Bioinformatika
Hidden Markov Models

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

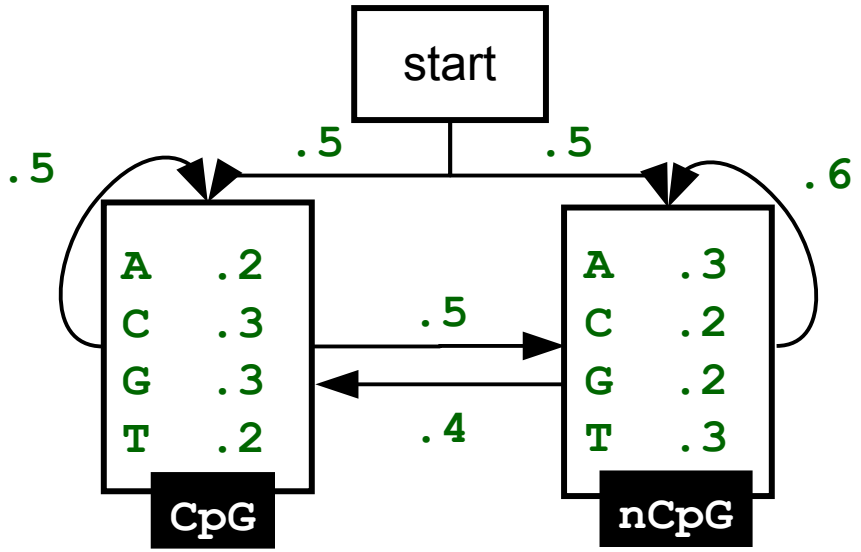
Motivation

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

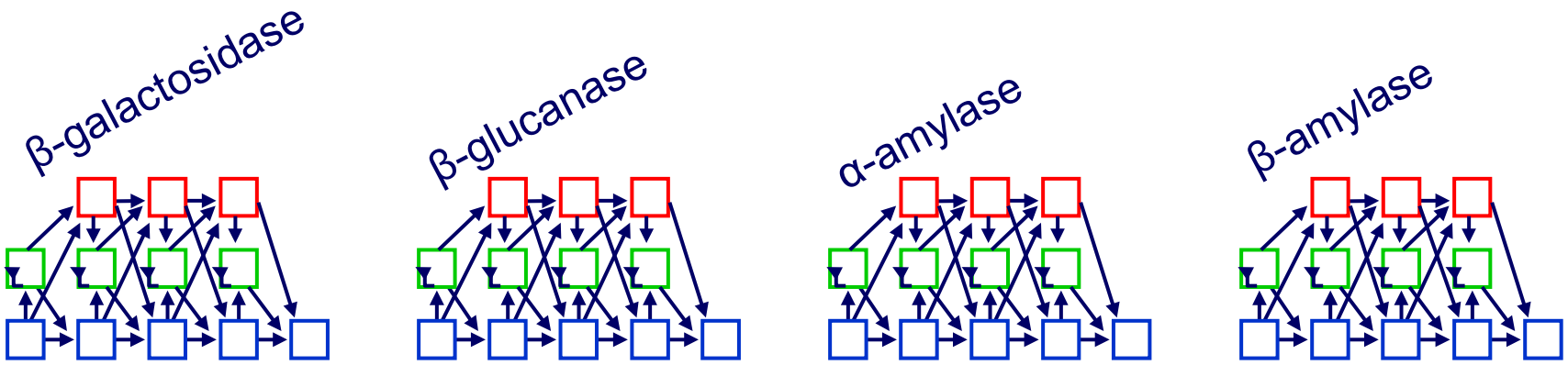
Notation

- Ex: Naïve model of CpG detection

$$p(x_i \ldots x_N ; s_i \ldots s_N) = p(s_0)\, p(x_0|s_0) \prod_{i=1}^{N} p(x_i|s_i)\, p(s_i|s_{i-1})$$



Forward algorithm



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

$$p(\alpha-amyl.)\,p(x_0\ldots x_N|\alpha-amyl.) \;<\; p(\beta-amyl.)\,p(x_0\ldots x_N|\beta-amyl.)$$

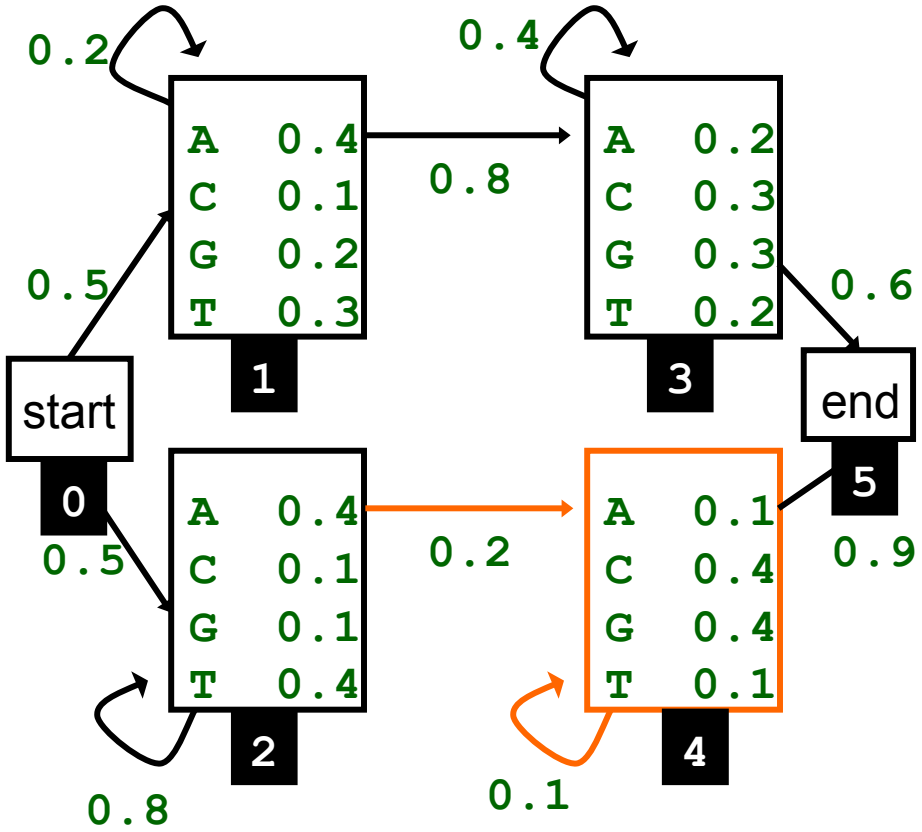
$$p(x_0\ldots x_N|\alpha-amyl.) \;=\; \sum_{s_0\ldots s_N\in S^N} p(x_0\ldots x_N; s_0\ldots s_N|\alpha-amyl.)$$

Forward algorithm

$$p(x_0 x_1 \ldots x_{N-1} x_N) = \sum_{s_0 s_1 \ldots s_{N-1} s_N \in S^N} p(x_0 x_1 \ldots x_{N-1} x_N ; s_0 s_1 \ldots s_{N-1} s_N)$$

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

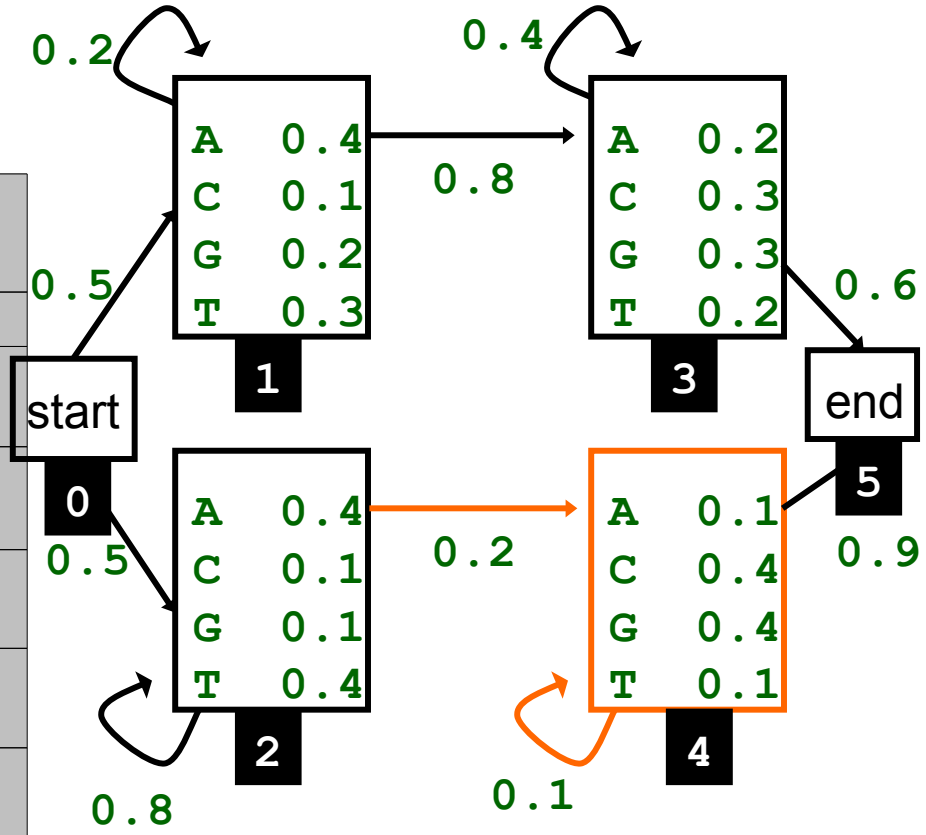
$$p(x_0 \ldots x_i) = \sum_{s_i \in S} p(x_i | s_i) p(x_0 \ldots x_{i-1}) p(s_i | s_{i-1})$$



Forward algorithm (ex.)

$$\sum_{s_0 \in S} p(\epsilon|s) = 1 \quad \text{if} \quad s_0 = \text{START else: } \sim 0$$

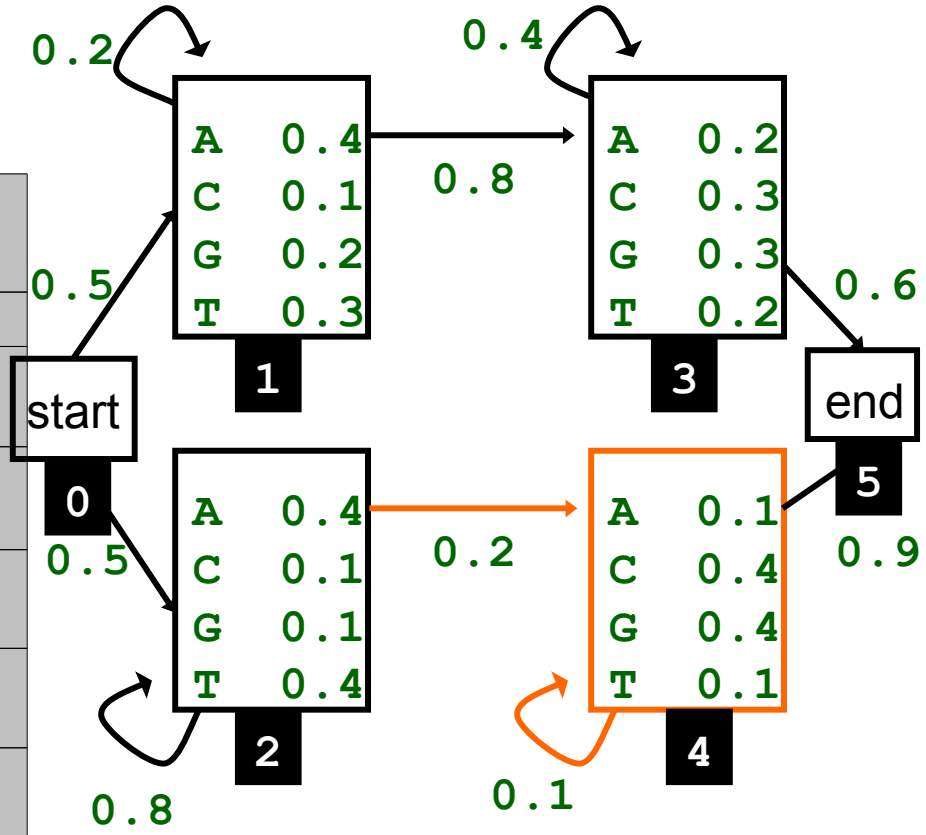
	ε	T	A	G	A	ε
0	1	0	0	0	0	0
1	0					
2	0					
3						
4						
5						



Forward algorithm (ex.)

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

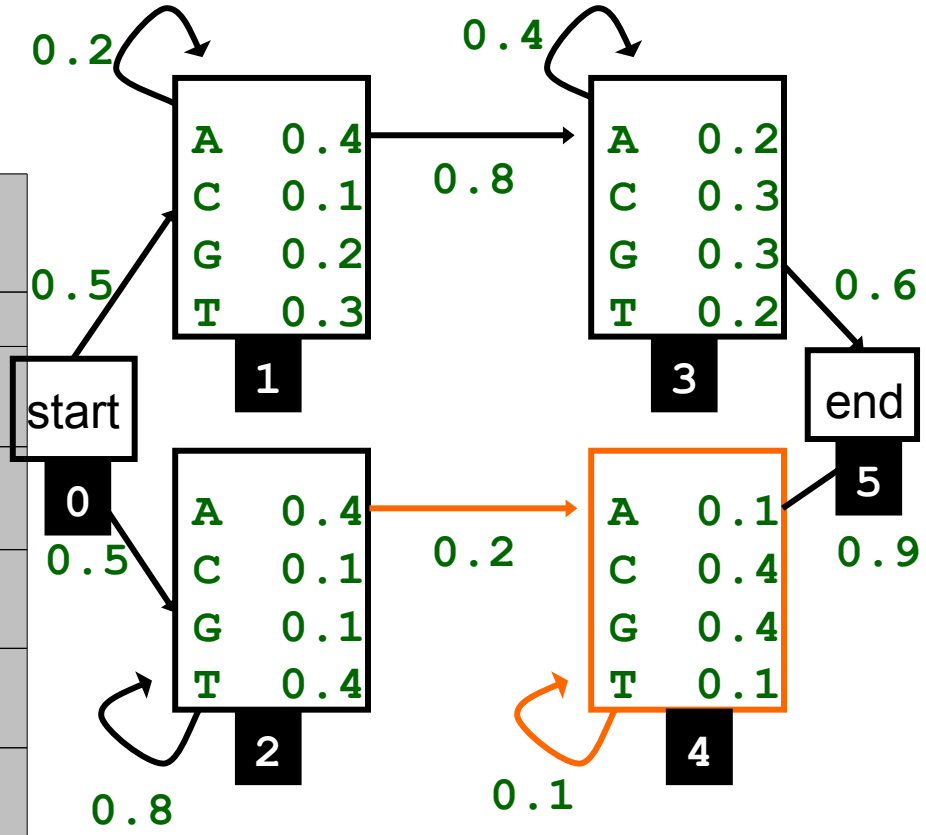
	ε	T	A	G	A	ε
0	1	0	0	0	0	0
1	0	.3 x **1** x .5 .3 x 0 x .2 .15				
2	0	.4 x **1** x .5 .4 x 0 x .8 .2				
3		0				
4		0				
5		0				



Forward algorithm (ex.)

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

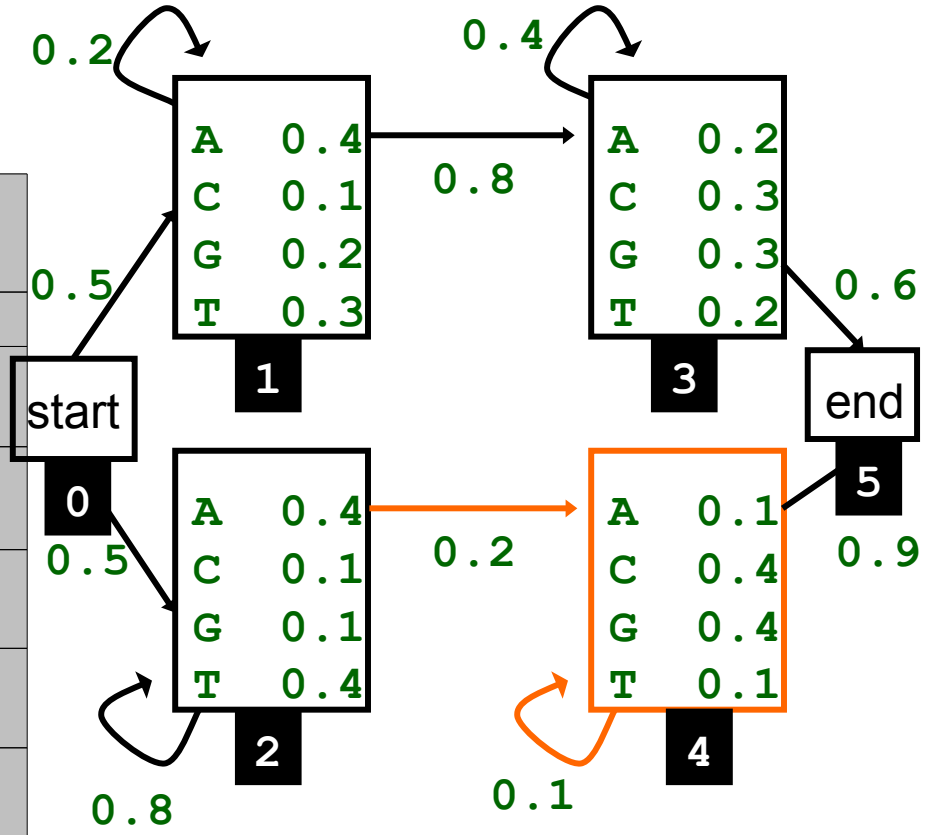
	ε	T	A	G	A	ε
0	1	0	0	0	0	0
1	0	.3 x **1** x .5 .3 x 0 x .2 .15	.4 x **0** x .5 .4 x .15 x .2 .012			
2	0	.4 x **1** x .5 .4 x 0 x .8 .2	.4 x **0** x .5 .4 x .2 x .8 .064			
3		0	.2 x .15 x .8 .2 x 0 x .4 .024			
4		0	.1 x .2 x .2 .1 x 0 x .1 .004			
5		0	0			



Forward algorithm (ex.)

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

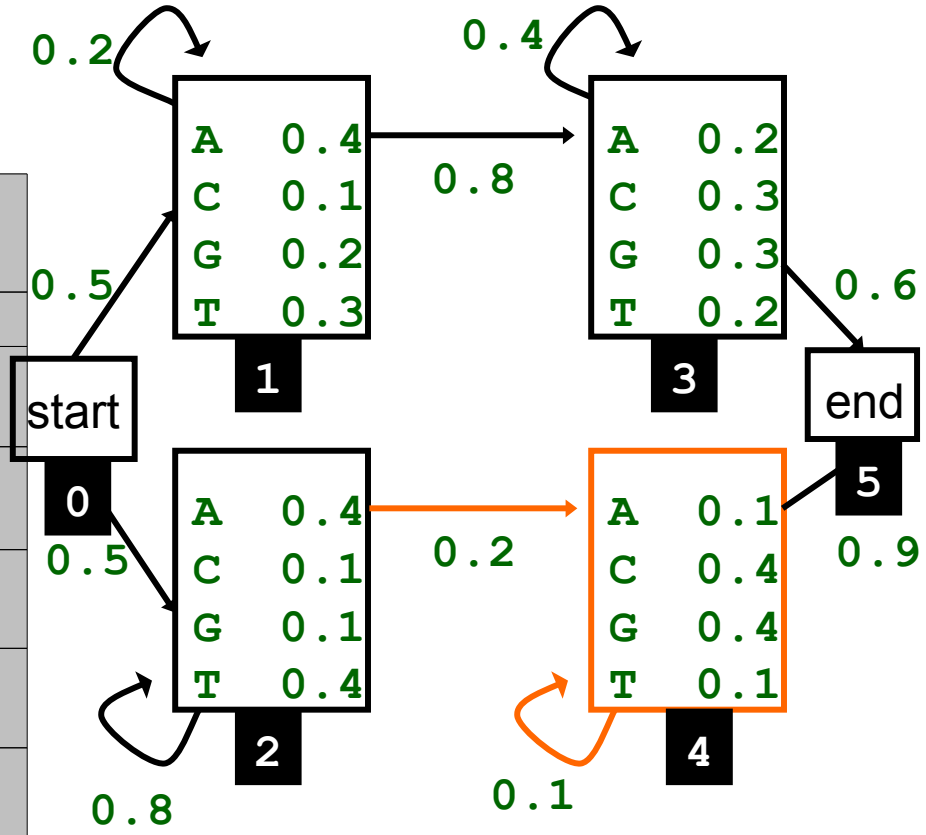
	ε	T	A	G	A	ε
0	1	0	0	0	0	
1	0	.3 x **1** x .5 .3 x 0 x .2 .15	.4 x **0** x .5 .4 x .15 x .2 .012	.2 x **0** x .5 .2 x .012 x .2 5e-4	.4 x **0** x .5 .4 x 5e-4 x .2 4e-5	
2	0	.4 x **1** x .5 .4 x 0 x .8 .2	.4 x **0** x .5 .4 x .2 x .8 .064	.1 x **0** x .5 .1 x .2 x .8 .016	.4 x **0** x .5 .4 x .016 x .8 .005	
3		0	.2 x .15 x .8 .2 x 0 x .4 .024	.3 x .012 x .8 .3 x .024 x .4 6e-4	.2 x 5e-4 x .8 .2 x 6e-4 x .4 1e-4	
4		0	.1 x .2 x .2 .1 x 0 x .1 .004	.4 x .064 x .2 .4 x .004 x .1 .005	.1 x .016 x .2 .1 x .005 x .1 4e-4	
5		0	0	0	0	



Forward algorithm (ex.)

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

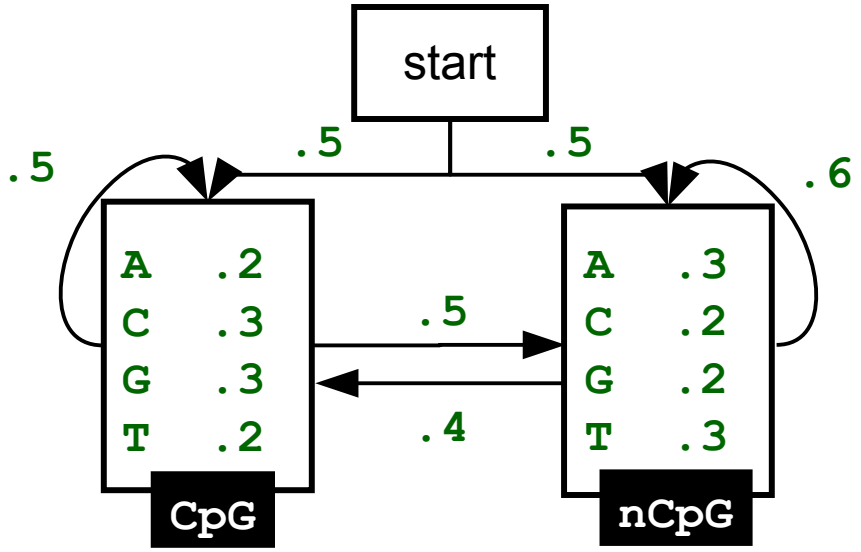
	ε	T	A	G	A	ε
0	1	0	0	0	0	0
1	0	.3 x **1** x .5 .3 x 0 x .2 .15	.4 x **0** x .5 .4 x .15 x .2 .012	.2 x **0** x .5 .2 x .012 x .2 5e-4	.4 x **0** x .5 .4 x 5e-4 x .2 4e-5	0
2	0	.4 x **1** x .5 .4 x 0 x .8 .2	.4 x **0** x .5 .4 x .2 x .8 .064	.1 x **0** x .5 .1 x .2 x .8 .016	.4 x **0** x .5 .4 x .016 x .8 .005	0
3		0	.2 x .15 x .8 .2 x 0 x .4 .024	.3 x .012 x .8 .3 x .024 x .4 6e-4	.2 x 5e-4 x .8 .2 x 6e-4 x .4 1e-4	0
4		0	.1 x .2 x .2 .1 x 0 x .1 .004	.4 x .064 x .2 .4 x .004 x .1 .005	.1 x .016 x .2 .1 x .005 x .1 4e-4	0
5		0	0	0	0	1e-4 x .6 4e-4 x .9 **4.2e-4**



Viterbi algorithm

- Given an observed sequence ***x.***
- What is the most likely path **s** through the model, i.e. sequence anotation?

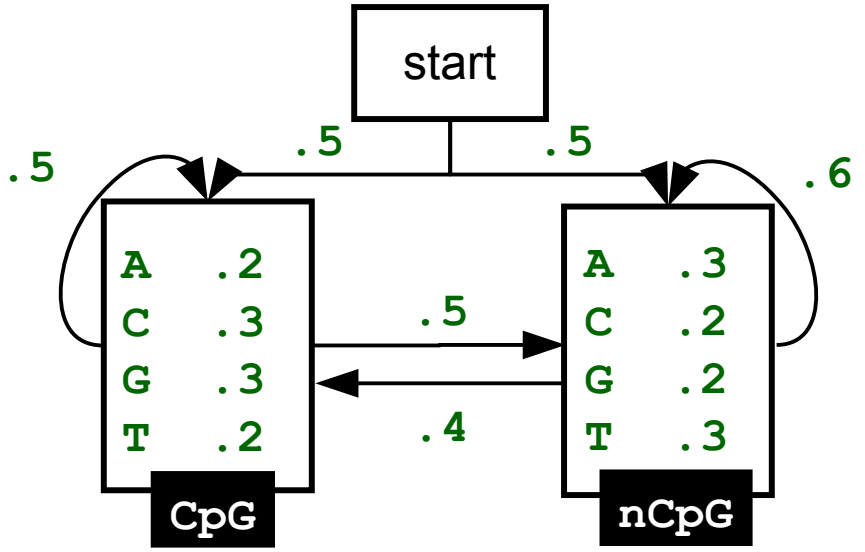
$$s^* = \arg\max_{s_0 \ldots s_N \in S^N} p(x_0 \ldots x_N ; s_0 \ldots s_N)$$



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									

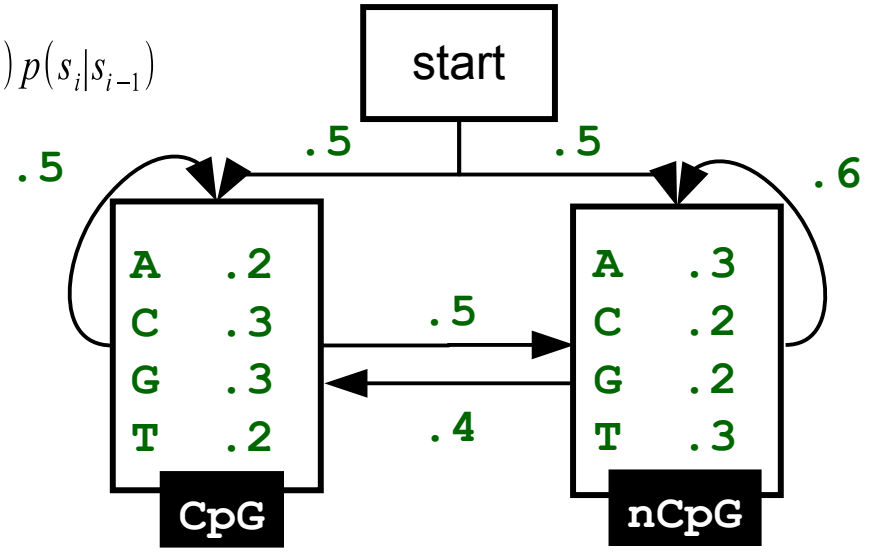
$$\max p(\epsilon|s_0) = 1 \quad \text{if} \quad s_0 = \text{START else: } \sim 0$$



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	.2 x **1** x .5 .2 x 0 x .5 .2 x 0 x .4 .1								
nCpG	0	.3 x **1** x .5 .3 x 0 x .5 .3 x 0 x .6 .15								

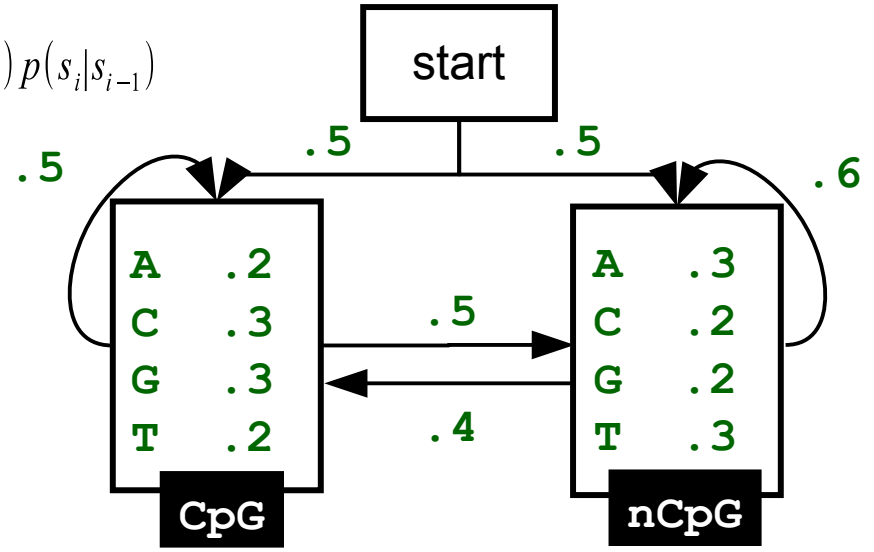
$$\max_{s_i \in S} p(x_0 \ldots x_i | s_i) = \max_{s_i \in S} p(x_i | s_i) \max_{s_{i-1} \in S} p(x_0 \ldots x_{i-1} | s_{i-1}) p(s_i | s_{i-1})$$



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	.2 x **1** x .5 .2 x 0 x .5 .2 x 0 x .4 .1	.2 x **0** x .5 .2 x .1 x .5 .2 x .15 x .4 .012							
nCpG	0	.3 x **1** x .5 .3 x 0 x .5 .3 x 0 x .6 .15	.3 x **0** x .5 .3 x .1 x .5 .3 x .15 x .6 .027							

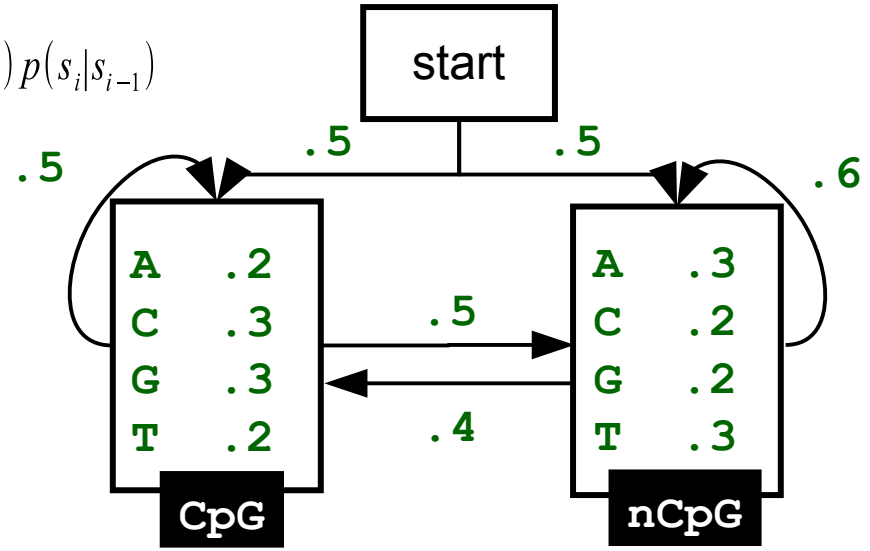
$$\max_{s_i \in S} p(x_0 \ldots x_i | s_i) = \max_{s_i \in S} p(x_i | s_i) \max_{s_{i-1} \in S} p(x_0 \ldots x_{i-1} | s_{i-1}) p(s_i | s_{i-1})$$



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	.2 x **1** x .5 .2 x 0 x .5 .2 x 0 x .4 .1	.2 x **0** x .5 .2 x .1 x .5 .2 x .15 x .4 .012	0 .3 x .012 x .5 .3 x .027 x .4 .0032	0 .3 x .0032 x .5 .3 x .0032 x .4 5e-4	0 .3 x .012 x .5 .3 x .027 x .4 5e-5				
nCpG	0	.3 x **1** x .5 .3 x 0 x .5 .3 x 0 x .6 .15	.3 x **0** x .5 .3 x .1 x .5 .3 x .15 x .6 .027	0 .2 x .012 x .5 .2 x .027 x .6 .0032	0 .2 x .0032 x .5 .2 x .0032 x .6 4e-4	0 .2 x .012 x .5 .2 x .027 x .6 4e-5				

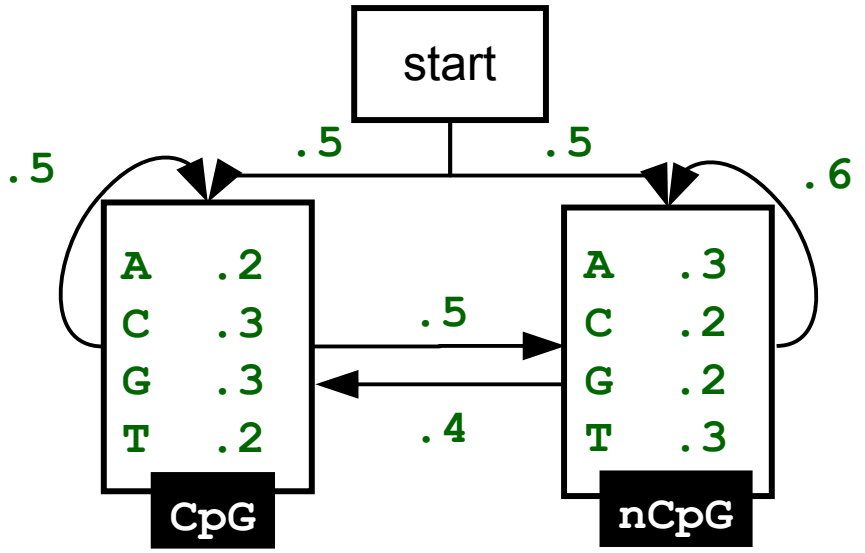
$$\max_{s_i \in S} p(x_0 \ldots x_i | s_i) = \max_{s_i \in S} p(x_i | s_i) \max_{s_{i-1} \in S} p(x_0 \ldots x_{i-1} | s_{i-1}) p(s_i | s_{i-1})$$



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							

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



Assignement

- Gene finding

http://www.biostat.wisc.edu/~craven/776/hw3.html

- Need not implement viterb algorithm. You can use an arbitrary solver.
- Alternative to gene finding assignement:
  - Viterbi and Forward implementation (or gene expression)
  - Sequence assembly