



Bioinformatika

Hidden Markov Models

Michael Asplund

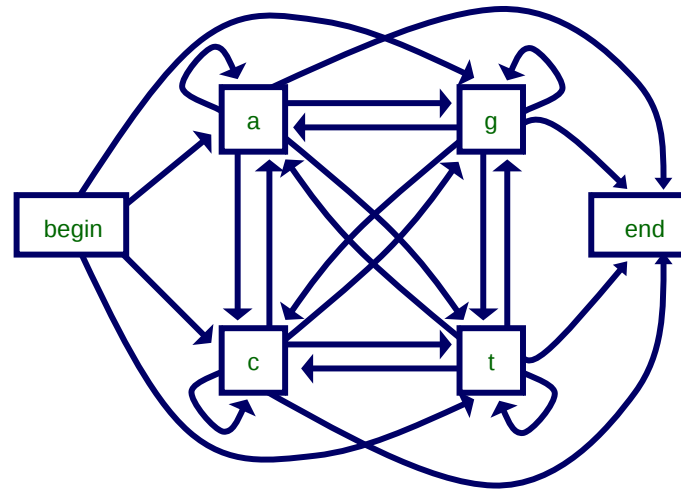
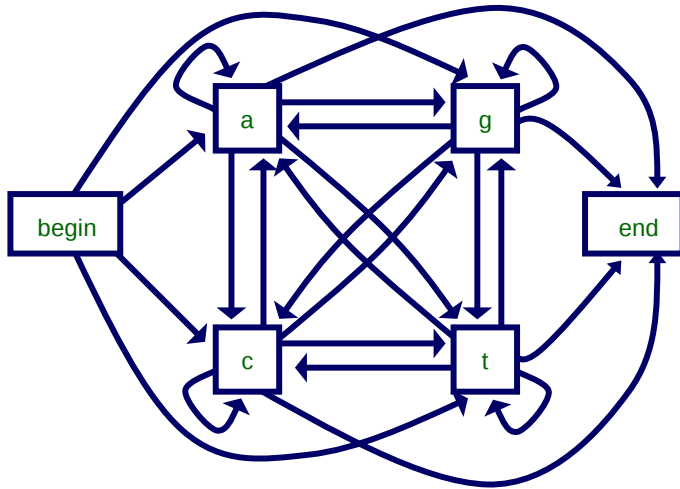
(some slides are courtesy of Mark Craven, U. of Wisconsin)

Motivation

- ↩ Sequence categorization into family of sequences (Forward alg.)
- ↩ Sequence anotation: CpG detection, gene finding (Viterbi alg.)
- ↩ Learning hidden parameters (Baum-Welsh alg.)

Motivation

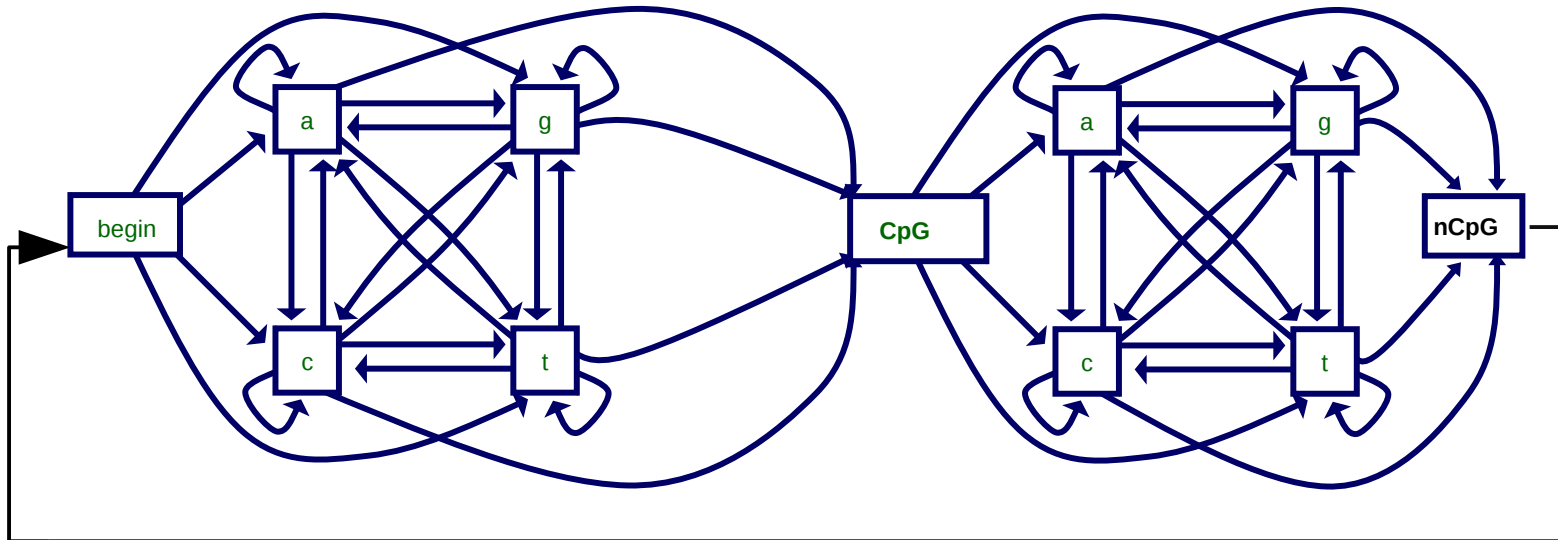
- Train two MMs: one to represent background sequence regions (*null*), another to represent CpG regions.



1. Given a test sequence, use two models to classify the sequence (*CpG* or *null*).
2. Given a test sequence, find CpG islands within. (?!)

Motivation

1. Train two MMs: one to represent background sequence regions (*null*), another to represent CpG regions.



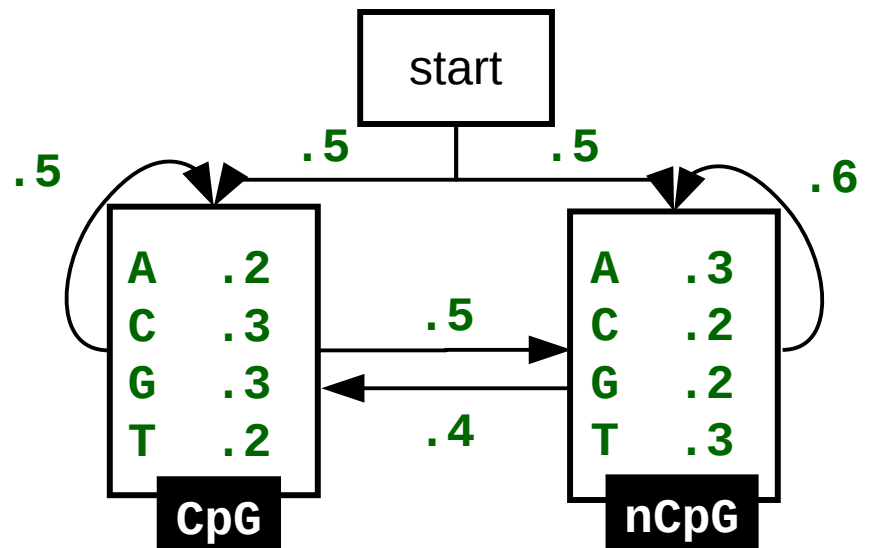
2. Join the 2 models into one HMM,
3. Segment given test sequence into CpG and non-CpG regions. **How?**

Viterbi algorithm

- ← Given an observed sequence \mathbf{x} .
- ← What is the most likely path \mathbf{s} through the model, i.e. sequence anotation?
- ← Ex: Naive model of CpG detection

$$\mathbf{s}^* = \arg \max_{s_0 \dots s_N \in S^N} p(x_0 \dots x_N; s_0 \dots s_N)$$

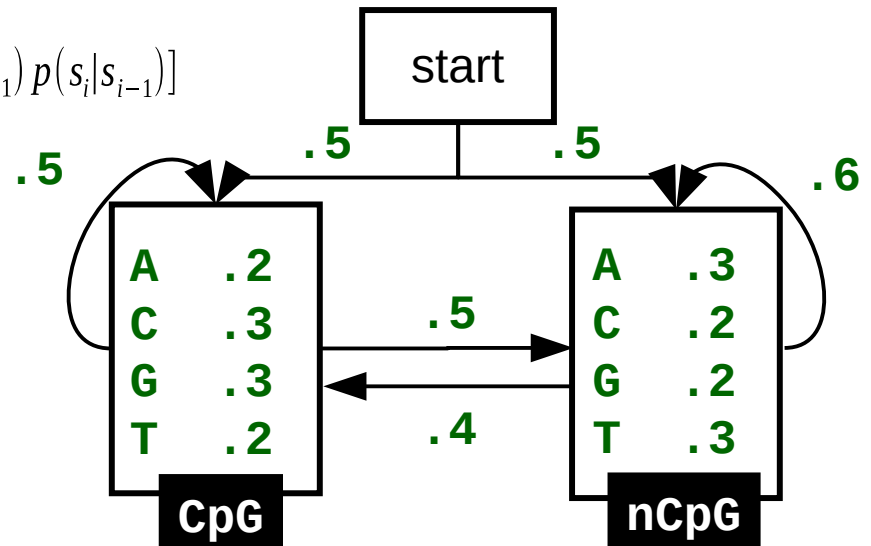
$$p(x_1 \dots x_N; s_1 \dots s_N) = \prod_{i=1}^N p(x_i | s_i) p(s_i | s_{i-1}),$$
$$p(s_0) = 1$$



Viterbi algorithm (ex.)

	ϵ	A	T	G	G	C	A	C	T	A
START	1	0	0	0	0	0	0	0	0	0
CpG	0									
nCpG	0									

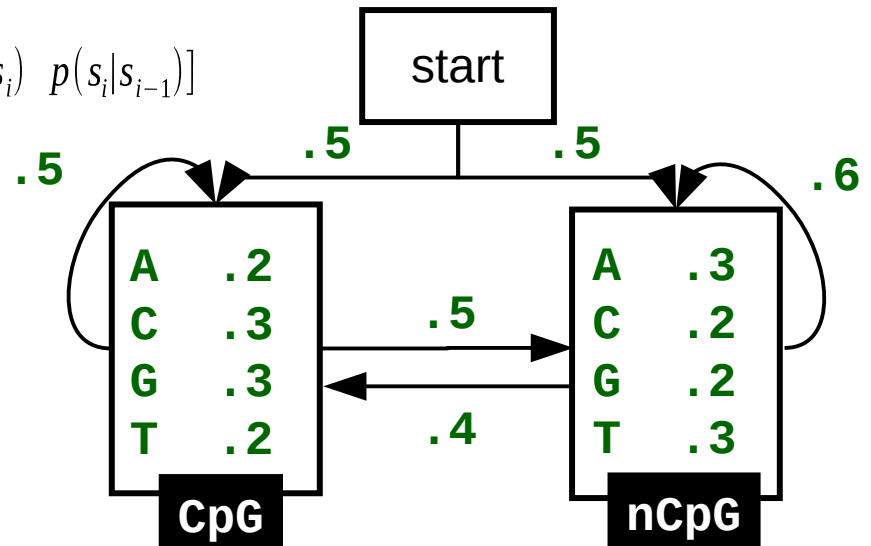
$$\max_{s_i \in S} p(x_0 \dots x_i | s_i) = \left[\max_{s_i \in S} p(x_i | s_i) \max_{s_{i-1} \in S} p(x_0 \dots x_{i-1} | s_{i-1}) p(s_i | s_{i-1}) \right]$$



Viterbi algorithm (ex.)

	ϵ	A	T	G	G	C	A	C	T	A
START	1	0	0	0	0	0	0	0	0	0
CpG	0	1 x .2 x .5 0 x .2 x .5 0 x .2 x .4 .1								
nCpG	0	1 x .3 x .5 0 x .3 x .5 0 .3 xx .6 .15								

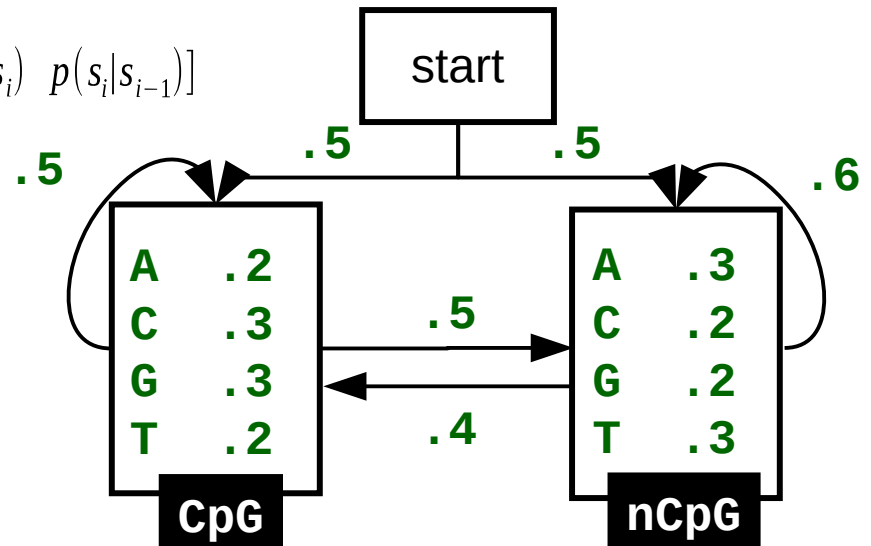
$$\max_{s_i \in S} p(x_0 \dots x_i | s_i) = \max_{s_{i-1} \in S} [p(x_0 \dots x_{i-1} | s_{i-1}) \max_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})]$$



Viterbi algorithm (ex.)

	ϵ	A	T	G	G	C	A	C	T	A
START	1	0	0	0	0	0	0	0	0	0
CpG	0	1 x .2 x .5 0 x .2 x .5 0 x .2 x .4 .1	0 x .2 x .5 .1 x .2 x .5 .15 x .2 x .4 .012							
nCpG	0	1 x .3 x .5 0 x .3 x .5 0 .3 x .6 .15	0 x .3 x .5 .1 x .3 x .5 .15 x .3 x .6 .027							

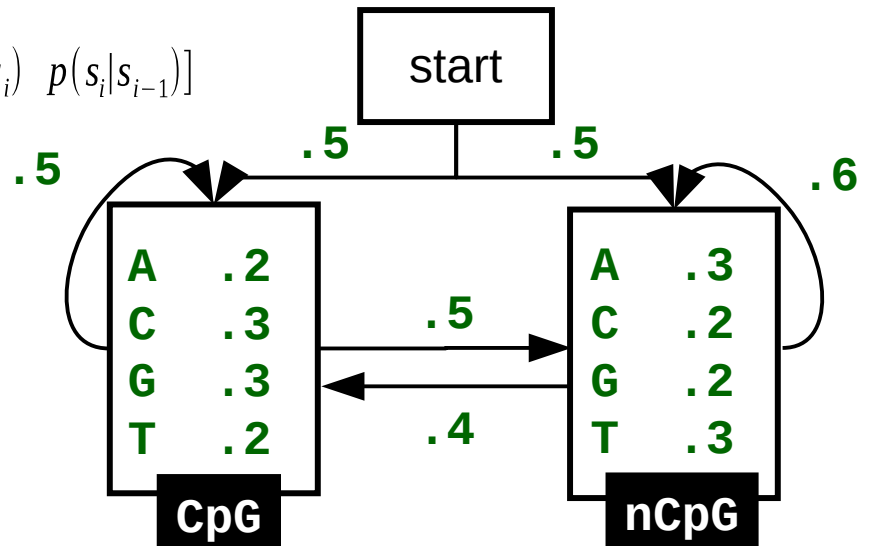
$$\max_{s_i \in S} p(x_0 \dots x_i | s_i) = \max_{s_{i-1} \in S} [p(x_0 \dots x_{i-1} | s_{i-1}) \max_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})]$$



Viterbi algorithm (ex.)

	ϵ	A	T	G	G	C	A	C	T	A
START	1	0	0	0	0	0	0	0	0	0
CpG	0	$1 \times .2 \times .5$ $0 \times .2 \times .5$ $0 \times .2 \times .4$ $.1$	$0 \times .2 \times .5$ $.1 \times .2 \times .5$ $.15 \times .2 \times .4$ $.012$	0	$.012 \times .3 \times .5$ $.027 \times .3 \times .4$ $.0032$	0 $.0032 \times .3 \times .5$ $.0032 \times .3 \times .4$ $5e-4$	0 $.012 \times .3 \times .5$ $.027 \times .3 \times .4$ $5e-5$			
nCpG	0	$1 \times .3 \times .5$ $0 \times .3 \times .5$ $0 \times .3 \times .6$ $.15$	$0 \times .3 \times .5$ $.1 \times .3 \times .5$ $.15 \times .3 \times .6$ $.027$	0	$.012 \times .2 \times .5$ $.027 \times .2 \times .6$ $.0032$	0 $.0032 \times .2 \times .5$ $.0032 \times .2 \times .6$ $4e-4$	0 $.012 \times .2 \times .5$ $.027 \times .2 \times .6$ $4e-5$			

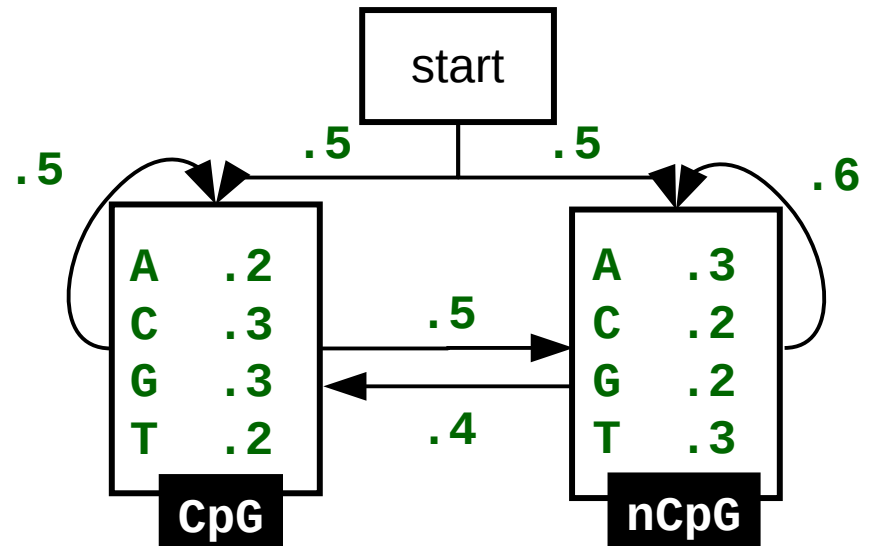
$$\max_{s_i \in S} p(x_0 \dots x_i | s_i) = \max_{s_{i-1} \in S} [p(x_0 \dots x_{i-1} | s_{i-1}) \max_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})]$$



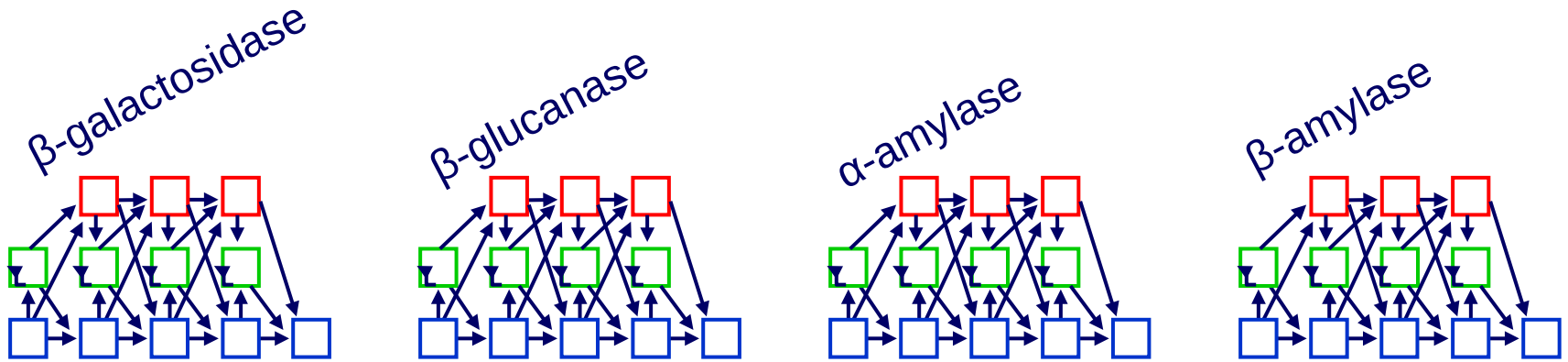
Viterbi algorithm (ex.)

	ϵ	A	T	G	G	C	A	C	T	A
START	0	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf	-inf
CpG	-inf	ln.2+0+ln.5 ln.2+ -inf +ln.5 ln.2+ -inf +ln.4 -2.30	ln.2+ -inf +ln.5 ln.2+0+ln.5 ln.2+ln.15+ln.4 -2.3							
nCpG	-inf	ln.3+0+ln.5 ln.3+ -inf +ln.5 ln.3+ -inf +ln.6 -1.9	-inf ln.3+ln.1+ln.5 ln.3+ln.15+ln.6 -1.9							

$$\arg \max_{s_i \in S} p(x_0 \dots x_i | s_i) = \arg \max_{s_i \in S} \log p(x_0 \dots x_i | s_i)$$



Forward algorithm



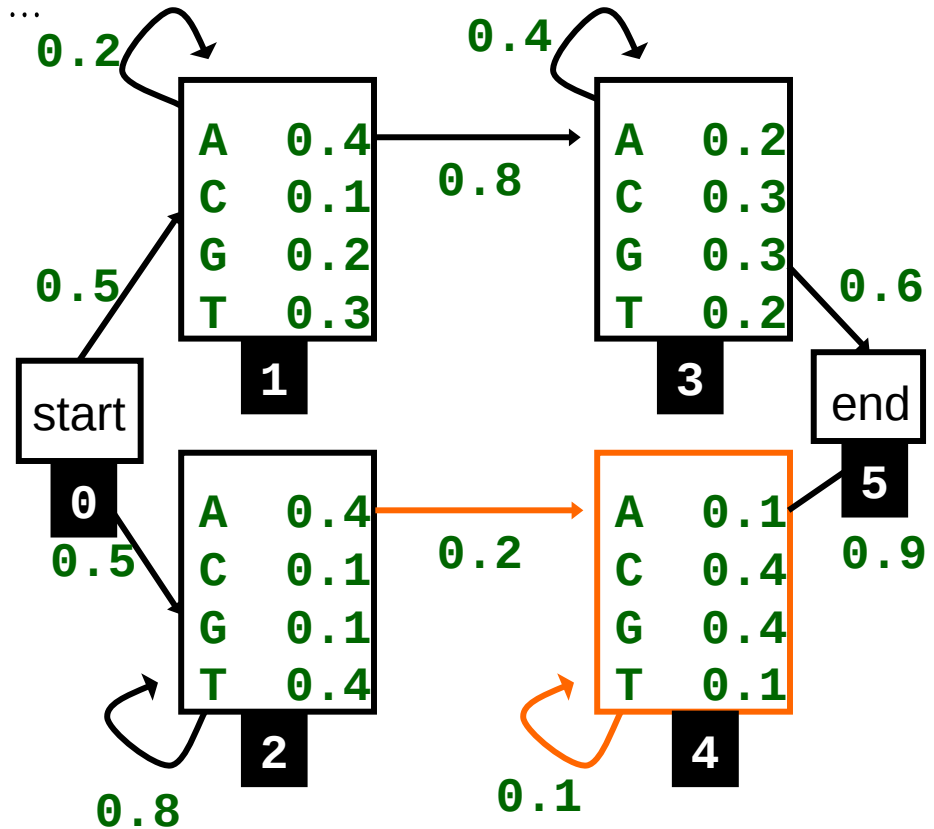
- ← Given **K** models of **K** sequence families.
- ← Categorize a new sequence **x**.

$$p(\alpha\text{-amyl.}) p(x_0 \dots x_N | \alpha\text{-amyl.}) < p(\beta\text{-amyl.}) p(x_0 \dots x_N | \beta\text{-amyl.})$$
$$p(x_0 \dots x_N | \alpha\text{-amyl.}) = \sum_{s_0 \dots s_N \in S^N} p(x_0 \dots x_N; s_0 \dots s_N | \alpha\text{-amyl.})$$

Forward algorithm

$$\begin{aligned}
 p(x_0 x_1 \dots x_{N-1} x_N) &= \sum_{s_0 s_1 \dots s_{N-1} s_N \in S^N} p(x_0 x_1 \dots x_{N-1} x_N; s_0 s_1 \dots s_{N-1} s_N) = \dots \\
 &\sum_{s_0 s_1 \dots s_{N-1} s_N \in S^N} p(x_0 | s_0) p(s_1 | s_0) p(x_1 | s_1) \dots p(s_N | s_{N-1}) p(x_N | s_N) = \dots \\
 &\sum_{s_0 s_1 \dots s_{N-1} \in S^{N-1}} \dots \sum_{s_N \in S} p(s_N | s_{N-1}) p(x_N | s_N) = \dots \\
 &\sum_{s_0 s_1 \dots s_{N-1} \in S^{N-1}} p(x_0 \dots x_{N-1}; s_0 \dots s_{N-1}) \sum_{s_N \in S} p(s_N | s_{N-1}) p(x_N | s_N)
 \end{aligned}$$

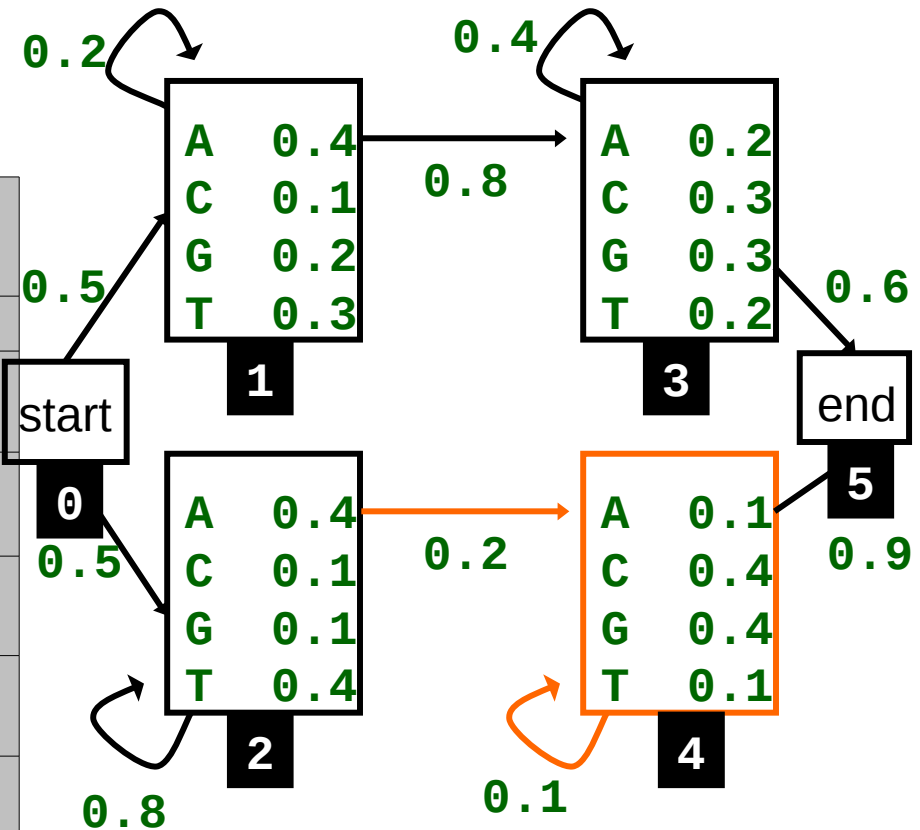
$$\begin{aligned}
 &\sum_{s_0 s_1 \dots s_{i-1} \in S^{i-1}} p(x_0 x_1 \dots x_{i-1}; s_0 s_1 \dots s_{i-1}) = \dots \\
 p(x_0 \dots x_i) &= p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})
 \end{aligned}$$



Forward algorithm (ex.)

$$p(x_0 \dots x_i) = p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})$$

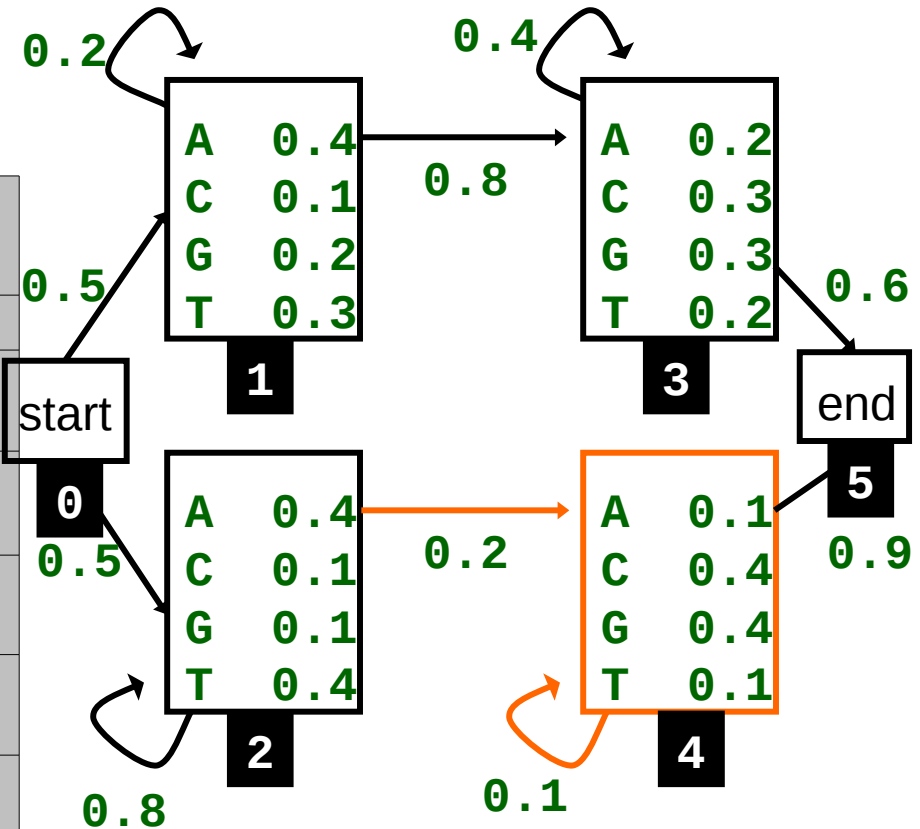
	ϵ	T	A	G	A	ϵ
0	1	0	0	0	0	
1	0					
2	0					
3						
4						
5		0	0	0	0	



Forward algorithm (ex.)

$$p(x_0 \dots x_i) = p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})$$

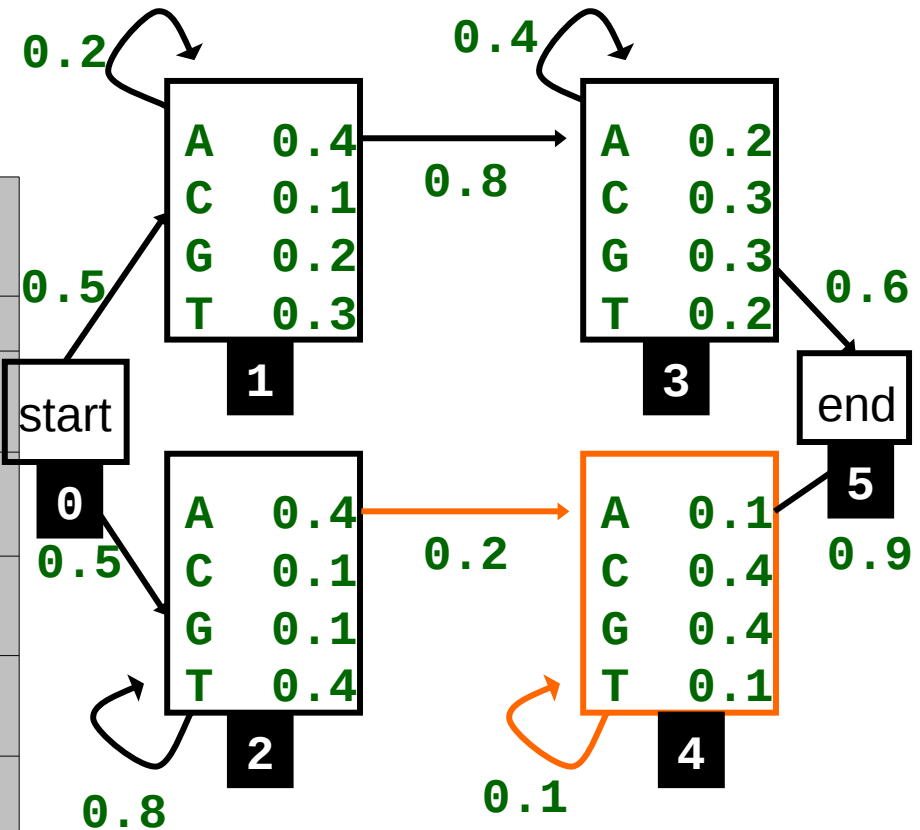
	ϵ	T	A	G	A	ϵ
0	1	0	0	0	0	
1	0	$\begin{matrix} 1 \times .3 \times .5 \\ 0 \times .3 \times .2 \\ .15 \end{matrix}$				
2	0	$\begin{matrix} 1 \times .4 \times .5 \\ 0 \times .4 \times .8 \\ .2 \end{matrix}$				
3		0				
4		0				
5		0	0	0	0	



Forward algorithm (ex.)

$$p(x_0 \dots x_i) = p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})$$

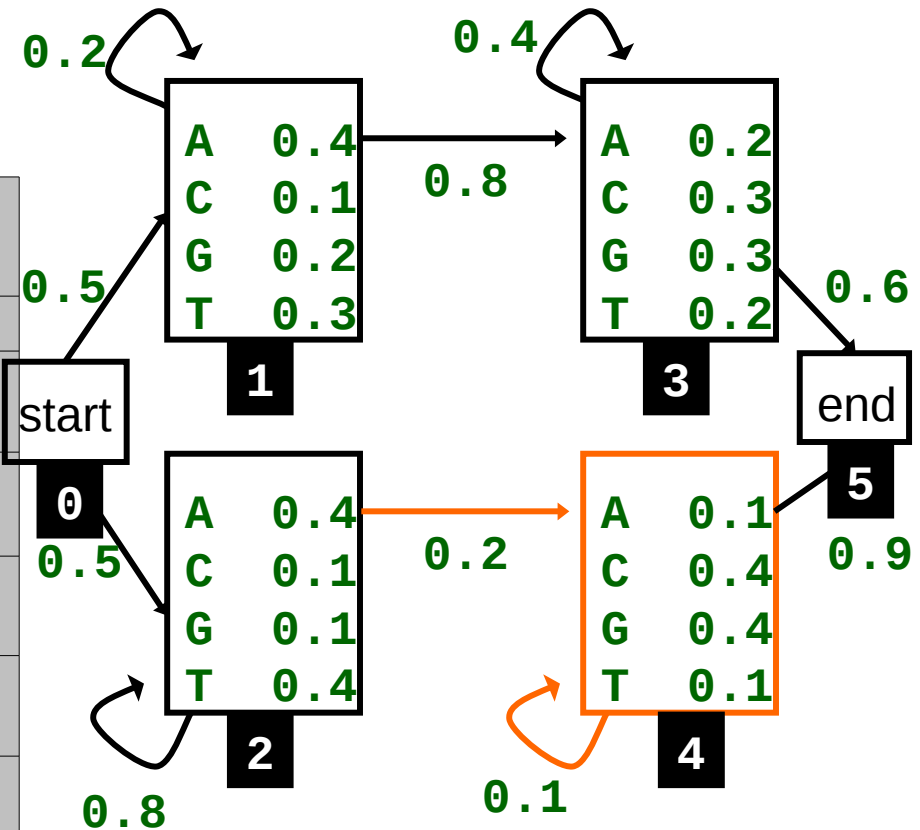
	ϵ	T	A	G	A	ϵ
0	1	0	0	0	0	
1	0	$1 \times .3 \times .5$ $0 \times .3 \times .2$.15	$0 \times .4 \times .5$ $.15 \times .4 \times .2$.012			
2	0	$1 \times .4 \times .5$ $0 \times .4 \times .8$.2	$0 \times .4 \times .5$ $.2 \times .4 \times .8$.064			
3		0	$.15 \times .2 \times .8$ $0 \times .2 \times .4$.024			
4		0	$.2 \times .1 \times .2$ $0 \times .1 \times .1$.004			
5		0	0	0	0	



Forward algorithm (ex.)

$$p(x_0 \dots x_i) = p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})$$

	ϵ	T	A	G	A	ϵ
0	1	0	0	0	0	
1	0	$1 \times .3 \times .5$ $0 \times .3 \times .2$.15	$0 \times .4 \times .5$ $.15 \times .4 \times .2$.012	$0 \times .2 \times .5$ $.012 \times .2 \times .2$ 5e-4		
2	0	$1 \times .4 \times .5$ $0 \times .4 \times .8$.2	$0 \times .4 \times .5$ $.2 \times .4 \times .8$.064	$0 \times .1 \times .5$ $.064 \times .1 \times .8$.00512		
3		0	$.15 \times .2 \times .8$ $0 \times .2 \times .4$.024	$.012 \times .3 \times .8$ $.024 \times .3 \times .4$.00576		
4		0	$.2 \times .1 \times .2$ $0 \times .1 \times .1$.004	$.064 \times .4 \times .2$ $.004 \times .4 \times .1$.00528	.	
5		0	0	0	0	



Forward algorithm (ex.)

$$p(x_0 \dots x_i) = p(x_0 \dots x_{i-1}) \sum_{s_i \in S} p(x_i | s_i) p(s_i | s_{i-1})$$

	ϵ	T	A	G	A	ϵ
0	1	0	0	0	0	
1	0	$1 \times .3 \times .5$ $0 \times .3 \times .2$.15	$0 \times .4 \times .5$ $.15 \times .4 \times .2$.012	$0 \times .2 \times .5$ $.012 \times .2 \times .2$ $5e-4$	$0 \times .4 \times .5$ $5e-4 \times .4 \times .2$ $4e-5$	0
2	0	$1 \times .4 \times .5$ $0 \times .4 \times .8$.2	$0 \times .4 \times .5$ $.2 \times .4 \times .8$.064	$0 \times .1 \times .5$ $.064 \times .1 \times .8$.00512	$0 \times .4 \times .5$ $5e-3 \times .4 \times .8$.0016	0
3		0	$.15 \times .2 \times .8$ $0 \times .2 \times .4$.024	$.012 \times .3 \times .8$ $.024 \times .3 \times .4$.00576	$5e-4 \times .2 \times .8$ $6e-3 \times .2 \times .4$ $6e-4$	0
4		0	$.2 \times .1 \times .2$ $0 \times .1 \times .1$.004	$.064 \times .4 \times .2$ $.004 \times .4 \times .1$.00528	$.005 \times .1 \times .2$ $.005 \times .1 \times .1$ $1.5e-4$	0
5		0	0	0	0	$6e-4 \times .6$ $1.5e-4 \times .9$ $4.6e-4$

