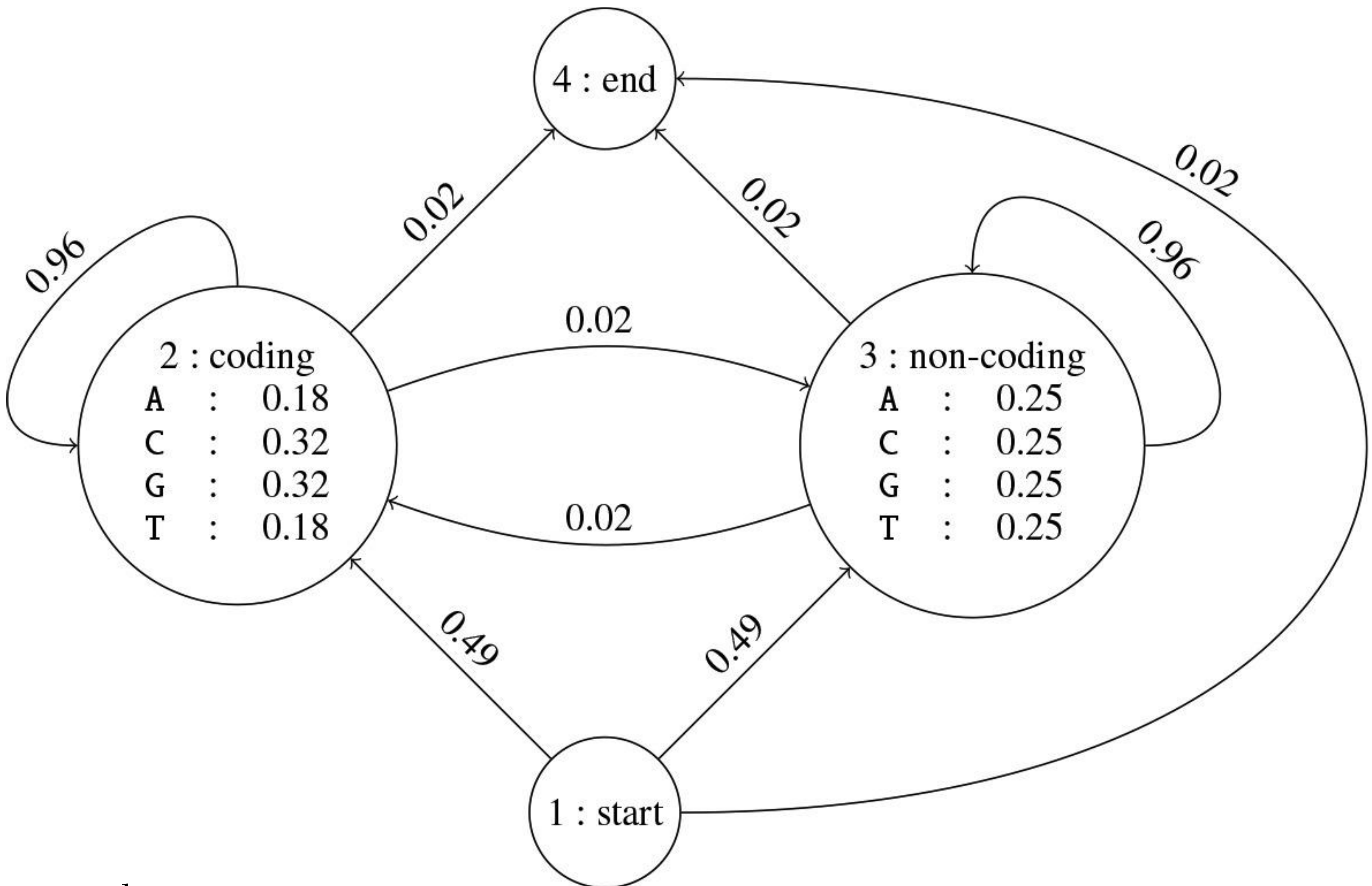
Algorithms in Genome Analysis, Spring 2023

Veli Mäkinen

Week 5

Hidden Markov Models (HMMs)

Max probable path = segmentation into coding and non-coding regions



coding   coding

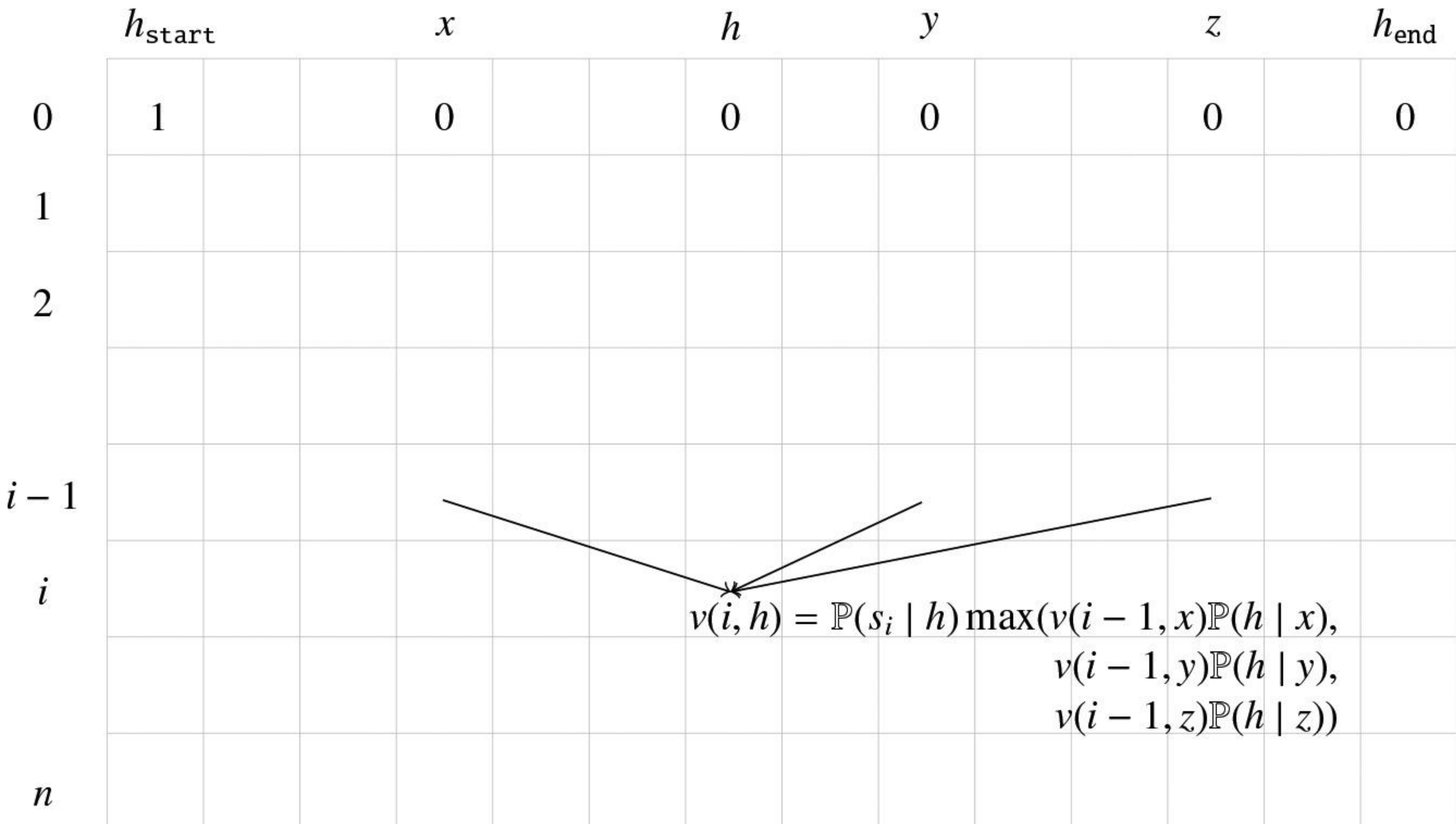
CACCGCTGCCGAAATTCTATGGTTAGTACTGTCATGATCTTCCGAGTCCGAATCAGA

HMM

- Directed graph with set H of hidden states emitting symbols and set T of transitions
- Emission probabilities p(c | h) in state h sum to 1
- Transition probabilities p(h | h’) to exit state h’ sum to 1
- Unique source (no transitions in) and sink (no transitions out)
- Given an input sequence, what is the most probable path from source to sink?
- Many other interesting questions to be studied under this framework

Viterbi

- Dynamic programming algorithm solving the most probable path problem
- $v(i,h)$ =max probability of a path emitting S[1..i] so that S[i] is emitted at state h
- $v(i,h) = p(S[i] \mid h) * \max_{(h',h)\in T} v(i-1,h') * p(h \mid h')$
- Initialization $v(0, start) = 1$
- Finalization
  $$v(|S|+1, end) = \max_{(h,end)\in T} v(|S|,h) * p(end \mid h)$$
  Most probable path can be traced back from $v(|S|+1, end)$
- For numerical accuracy, log probabilities are used so that all multiplications become summations


$h_{\texttt{start}}$
x
h
y
z
h_{end}
0
1
2
$i-1$
i
n
1
0
0
0
0
0
$v(i,h) = \mathbb{P}(s_i \mid h) \max(v(i-1,x)\mathbb{P}(h \mid x),$
$v(i-1,y)\mathbb{P}(h \mid y),$
$v(i-1,z)\mathbb{P}(h \mid z))$

Training

- With labeled training data one can use the observed frequencies to fix the emission and transition probabilities
- Without labels, common ways to proceed are
  - Viterbi training: Set initial probabilities based on background knowledge, find a most probable path P for input sequence S, label S according to P and use this as labeled data. Iterate.
  - Expectation maximation (EM): As above, but take all paths into account in proportion to their probabilities, calculating expected probability for each emission / transition. Iterate.

EM through Baum-Welsch 1/2

- Uses forward-backward variant of viterbi to find EM estimates
- $\mathrm{f}(i,h)$ =sum of probabilities of a paths emitting S[1..i] so that S[i] is emitted at state h
- $\mathrm{f}(i,h) = p(S[i] \mid h) * \sum_{(h',h)\in T} f(i-1,h') * p(h \mid h')$
- $\mathrm{b}(i,h)$ =sum of probabilities of a paths emitting S[i..|S|] so that S[i] is emitted at state h
- $\mathrm{b}(i,h) = p(S[i] \mid h) * \sum_{(h,h')\in T} b(i+1,h') * p(h' \mid h)$
- Initialization: $\mathrm{f}(0,start) = b(|S|+1,end) = 1$
- Finalization:
  - $\mathrm{f}(|S|+1,end) = \sum_{(h,end)\in T} f(|S|,h) * p(end \mid h)$
  - $\mathrm{b}(0,start) = \sum_{(start,h)\in T} b(1,h) * p(h \mid start)$

EM through Baum-Welch 2/2

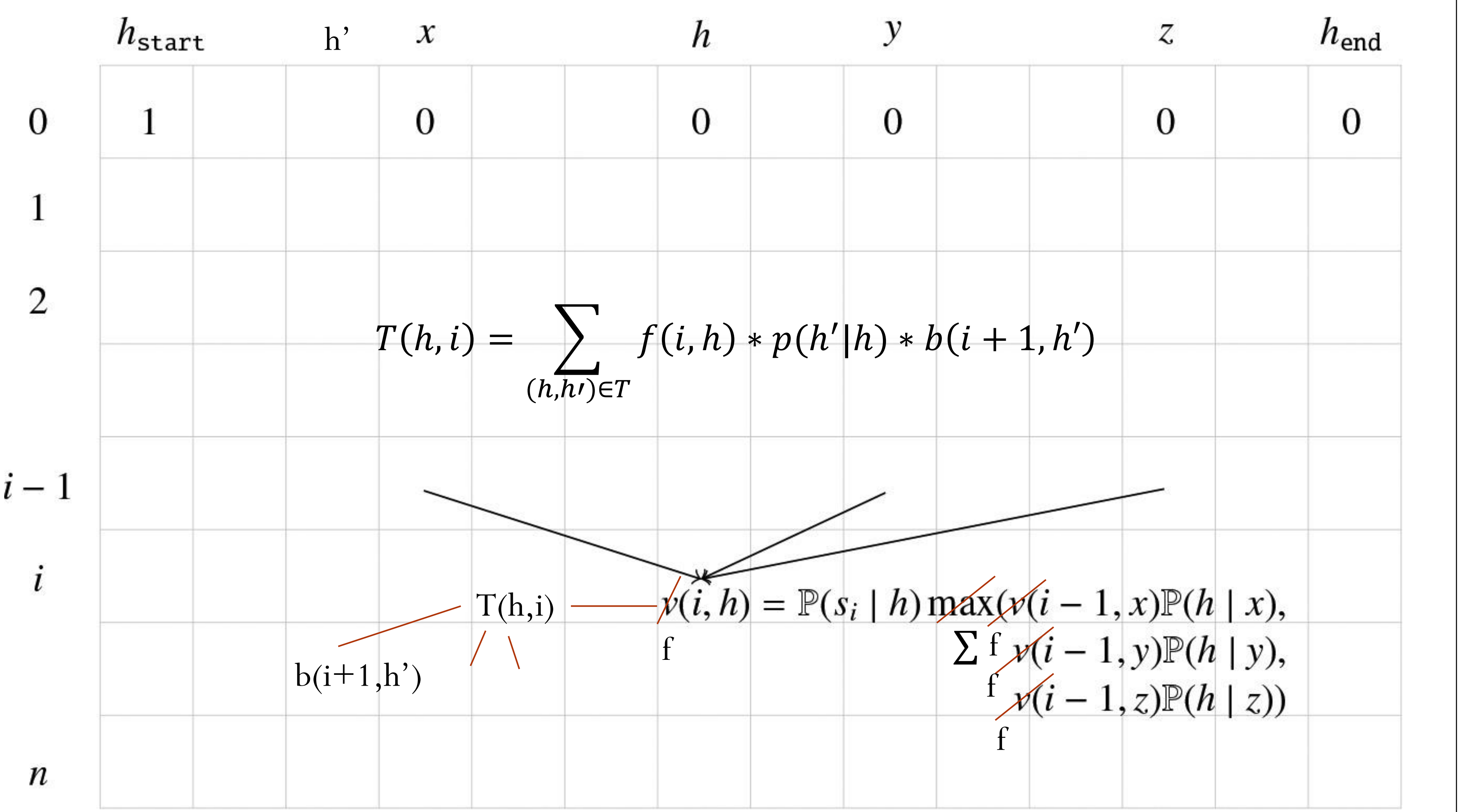
- Note: $\mathrm{f}(|S|+1, end) = b(0, start)$ is the total probability of emitting S
- The total probability $\mathrm{T}(h, i)$ of emitting S[i] at state h is

$$\sum_{(h',h)\in T} f(i-1, h') * p(h|h') * b(i, h)$$

- Expected emission count EC(h,c) of seeing c being emitted at state h is thus

$$\frac{\sum_{\{i:S[i]=c\}} \mathrm{T}(h, i)}{f(|S|+1, end)}$$

- One can then set $p(c \mid h) = \frac{EC(h,c)}{\sum_{\{c' \in \Sigma\}} EC(h,c')}$
- Derivation for $p(h \mid h')$ is left as an exercise



h_{start}
h'
x
h
y
z
h_{end}
0
1
2
$i-1$
i
n
1
0
0
0
0
0
$T(h,i) = \sum_{(h,h')\in T} f(i,h) * p(h'|h) * b(i+1,h')$
T(h,i)
b(i+1,h')
$v(i,h) = \mathbb{P}(s_i \mid h) \max(v(i-1,x)\mathbb{P}(h \mid x),$
$v(i-1,y)\mathbb{P}(h \mid y),$
$v(i-1,z)\mathbb{P}(h \mid z))$
f
$\sum$
f
f
f