Bioinformatika
Hidden Markov Models

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

Motivation

- What is a gene?

Motivation

- What is a gene?
- Can you formalize a gene as a regular expression?

Motivation

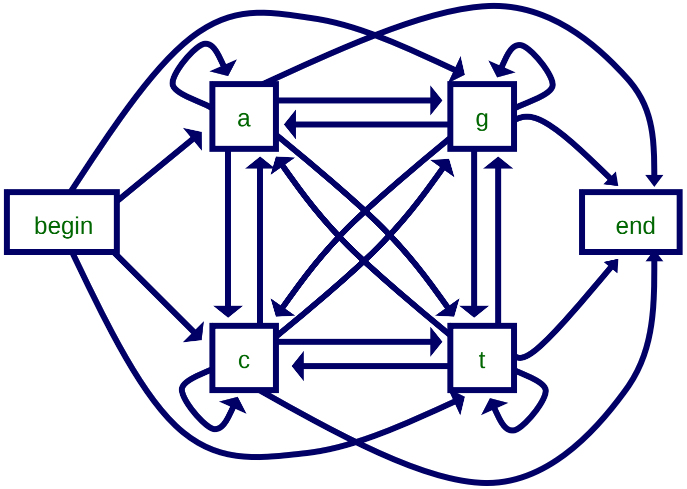
- What is a gene?
- Can you formalize a gene as a regular expression?
- What is HMM?

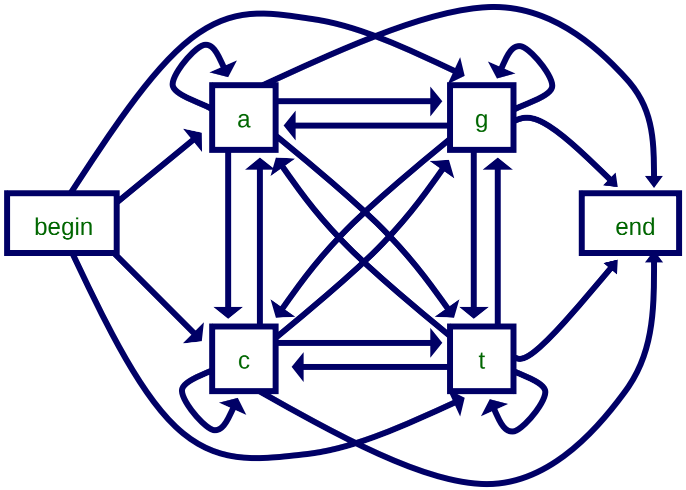
Motivation

- What is a gene?
- Can you formalize a gene as a regular expression?
- What is HMM?
- What are the general tasks with HMMs?

Motivation

1. Train two MMs: one to represent represent CpG regions, the other to the background (nCpG)



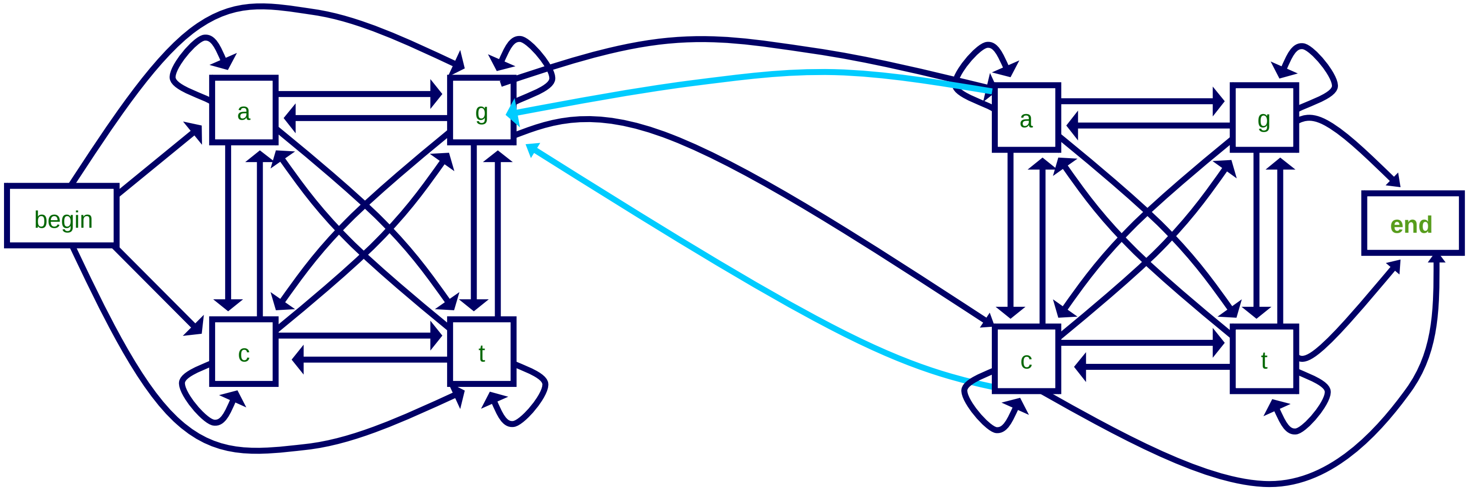


➔ Given a new sequence, use two models to *classify* the sequence (CpG or nCpG).

➔ Given a new sequence, find the CpG islands within (**?!?**)

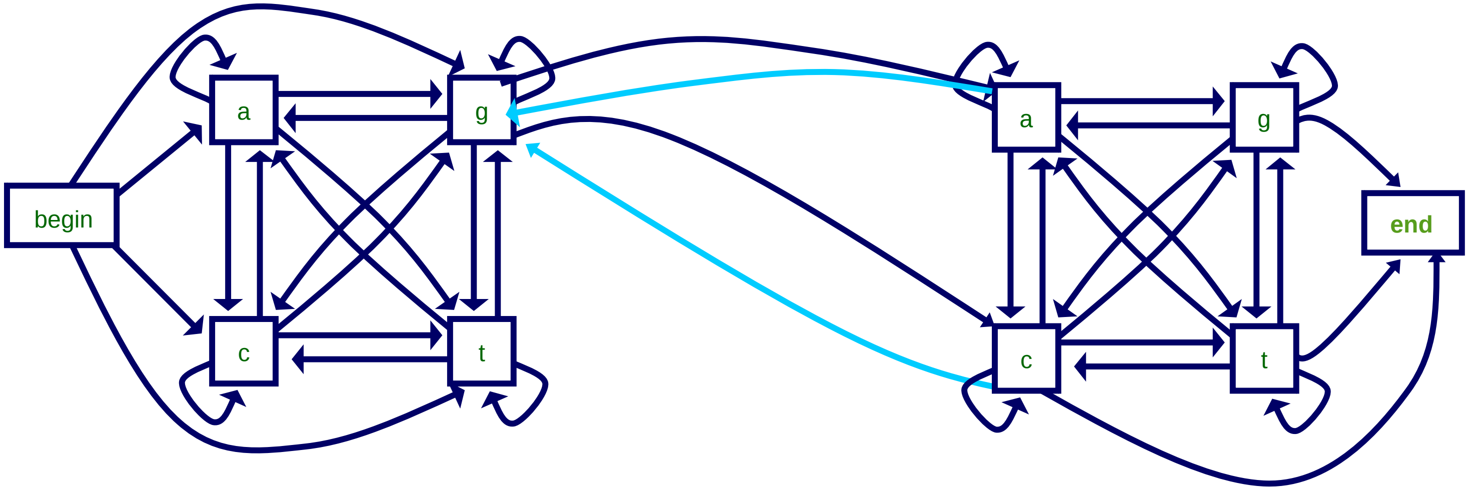
Motivation

1. Train two MMs: one to represent represent CpG regions, the other to the background (nCpG)



Motivation

1. Train two MMs: one to represent represent CpG regions, the other to the background (nCpG)



2. Join the 2 models into one HMM:

   - {a, c, t, g} → {a_{CpG}, a_{nCpG}, c_{CpG}, c_{nCpG}, t_{CpG}, t_{nCpG}, g_{CpG}, g_{nCpG}}

3. Segment a sequence as a maximum likely walk through the state space.

Hidden Markov Model

$M = (A, S, P_t, P_e)$

- $A = \{a, c, t, g\}$
- $S = \{s_1, \ldots, s_K\}$
- $P_t : S \times S \rightarrow [0, 1]$
- $P_e : S \times A \rightarrow [0, 1]$

$P(x_1, \ldots, x_L; s_1, \ldots, s_L) =$

$= P(s_1) \cdot P(x_1|s_1) \cdot P(x_2|s_2) \cdot P(s_2|s_1) \cdot$

$\cdot \ldots \cdot P(x_L|s_L) \cdot P(s_L|s_{L-1})$

with $x_i \in A$, $s_i \in S$

Sequence Annotation

Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- model $M = (A, S, P_t, P_e)$

Find:

- max. likely labeling $\mathbf{s} \in S^L$ → Viterbi alg.

Sequence Annotation

Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- model $M = (A, S, P_t, P_e)$

Find:

- max. likely labeling $\mathbf{s} \in S^L$ → Viterbi alg.

But how have the P_t, P_e been learnt??

Sequence Annotation

Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- model $M = (A, S, P_t, P_e)$

Find:

- max. likely labeling $\mathbf{s} \in S^L$ → Viterbi alg.

But how have the P_t, P_e been learnt??

- Supervised: $T = \{(\mathbf{x}_i,\mathbf{s}_i)\}_{i=1\ldots N}$ where $\mathbf{x}_i \in A^*$, $\mathbf{s}_i \in S^*$

Sequence Annotation

Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- model $M = (A, S, P_t, P_e)$

Find:

- max. likely labeling $\mathbf{s} \in S^L$ → Viterbi alg.

But how have the P_t, P_e been learnt??

- Supervised: $T = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1\ldots N}$ where $\mathbf{x}_i \in A^*$, $\mathbf{s}_i \in S^*$
- Unsupervised: $T = \{\mathbf{x}_i\}_{i=1\ldots N}$ where $\mathbf{x}_i \in A^*$
  – Expectation-Maximization → Baum-Welsh alg. (**later**)

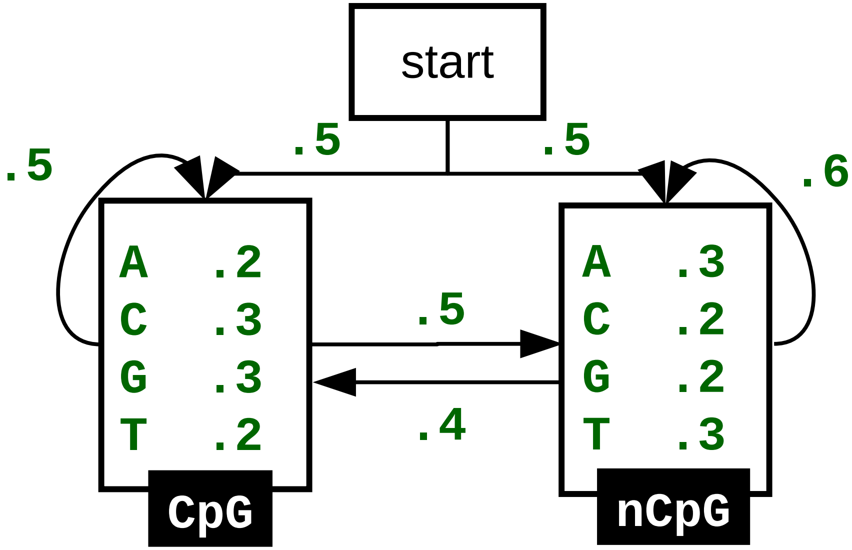
Viterbi algorithm

Ex: Naive model of CpG detection

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

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

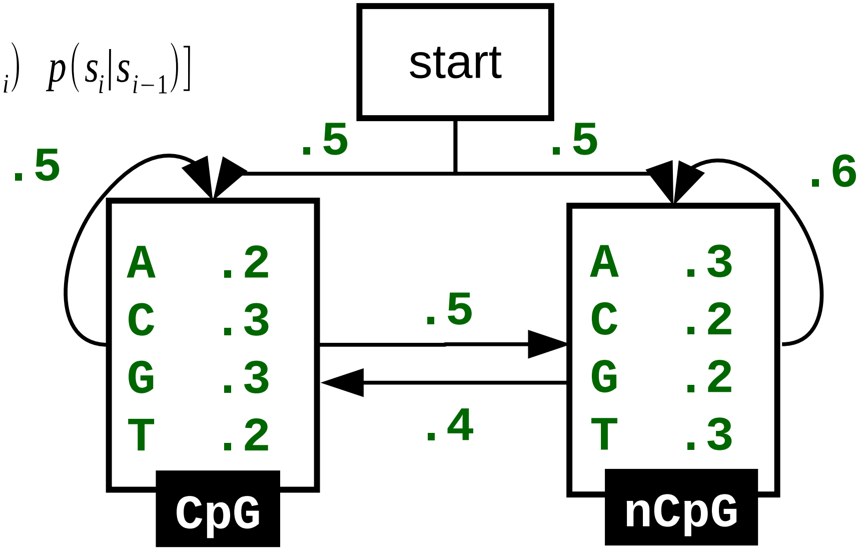
$$p\left(s_0\right) = 1$$



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									

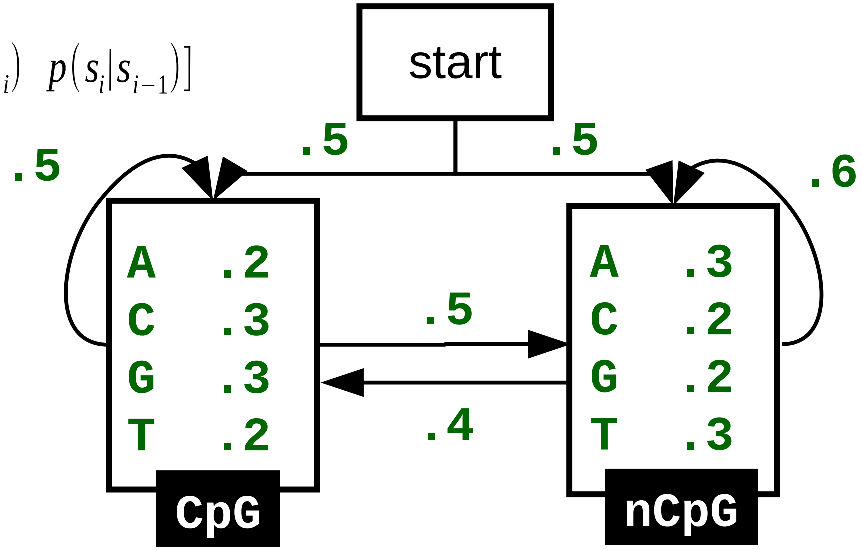
$$\max_{s_i \in S} p(x_0 \ldots x_i | s_i) = \max_{s_{i-1} \in S} \left[p(x_0 \ldots x_{i-1} | s_{i-1}) \ \max_{s_i \in S} p(x_i | s_i) \ \ p(s_i | s_{i-1}) \right]$$



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								

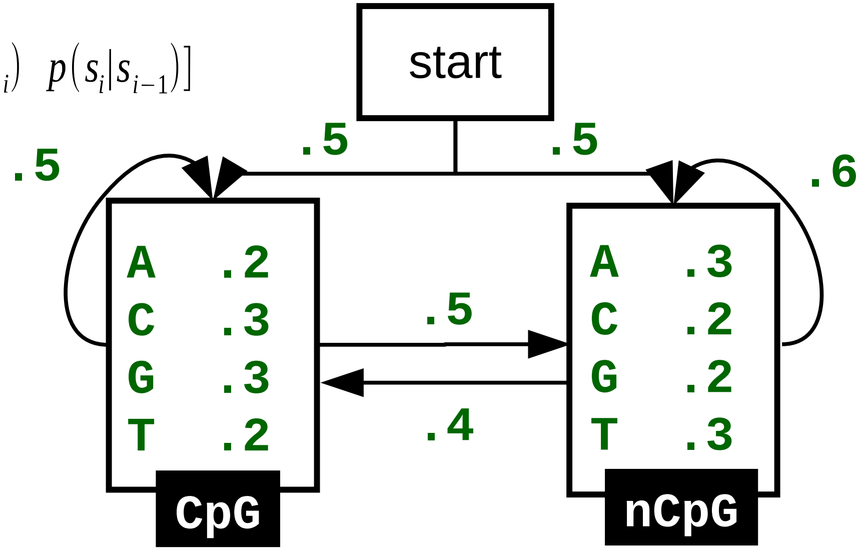
$$\max_{s_i \in S} p\left(x_0 \ldots x_i | s_i\right) = \max_{s_{i-1} \in S}\left[p\left(x_0 \ldots x_{i-1} | s_{i-1}\right) \ \max_{s_i \in S} p\left(x_i | s_i\right) \ \ p\left(s_i | s_{i-1}\right)\right]$$



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	**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 xx .6 .15	**0** x .3 x .5 .1 x .3 x .5 .15 x .3 x .6 .027							

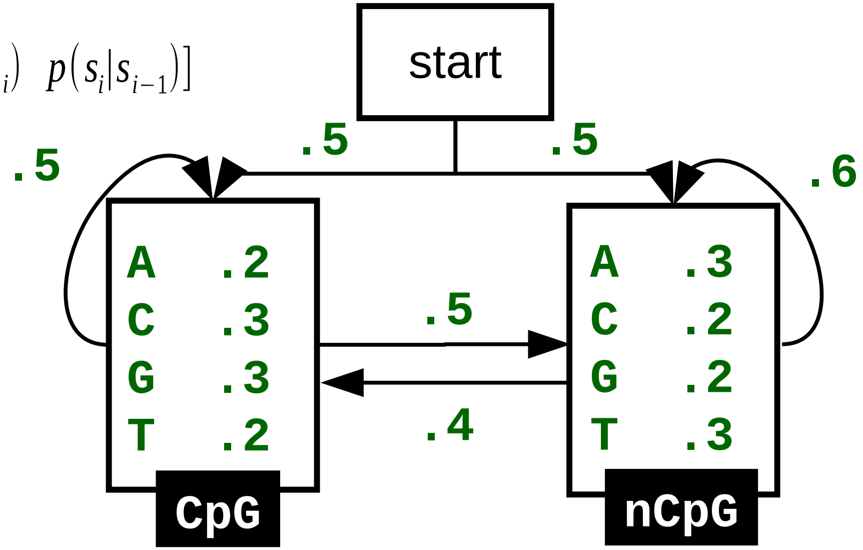
$$\max_{s_i\in S} p(x_0\ldots x_i|s_i) = \max_{s_{i-1}\in S}\left[p(x_0\ldots x_{i-1}|s_{i-1}) \max_{s_i\in S} p(x_i|s_i) \quad p(s_i|s_{i-1})\right]$$



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

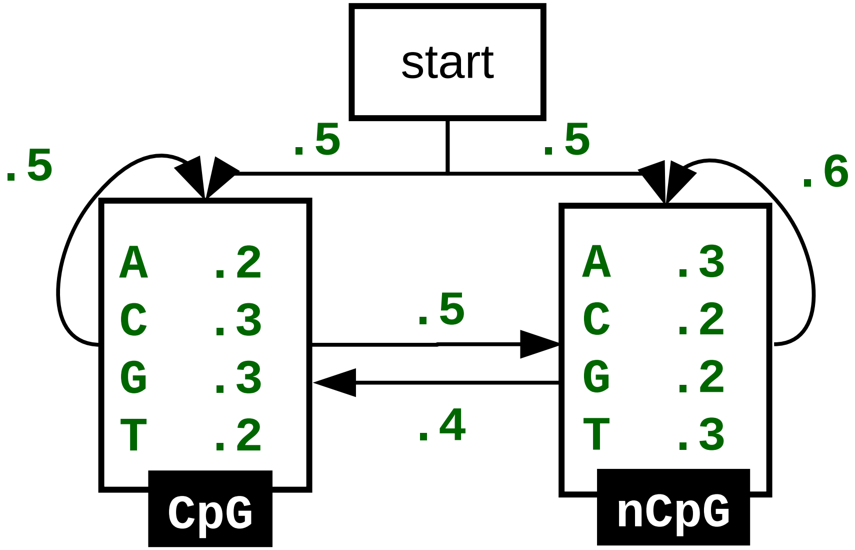
$$\max_{s_i\in S} p(x_0\ldots x_i|s_i) \;=\; \max_{s_{i-1}\in S}\left[p(x_0\ldots x_{i-1}|s_{i-1})\;\; \max_{s_i\in S} p(x_i|s_i)\;\; p(s_i|s_{i-1})\right]$$



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							

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

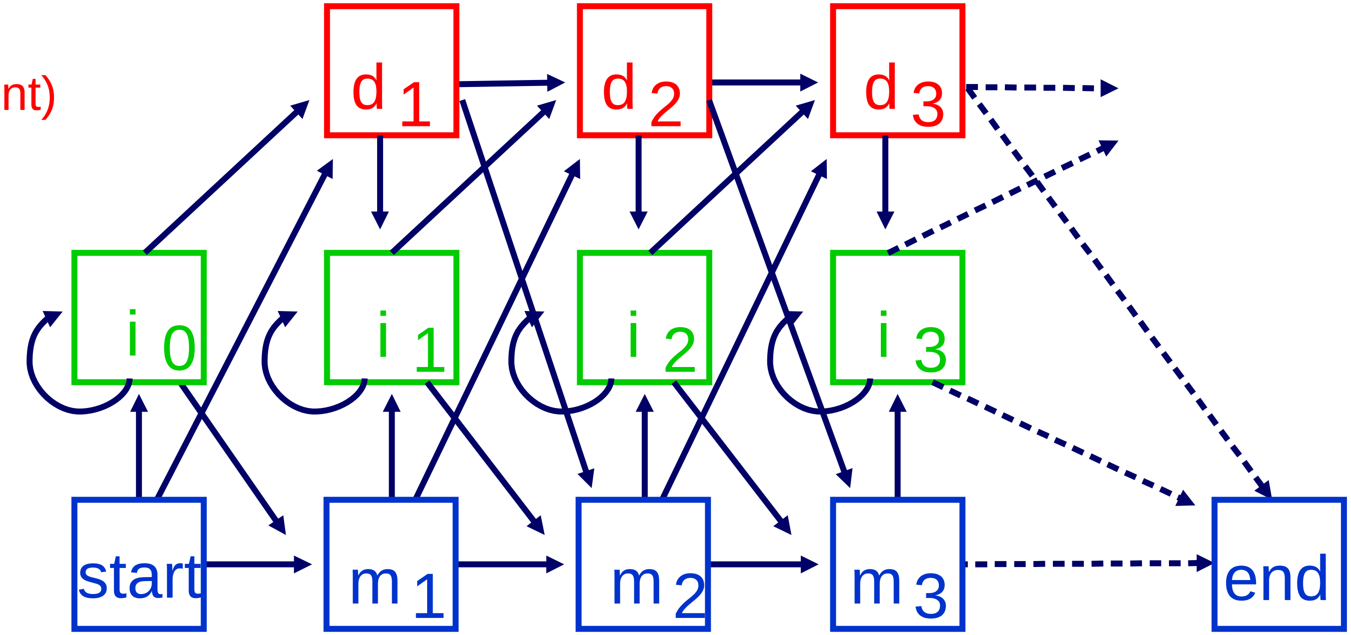


Assignment – Gene Finding

- http://www.biostat.wisc.edu/~craven/776/hw3.html
- **You can use an existing implementation of Viterbi alg.**
- 15 pt.

Profile HMM

```
ATTGCC- A TT--
ATGGCC- A TT--
ATC-CA- A TTTT
ATCTTC- - TT--
ATTGCCG A TT--
```



Delete states (silent)

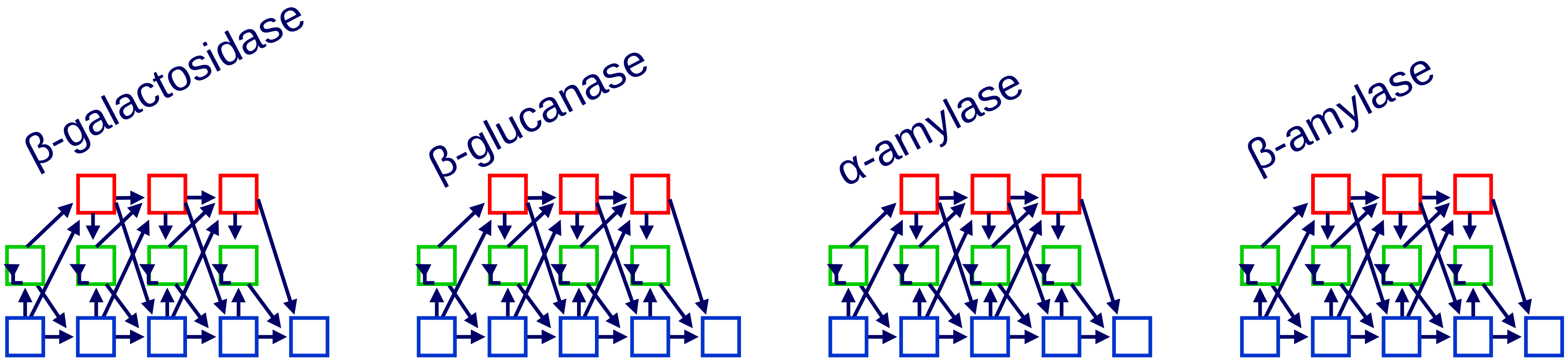
Insert states

Match states

Profile HMM – Excercise

```
AG---C
A-AG-C
AG-AA-
--AAAC
AG---C
12-3-4
```

Sequence Categorization



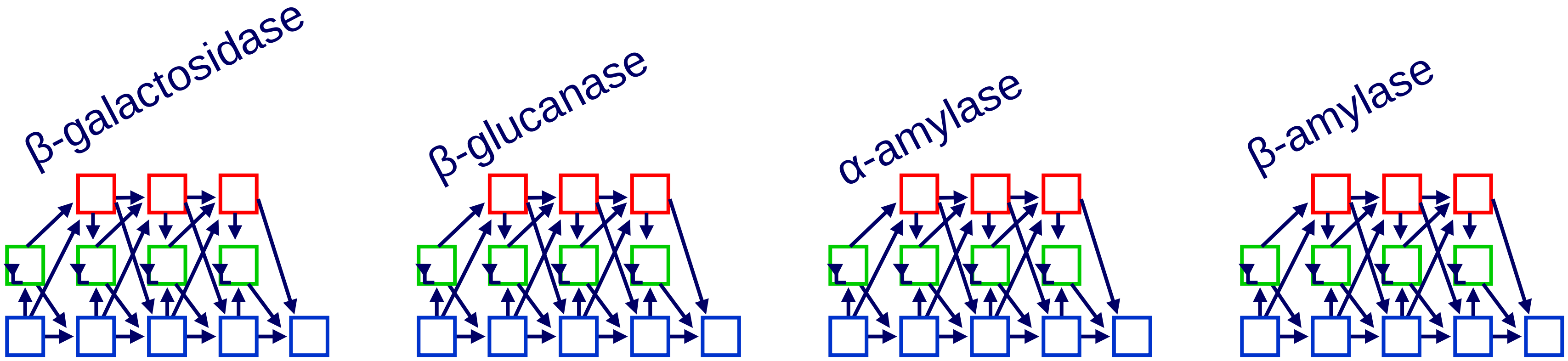
Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- Set of K models $\{M_k = (A, S, P_t, P_e)\}$ of K families

Do:

- categorize **x** into one of the families

Sequence Categorization



Given:

- observed sequence $\mathbf{x} \in \{a, c, t, g\}^L$
- Set of *K* models $\{M_k = (A, S, P_t, P_e)\}$ of *K* families

Do:

- categorize **x** into one of the families

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

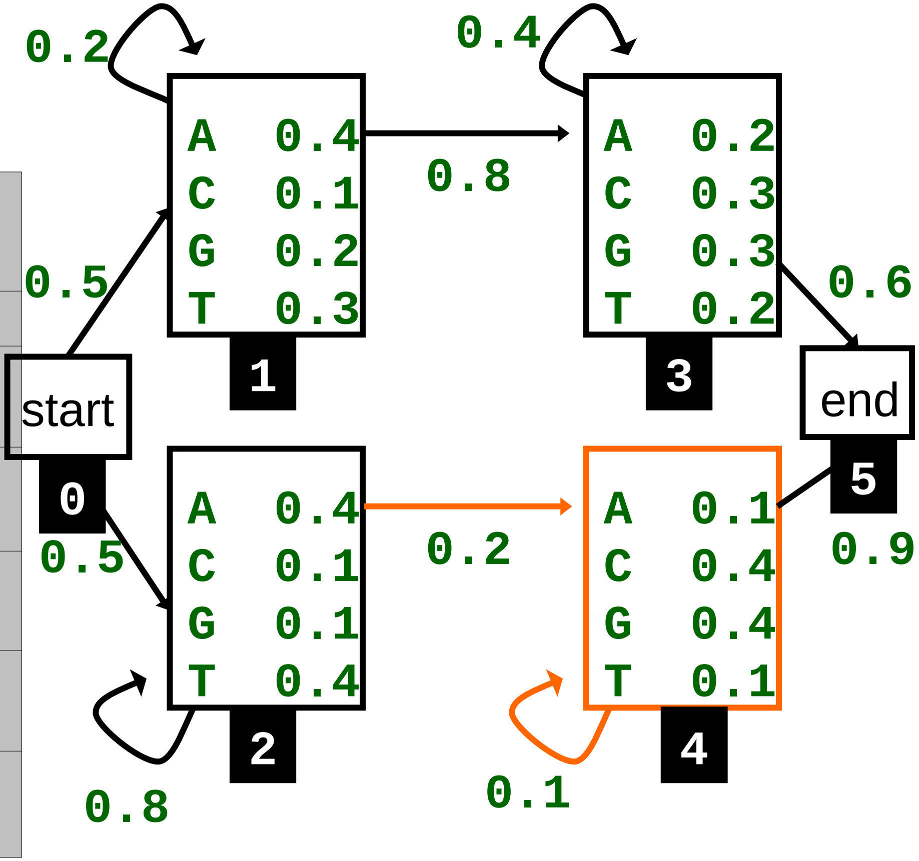
$$p(\alpha-amyl.)\,p(x_0 \ldots x_N|\alpha-amyl.) < p(\beta-gluc.)\,p(x_0 \ldots x_N|\beta-gluc.)$$

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

Forward algorithm (ex.)

$$\sum_{s_1\ldots s_i} p\left(x_0\ldots x_i, s_1\ldots s_i\right) = \sum_{s_i\in S}\sum_{s_1\ldots s_{i-1}} p\left(x_1\ldots x_{i-1}, s_1\ldots s_{i-1}\right) p\left(x_i|s_i\right) p\left(s_i|s_{i-1}\right)$$

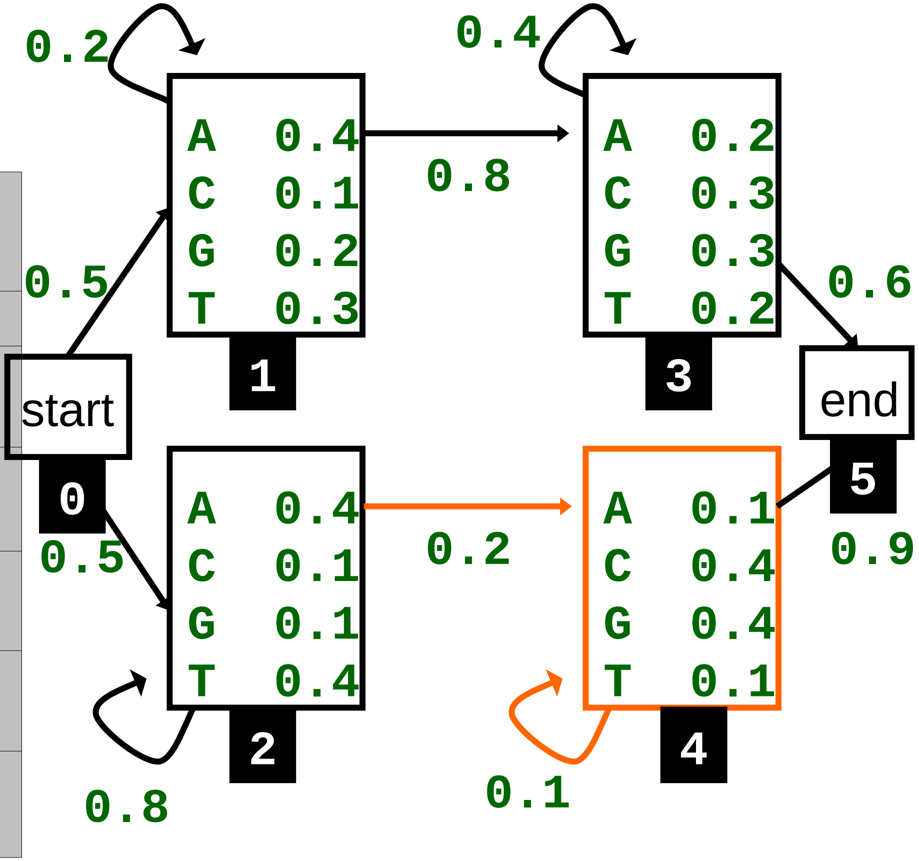
	ε	T	A	G	A	ε
0	**1**	0	0	0	0	
1	0					
2	0					
3						
4						
5		0	0	0	0	



Forward algorithm (ex.)

$$\sum_{s_1\dots s_i} p\left(x_0\dots x_i, s_1\dots s_i\right) = \sum_{s_i\in S}\sum_{s_1\dots s_{i-1}} p\left(x_1\dots x_{i-1}, s_1\dots s_{i-1}\right) p\left(x_i|s_i\right) p\left(s_i|s_{i-1}\right)$$

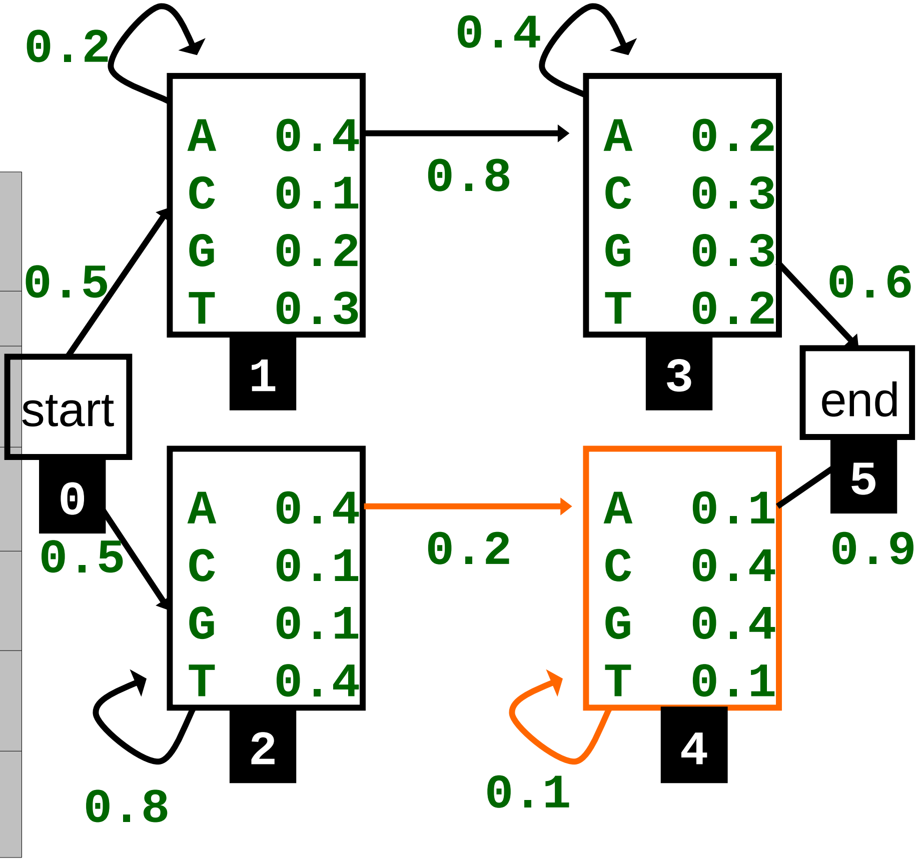
	ε	T	A	G	A	ε
0	**1**	0	0	0	0	
1	0	**1** x .3 x .5 0 x .3 x .2 .15				
2	0	**1** x .4 x .5 0 x .4 x .8 .2				
3		0				
4		0				
5		0	0	0	0	



Forward algorithm (ex.)

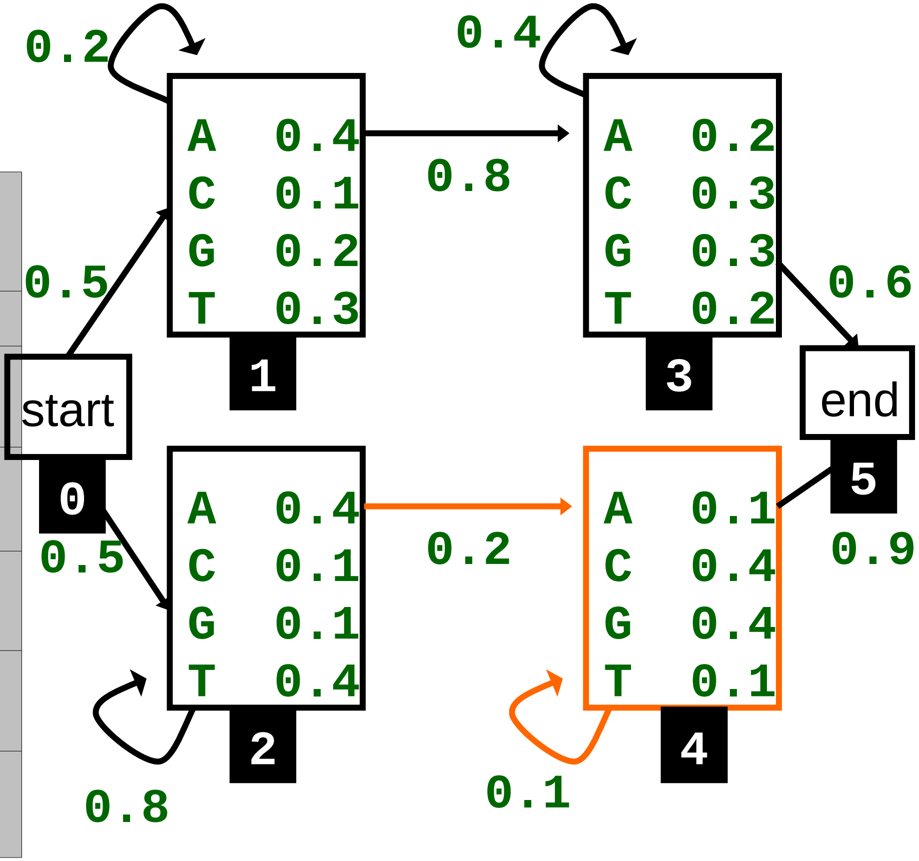
$$\sum_{s_1\ldots s_i} p(x_0\ldots x_i, s_1\ldots s_i) = \sum_{s_i\in S}\sum_{s_1\ldots s_{i-1}} p(x_1\ldots x_{i-1}, s_1\ldots s_{i-1})\,p(x_i|s_i)\,p(s_i|s_{i-1})$$

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



Forward algorithm (ex.)

$$\sum_{s_1\ldots s_i} p\left(x_0\ldots x_i, s_1\ldots s_i\right) = \sum_{s_i\in S}\sum_{s_1\ldots s_{i-1}} p\left(x_1\ldots x_{i-1}, s_1\ldots s_{i-1}\right) p\left(x_i|s_i\right) p\left(s_i|s_{i-1}\right)$$

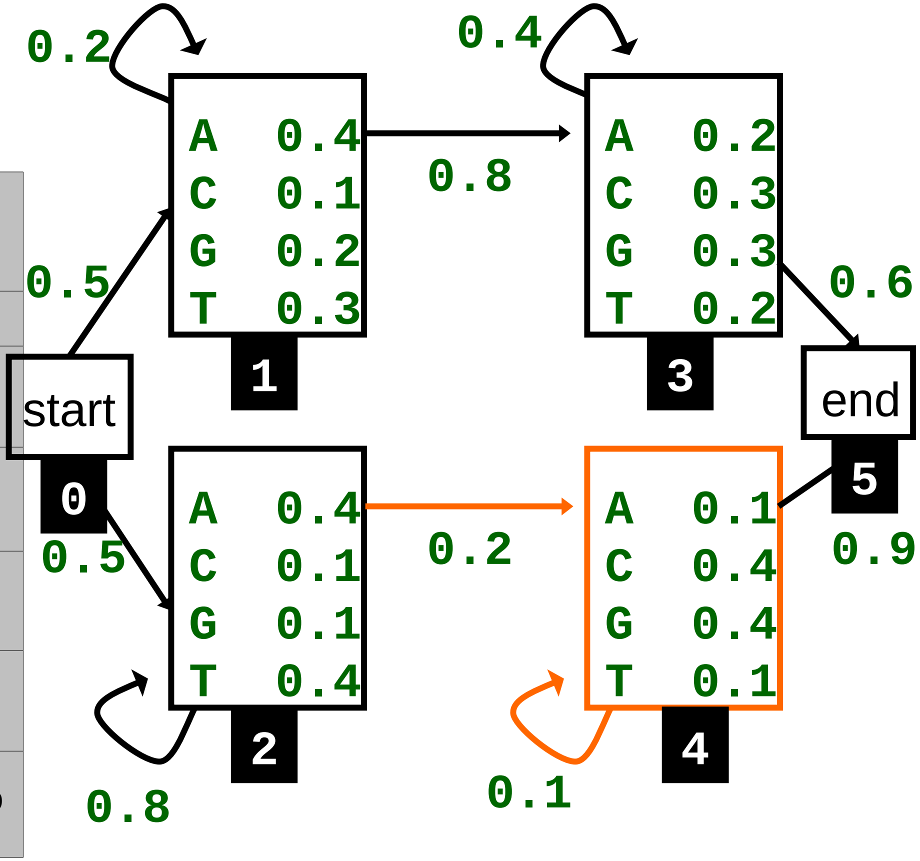


	ε	T	A	G	A	ε
0	**1**	0	0	0	0	
1	0	**1** x .3 x .5 0 x .3 x .2 .15	**0** x .4 x .5 .15 x .4 x .2 .012	**0** x .2 x .5 .012 x .2 x .2 5e-4		
2	0	**1** x .4 x .5 0 x .4 x .8 .2	**0** x .4 x .5 .2 x .4 x .8 .064	**0** x .1 x .5 .064 x .1 x .8 .00512		
3		0	.15 x .2 x .8 0 x .2 x .4 .024	.012 x .3 x .8 .024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

Forward algorithm (ex.)

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

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



Sum-up

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