Bioinformatika
Hidden Markov Models

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

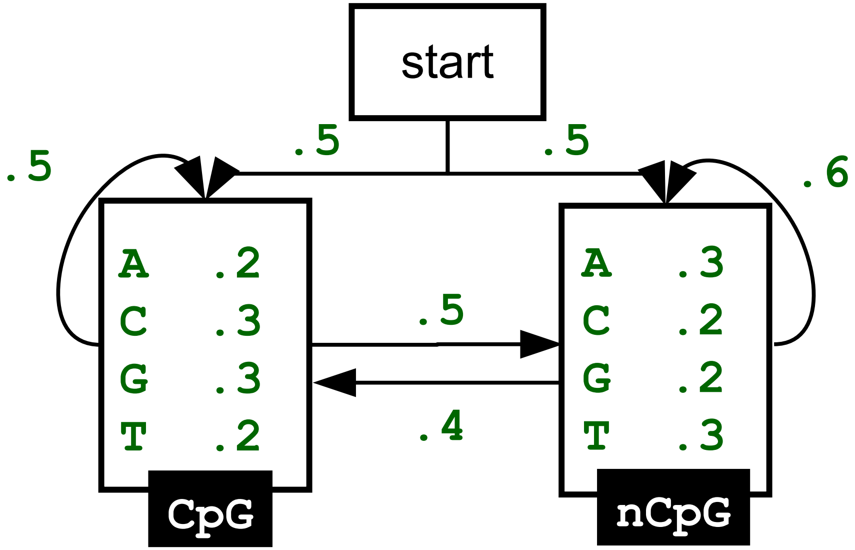
Motivation

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

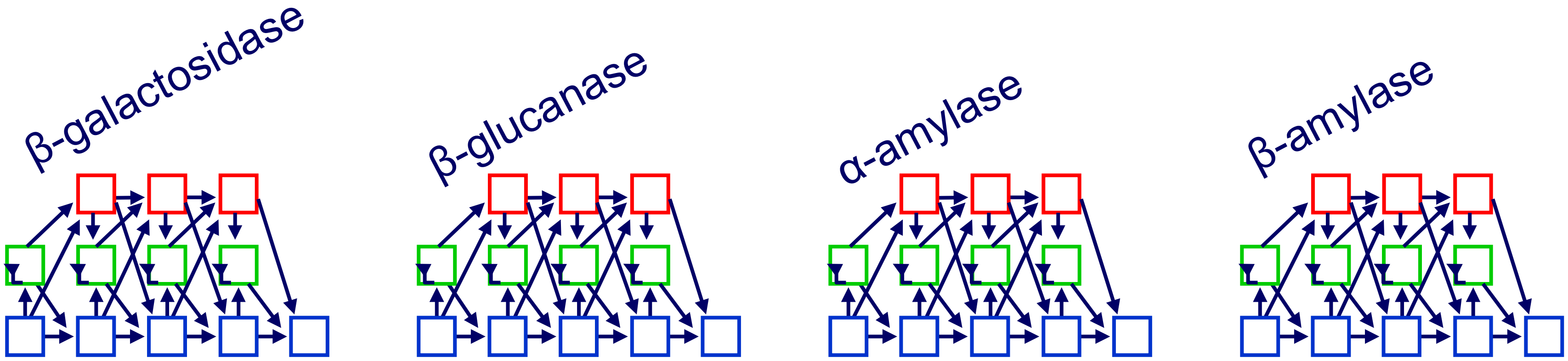
Notation

- Ex: Naïve model of CpG detection

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



Forward algorithm



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

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

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

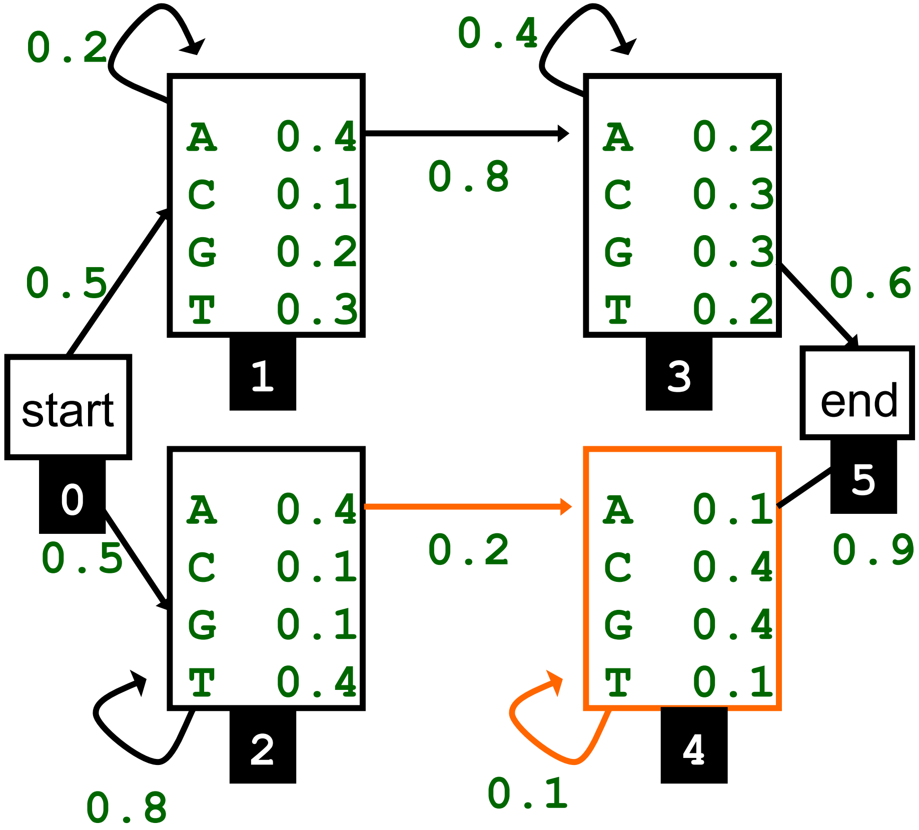
Forward algorithm

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

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

$$\ldots \sum_{s_N \in S} p(s_N | s_{N-1})\, p(x_N | s_N)$$

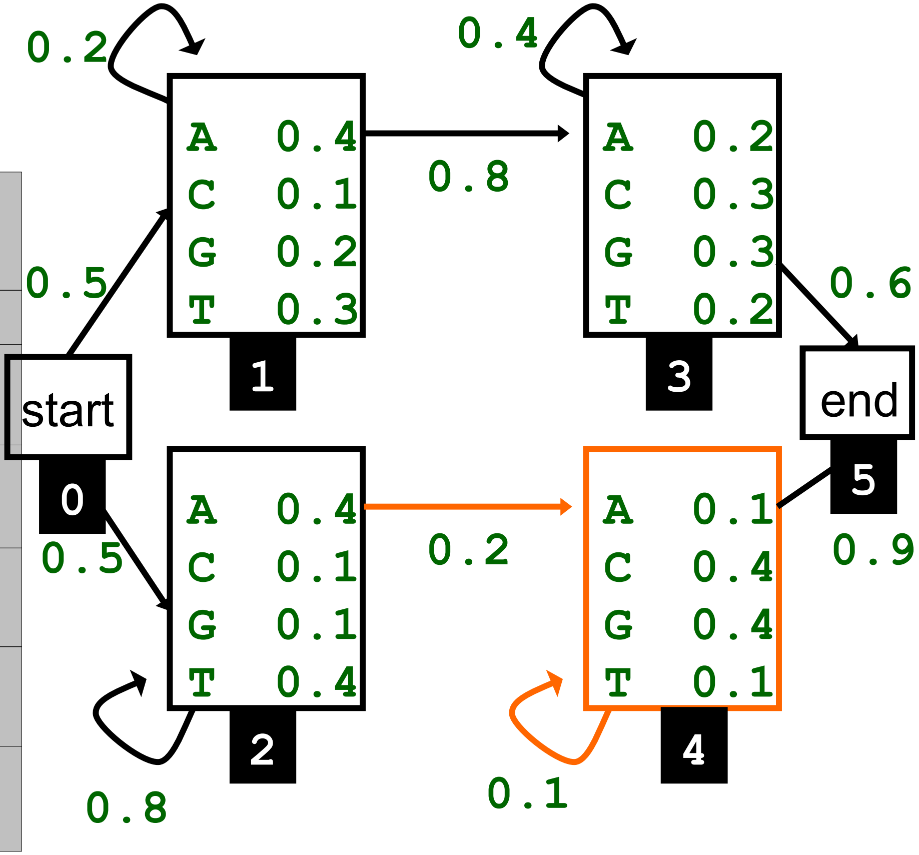
$$\boxed{p(x_0 \ldots x_i) = \sum_{s_i \in S} p(x_i | s_i)\, p(x_0 \ldots x_{i-1})\, p(s_i | s_{i-1})}$$



Forward algorithm (ex.)

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

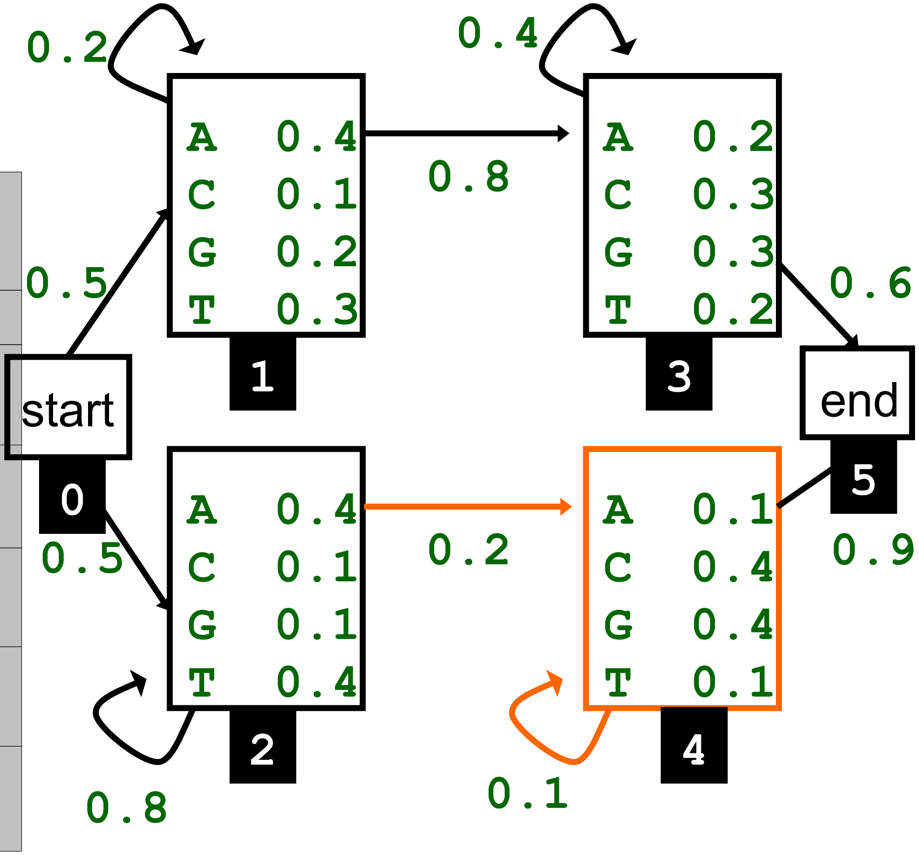
	ε	T	A	G	A	ε
0	1	0	0	0	0	0
1	0					
2	0					
3						
4						
5						



Forward algorithm (ex.)

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

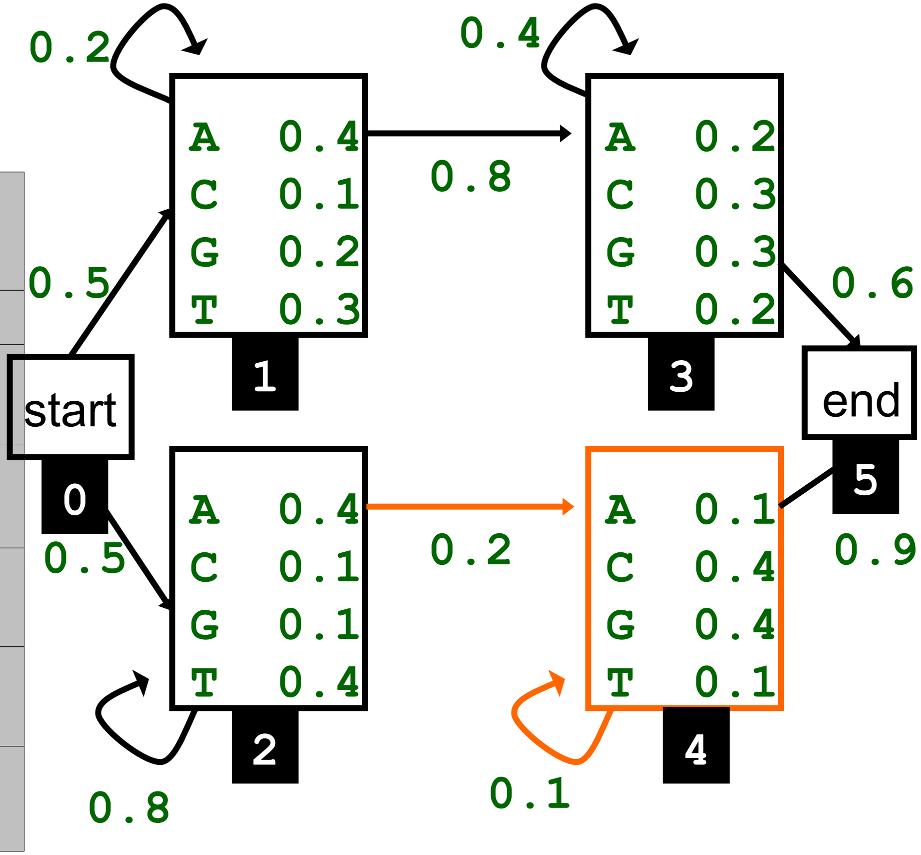
	ε	T	A	G	A	ε
0	**1**	0	0	0	0	0
1	0	.3 x **1** x .5 .3 x 0 x .2 .15				
2	0	.4 x **1** x .5 .4 x 0 x .8 .2				
3		0				
4		0				
5		0				



Forward algorithm (ex.)

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

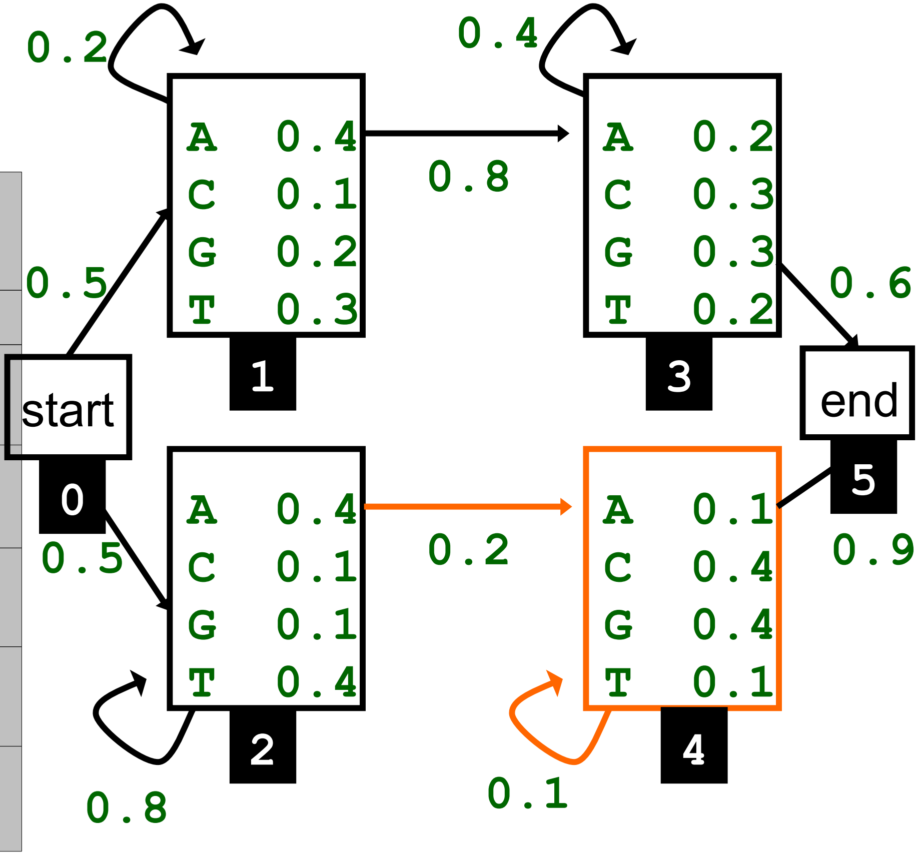
	ε	T	A	G	A	ε
0	**1**	0	0	0	0	0
1	0	.3 x **1** x .5 .3 x 0 x .2 .15	.4 x **0** x .5 .4 x .15 x .2 .012			
2	0	.4 x **1** x .5 .4 x 0 x .8 .2	.4 x **0** x .5 .4 x .2 x .8 .064			
3		0	.2 x .15 x .8 .2 x 0 x .4 .024			
4		0	.1 x .2 x .2 .1 x 0 x .1 .004			
5		0	0			



Forward algorithm (ex.)

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

	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 .064 x .8 .00512	.4 x **0** x .5 .4 x 5e-3 x .8 .0016	
3		0	.2 x .15 x .8 .2 x 0 x .4 .024	.3 x .012 x .8 .3 x .024 x .4 .00576	.2 x 5e-4 x .8 .2 x 6e-3 x .4 6e-4	
4		0	.1 x .2 x .2 .1 x 0 x .1 .004	.4 x .064 x .2 .4 x .004 x .1 .00528	.1 x .005 x .2 .1 x .005 x .1 1.5e-4	
5		0	0	0	0	



start 0

0.5

1
A 0.4
C 0.1
G 0.2
T 0.3

0.2

0.8

3
A 0.2
C 0.3
G 0.3
T 0.2

0.4

0.6

end 5

0.5

2
A 0.4
C 0.1
G 0.1
T 0.4

0.8

0.2

4
A 0.1
C 0.4
G 0.4
T 0.1

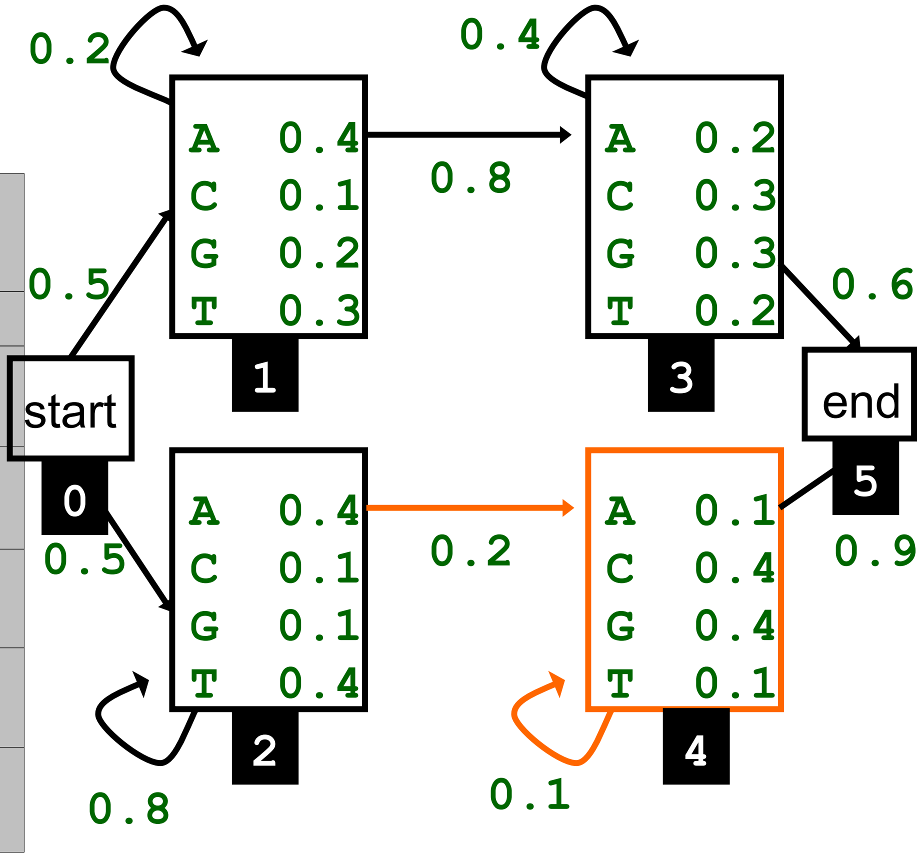
0.1

0.9

Forward algorithm (ex.)

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

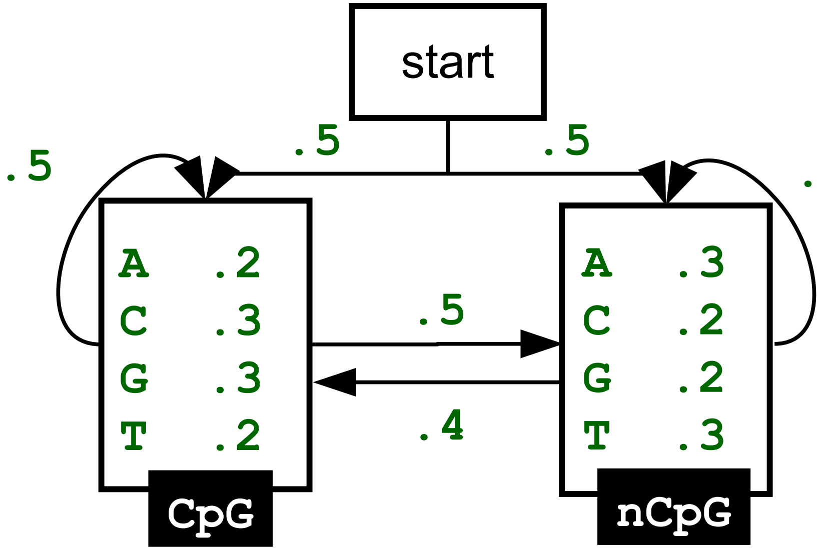
	ε	T	A	G	A	ε
0	**1**	0	0	0	0	
1	0	.3 x **1** x .5 .3 x 0 x .2 .15	.4 x **0** x .5 .4 x .15 x .2 .012	.2 x **0** x .5 .2 x .012 x .2 5e-4	.4 x **0** x .5 .4 x 5e-4 x .2 4e-5	0
2	0	.4 x **1** x .5 .4 x 0 x .8 .2	.4 x **0** x .5 .4 x .2 x .8 .064	.1 x **0** x .5 .1 x .064 x .8 .00512	.4 x **0** x .5 .4 x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



Viterbi algorithm

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

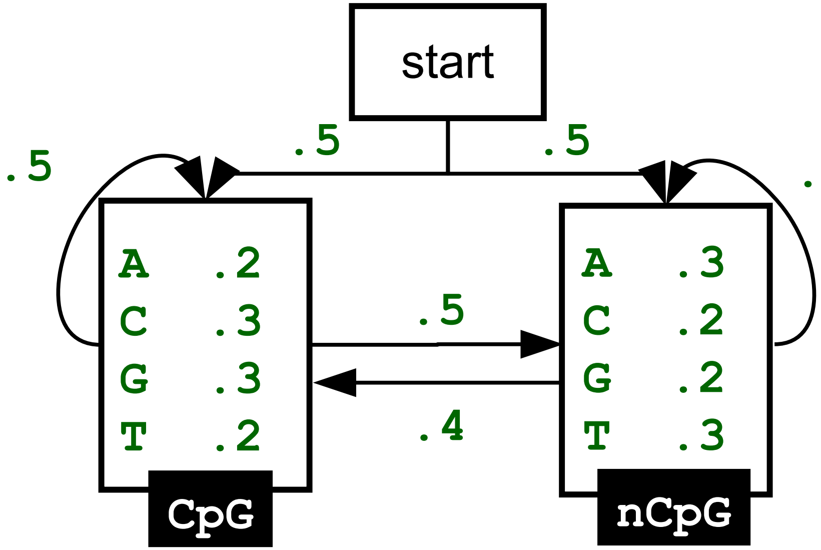
$$\boldsymbol{s}^* = \underset{s_0 \ldots s_N \in S^N}{\arg\max}\, p\left(x_0 \ldots x_N; s_0 \ldots s_N\right)$$



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									
nCpG	0									

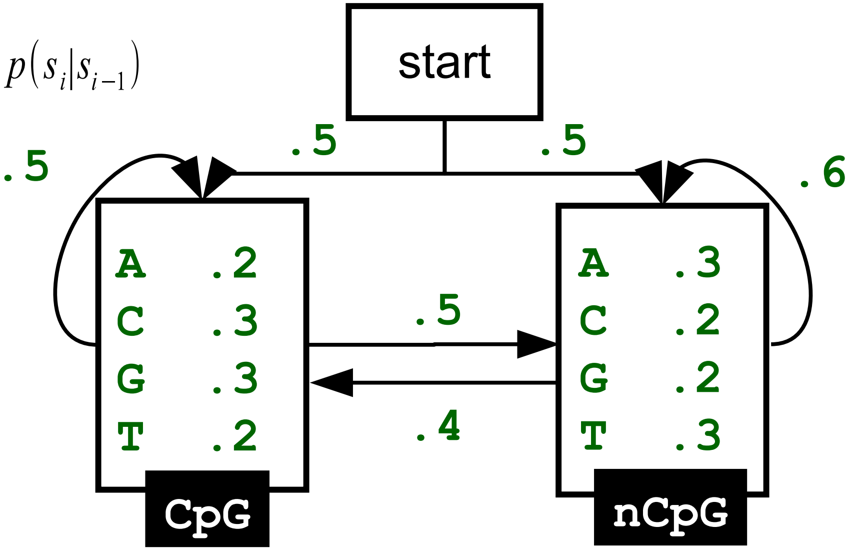
$$\max p(\epsilon|s_0) = 1 \text{ if } s_0\text{=START else: }\sim 0$$



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								

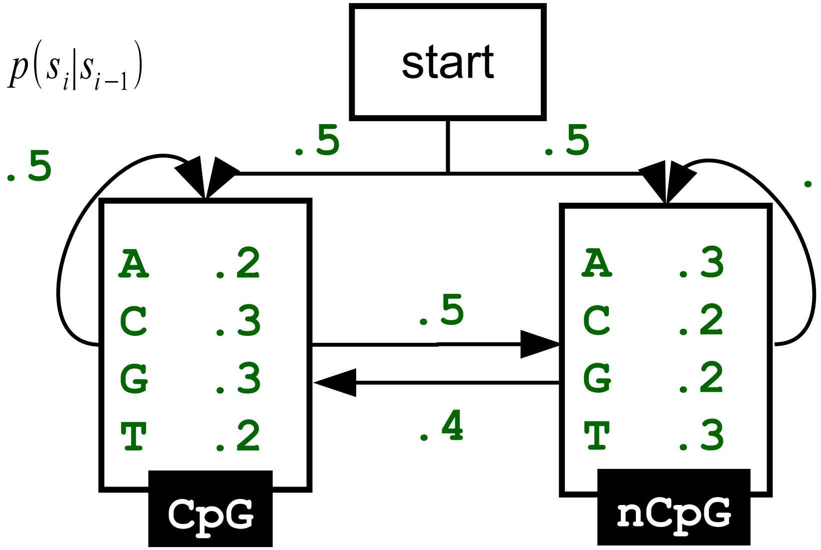
$$\max_{s_i \in S} p\left(x_0 \ldots x_i | s_i\right) = \max_{s_i \in S} p\left(x_i | s_i\right) \max_{s_{i-1} \in S} p\left(x_0 \ldots x_{i-1} | s_{i-1}\right) p\left(s_i | s_{i-1}\right)$$



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	.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							

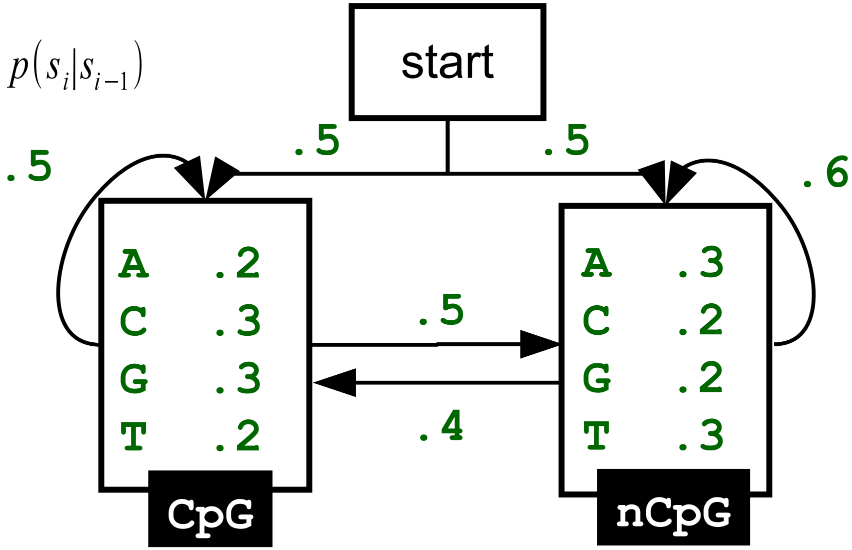
$$\max_{s_i \in S} p\left(x_0 \ldots x_i \middle| s_i\right) = \max_{s_i \in S} p\left(x_i \middle| s_i\right) \max_{s_{i-1} \in S} p\left(x_0 \ldots x_{i-1} \middle| s_{i-1}\right) p\left(s_i \middle| s_{i-1}\right)$$



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	.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				

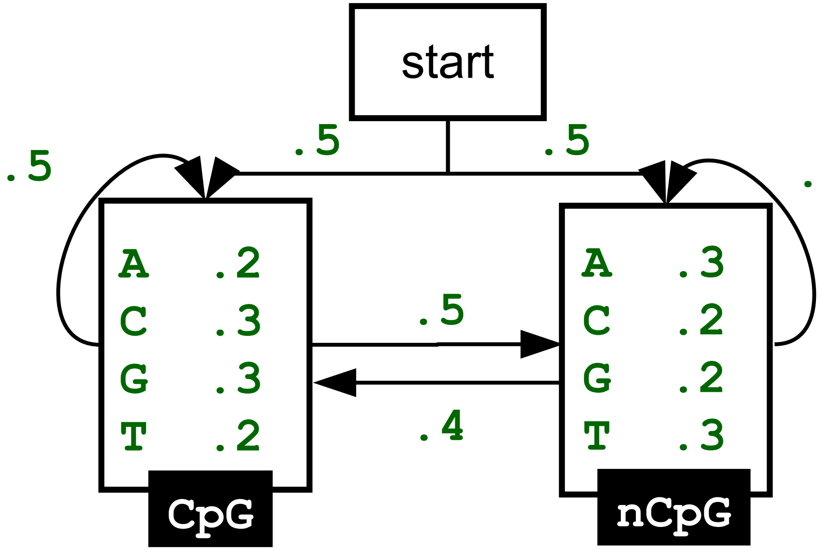
$$\max_{s_i\in S} p\left(x_0\ldots x_i|s_i\right) = \max_{s_i\in S} p\left(x_i|s_i\right) \max_{s_{i-1}\in S} p\left(x_0\ldots x_{i-1}|s_{i-1}\right) p\left(s_i|s_{i-1}\right)$$



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							

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



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