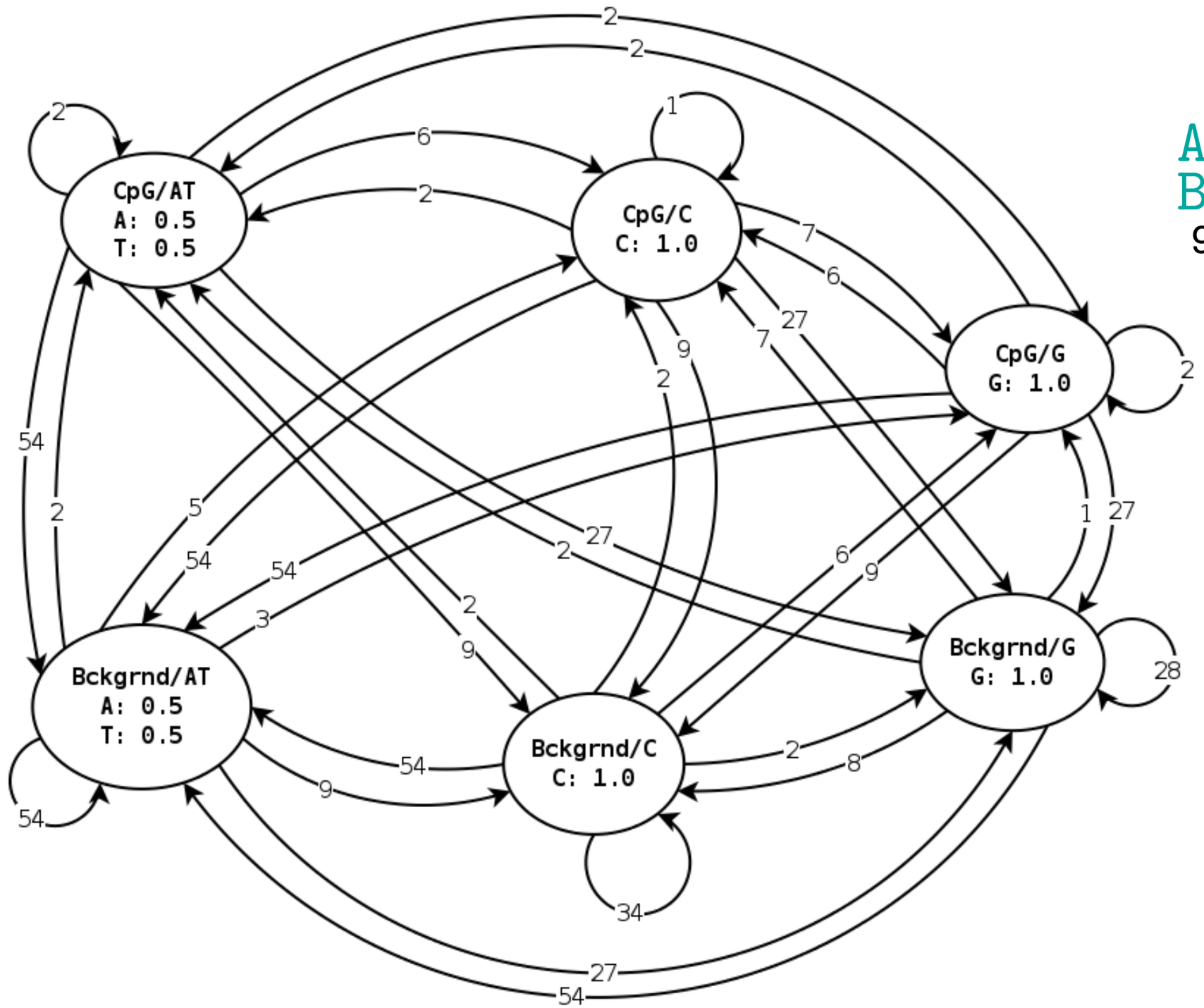
CG < CC, GC, GG

$$E = \begin{array}{c} \\ \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{array} \begin{array}{c} \begin{array}{cccccc} \text{BA} & \text{BC} & \text{BG} & \text{CA} & \text{CC} & \text{CG} \end{array} \\ \begin{pmatrix} 0.5 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0.5 & 0 & 0 & 0.5 & 0 & 0 \end{pmatrix} \end{array}$$

Example: CpG islands

ACGCGTCGCGA
BCCCCCCCCCCB
5 7 6 7 2 6 7 6 7 54 = 2.8e-12

ACGCGTCGCGA
BBBBBBBBBBB
9 2 8 2 54 9 2 8 2 54 = 2.4e-12



Hidden Markov Models (HMM)

Questions:	Answers:
• What is the sequence $\mathbf{S}$ of hidden states most likely to generate the observed symbols $\mathbf{T}$?	• Viterbi algorithm
• What is the probability of the observed sequence $\mathbf{T}$?	• Forward / Backward algorithms
• What are the parameters E, M (emission and transition probabilities) that maximize the probability of $\mathbf{T}$?	• Baum-Welch algorithm

Hidden Markov Models (HMM)

Questions:

- What is the sequenc[illegible] generate the observed symbols $\mathbf{I}$?

Decoding the observations

- What is the probabili[illegible]

Evaluating the model

- What are the parameters E, M (emission and transition probabilities) that ma[illegible]

Optimizing the model

Answers:

- Vitarbi algorithm
- Forward / Backward [illegible]hms
- [illegible]-Welch algorithm

HMM formulas

Probability of hidden sequence:

initial state

transition

$$P(\mathbf{S}) = P(S_0) \prod_{n=1}^{N} M(S_n, S_{n-1})$$

Probability of observed sequence (knowing hidden states):

emission from hidden state

$$P(\mathbf{T}|\mathbf{S}) = \prod_{n=1}^{N} P(T_n|S_n) = \prod_{n=1}^{N} E(T_n, S_n)$$

Bayes theorem:

$$P(\mathbf{S}|\mathbf{T}) = \frac{P(\mathbf{T}|\mathbf{S})P(\mathbf{S})}{P(\mathbf{T})} = \frac{P(\mathbf{T}|\mathbf{S})P(\mathbf{S})}{\sum_{\sigma} P(\mathbf{T}|\sigma)P(\sigma)}$$

Numerical stability:

$$\log P(\mathbf{S}) = \log P(S_0) + \sum_{n=1}^{N} \log M(S_n, S_{n-1})$$

Viterbi algorithm

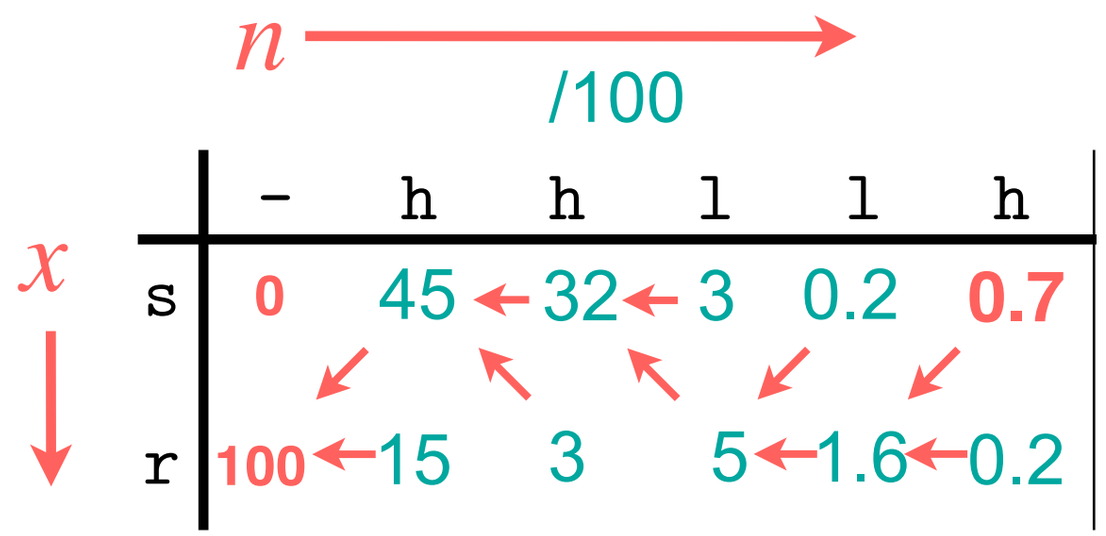
Example

$\mathbf{T} = \texttt{hhllh}$

$$V_{n,x} = E(T_n, x) \max_y M(x, y) V_{n-1,y}$$

$$M = \begin{array}{c} \\ \text{sun} \\ \text{rain} \end{array} \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.8 & 0.5 \\ 0.2 & 0.5 \end{pmatrix} \end{array}$$

$$E = \begin{array}{c} \\ \text{high} \\ \text{low} \end{array} \begin{array}{c} \begin{array}{cc} \text{sun} & \text{rain} \end{array} \\ \begin{pmatrix} 0.9 & 0.3 \\ 0.1 & 0.7 \end{pmatrix} \end{array}$$



Most probable sequence: `ssrrs`

Forward / Backward algorithm

- What is the probability of my observation $\mathbf{T}$? $P(\mathbf{T}) = \sum_{\mathbf{s}} P(\mathbf{T}\,|\,\mathbf{s})P(\mathbf{s})$
- Forward score (replace max by sum in Viterbi):

$$F_{n,x} = E(T_n, x)\sum_y M(x,y)F_{n-1,y} = P(T_1, \ldots, T_n, S_n = x)$$

- Backward (similar to Forward, move from right to left):

$$B_{n,x} = \sum_y E(T_{n+1}, y)M(y,x)B_{n+1,y} = P(T_{n+1}, \ldots, T_N\,|\,S_n = x)$$

$$\frac{P(\mathbf{T}\,|\,S_n = x)\cdot P(S_n = x)}{P(\mathbf{T})}$$

$P(\ldots T_n, S_n) = P(\ldots T_n\,|\,S_n)P(S_n)$

$P(\ldots T_n, T_{n+1}\ldots\,|\,S_n) = P(\ldots T_n\,|\,S_n)P(T_{n+1}\ldots\,|\,S_n)$

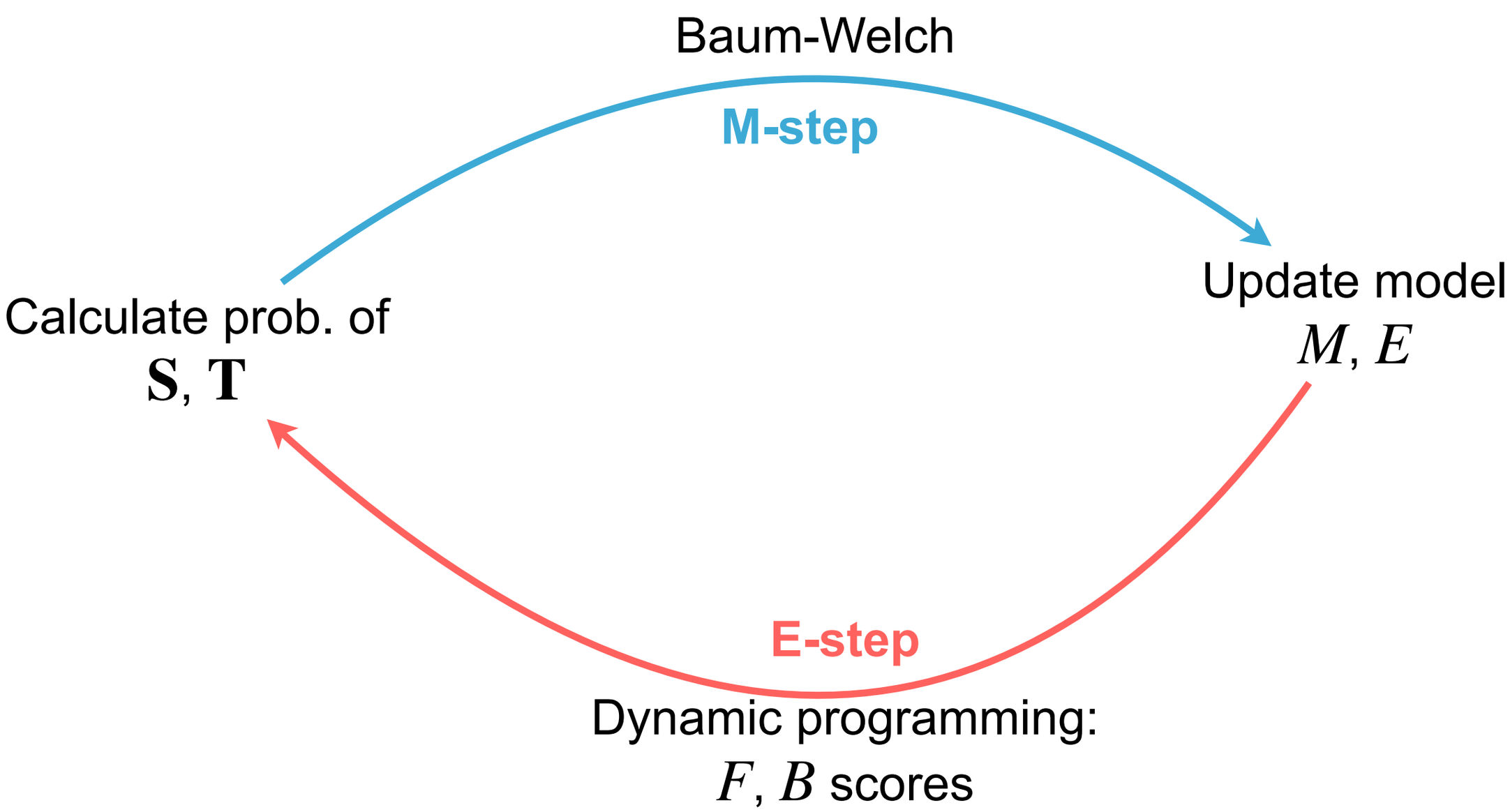
$$\frac{F_{n,x}\cdot B_{n,x}}{\sum_y F_{N,y}} = \frac{P(T_1, \ldots, T_n, S_n = x)\cdot P(T_{n+1}, \ldots, T_N\,|\,S_n = x)}{P(\mathbf{T})} = P(S_n = x\,|\,\mathbf{T})$$

Bayes

Probability of any hidden state given observed $\mathbf{T}$

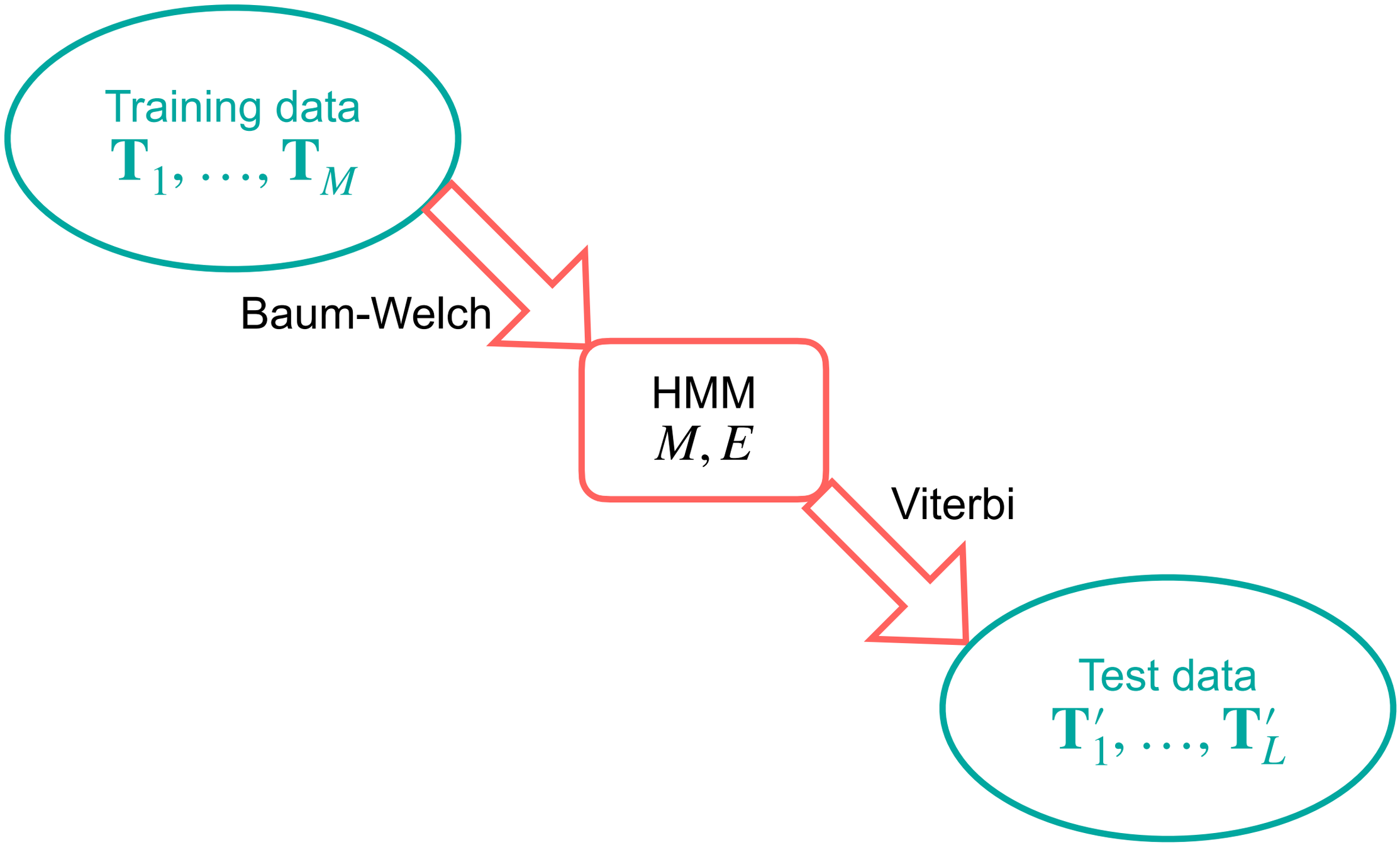
Baum-Welch optimization

Example of Expectation-Maximization (EM) algorithm

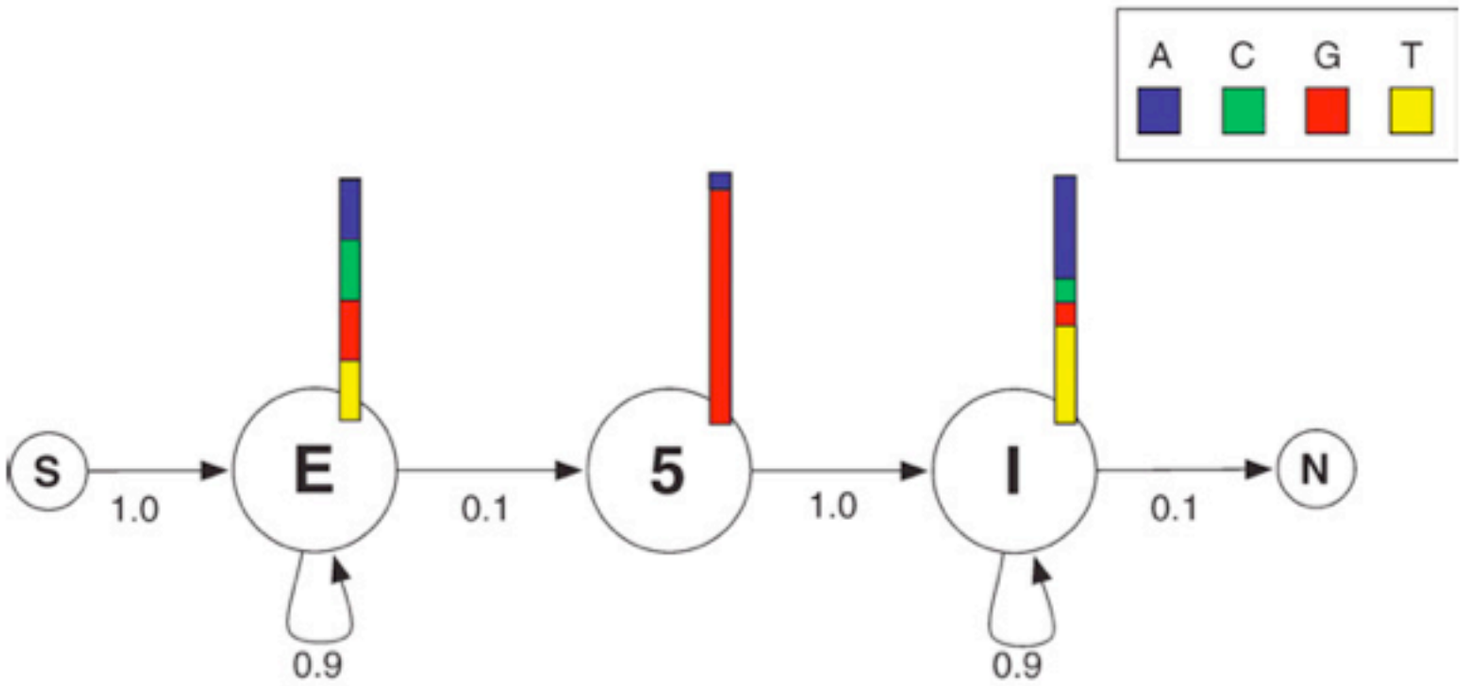


⇒ converges to a model which maximizes $P(\mathrm{T})$

Model "learning"



Modeling spliced genes

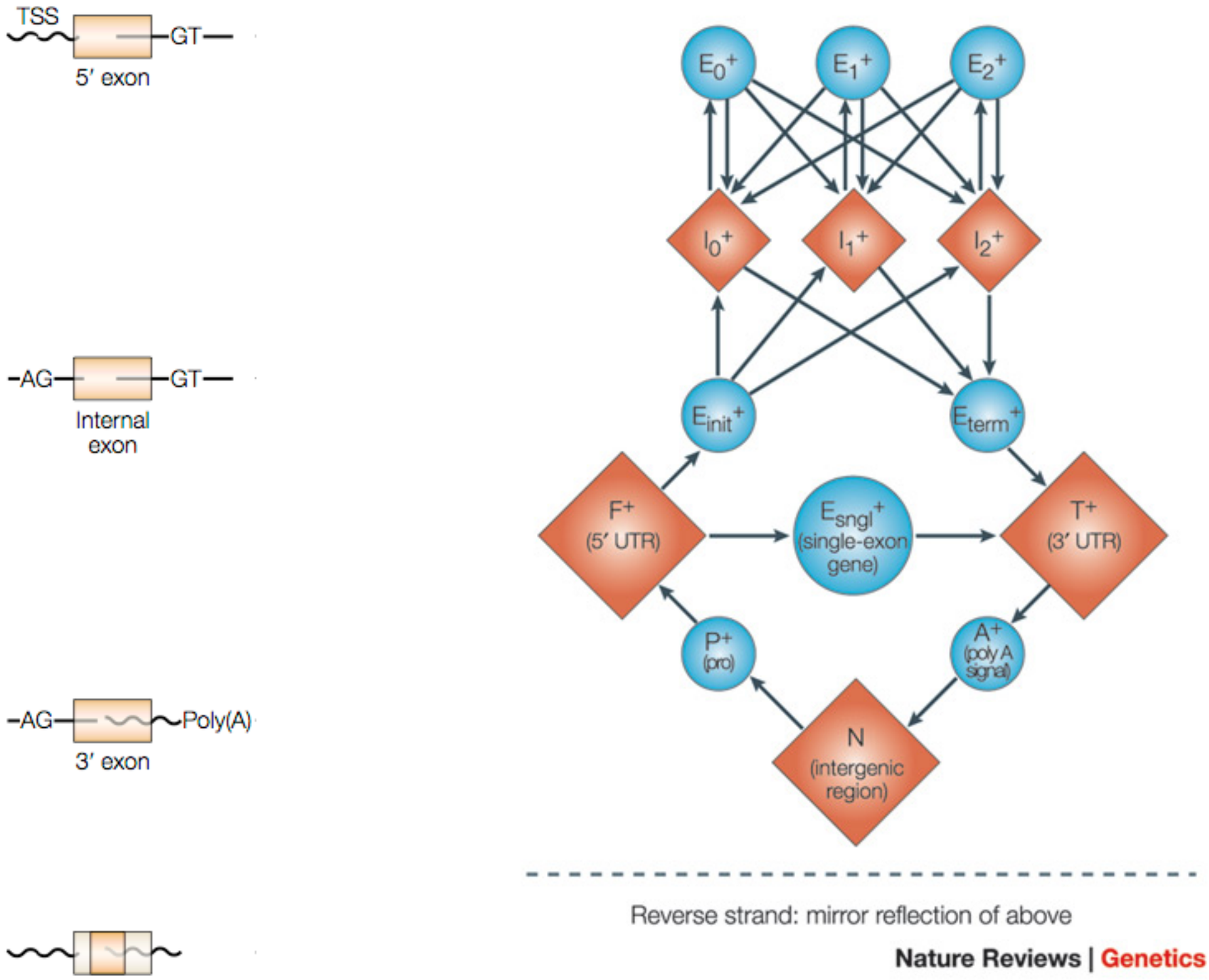


S = ...EEEEEEEE5IIIIIII...
T = ...CAGTGTAAGTATCATT...

Schuster-Böckler B, Bateman A,
Current protocols in bioinformatics (2007)

Genscan programm

wikipedia:GENSCAN



Zhang MQ, Nat Rev Gen (2002)