Hidden Markov models in gene finding

Broňa Brejová

Bioinformatics research group

David R. Cheriton School of Computer Science

University of Waterloo

Topics for today

- What is gene finding (biological background)
- HMMs in gene finding – tools of the trade:
  - Viterbi algorithm
  - Higher order states
  - State transition diagram

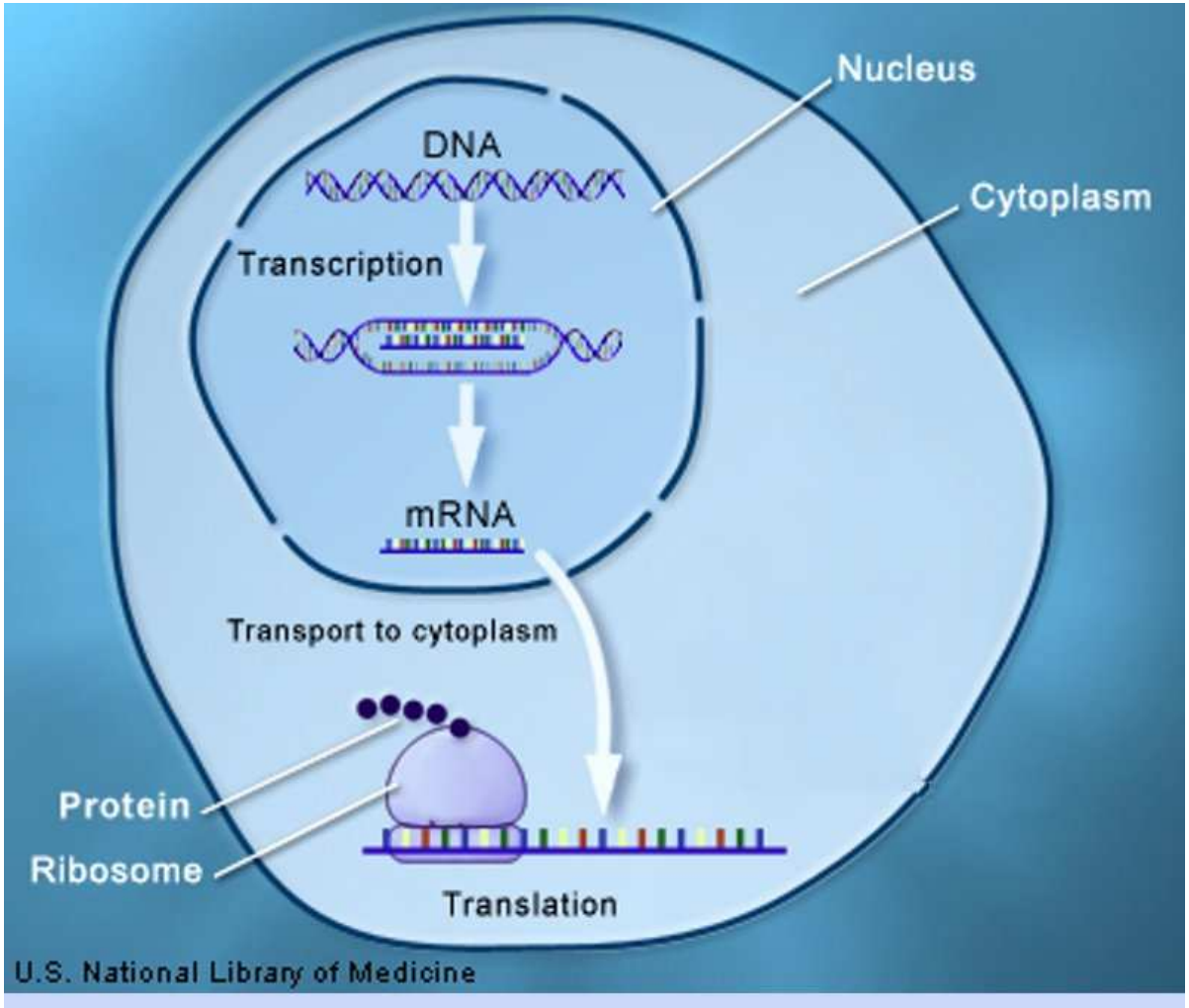
What is a genome?

```
cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgtttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
ggtccaagcaaaggctcttttcagagccaccacctttcaagtaaagtagctgtaagaaa
ccaatttaagacaaaagggaatgcattgggagcactttcgttttaatgctactgaaggc
```

Human genome: 3 billion letters (sequenced in 2001)

What do they mean?

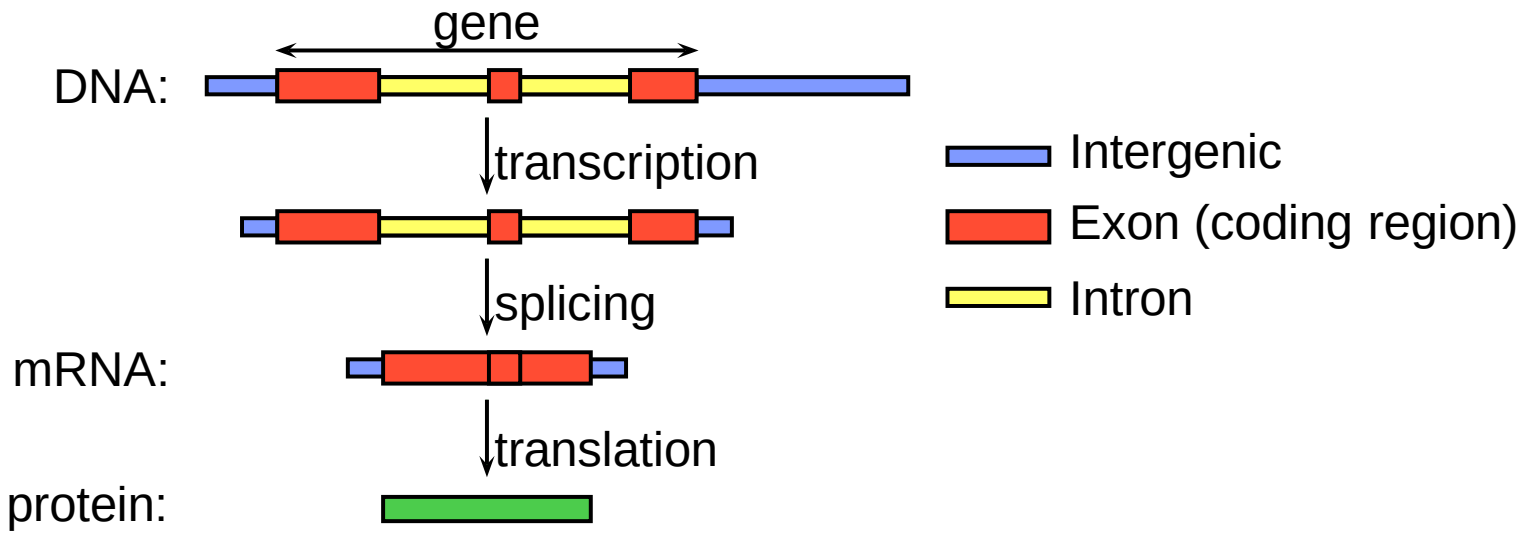
What are genes?



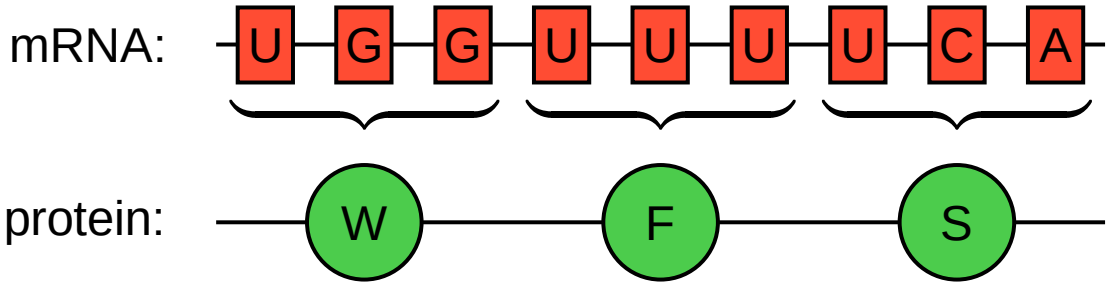
Genes are portions of genome encoding proteins

Closer look on genes

Process of protein production:



Translation: triple of letters in mRNA $\longrightarrow$ one amino acid in protein

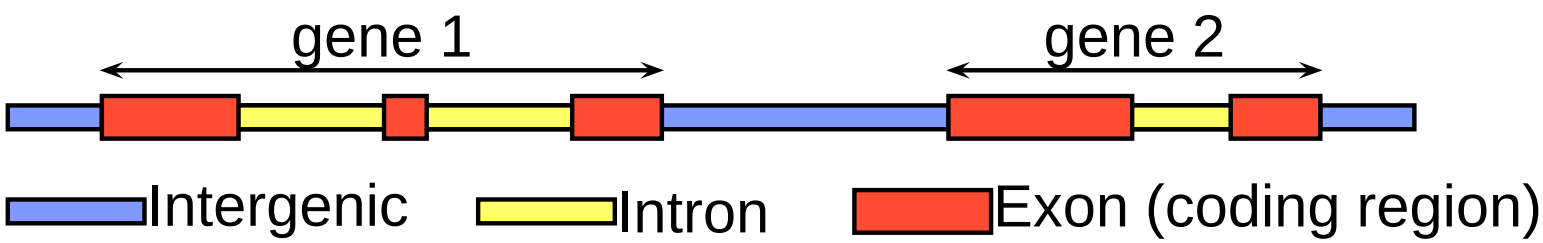


Task of gene finding

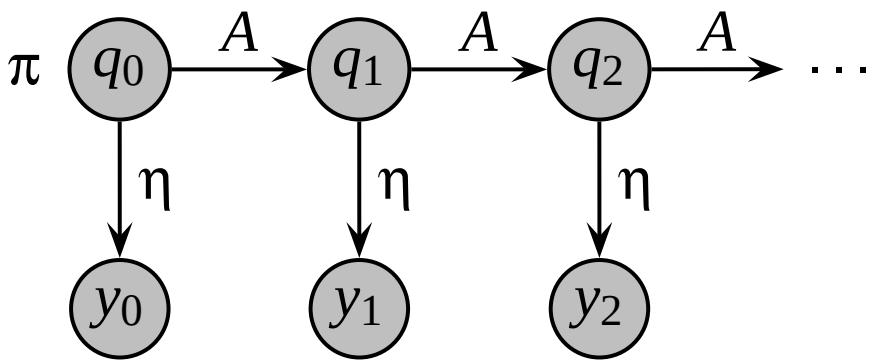
```
cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgtttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
```

Task of gene finding

```
cggtgaaactgcacgattgttgctggcttaaagatagaccaatcagagtgtgtaacgtca
tatttagcgtcttctatcatccaatcactgcactttacacactataaatagagcagctca
tgggcgtatttgcgctagtgttgggtgttccgctgtgctgtttttccgtcatggctcgca
ctaagcaaactgctcggaagtctactggtggcaaggcgccacgcaaacagttggccacta
aggcagcccgcaaaagcgctccggccaccggcggcgtgaaaaagccccaccgctaccggc
cgggcaccgtggctctgcgcgagatccgccgttatcagaagtccactgaactgcttattc
gtaaactacctttccagcgcctggtgcgcgagattgcgcaggactttaaaacagacctgc
gtttccagagctccgctgtgatggctctgcaggaggcgtgcgaggcctacttggtagggc
tatttgaggacactaacctgtgcgccatccacgccaagcgcgtcactatcatgcccaagg
acatccagctcgcccgccgcatccgcggagagagggcgtgattactgtggtctctctgac
```



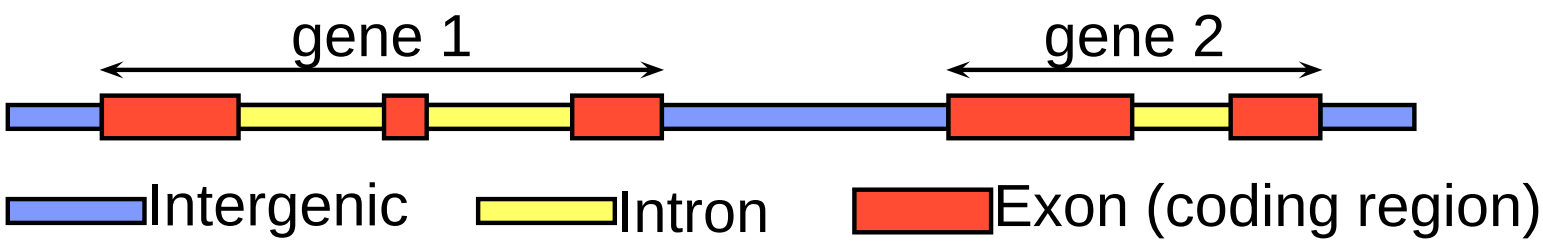
Toy Hidden Markov model for gene finding



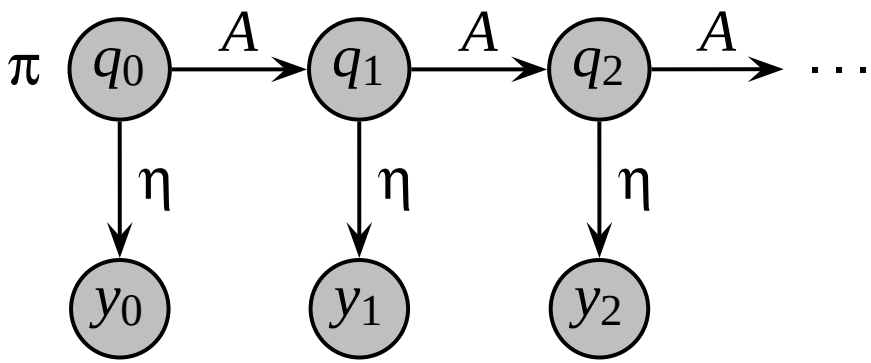
$q_t \in \{$ ■, ■, ■ $\}$

$y_t \in \{a, c, g, t\}$

$p(q, y)$: joint prob. over genomic sequences y and gene structures q



Toy Hidden Markov model for gene finding



$q_t \in \{$ ■, ■, ■ $\}$

$y_t \in \{a, c, g, t\}$

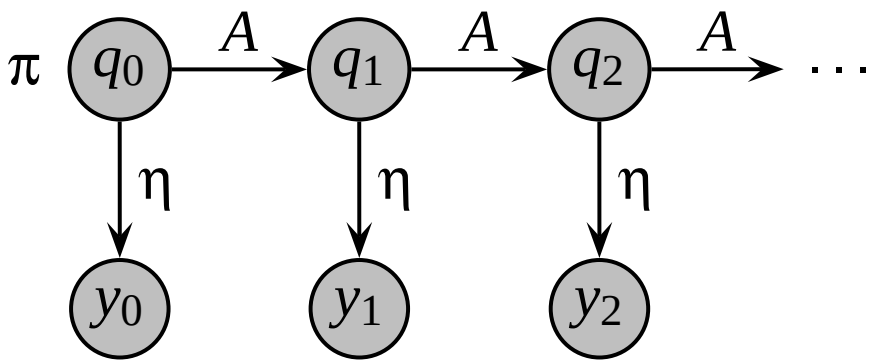
A	■	■	■
■	0.99	0.007	0.003
■	0.01	0.99	0
■	0.001	0	0.999

$p(q_t|q_{t-1})$

η	a	c	g	t
■	0.24	0.27	0.28	0.21
■	0.26	0.22	0.22	0.30
■	0.27	0.23	0.23	0.27

$p(y_t|q_t)$

HMMs for gene finding: training



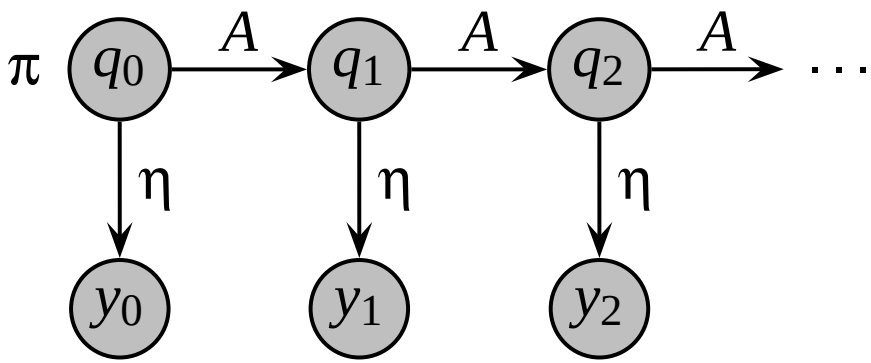
$\mathtt{q}_t \in \{$ ■, □, ■ $\}$

$\mathtt{y}_t \in \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$

Training:

- Training set of sequences with known gene structures
- Fully observed model (both $\mathtt{q}$ and $\mathtt{y}$ known)
- Maximum likelihood estimation by relative frequencies

HMMs for gene finding: inference



$q_t \in \{$ ■, ■, ■ $\}$

$y_t \in \{a, c, g, t\}$

Inference: Viterbi algorithm

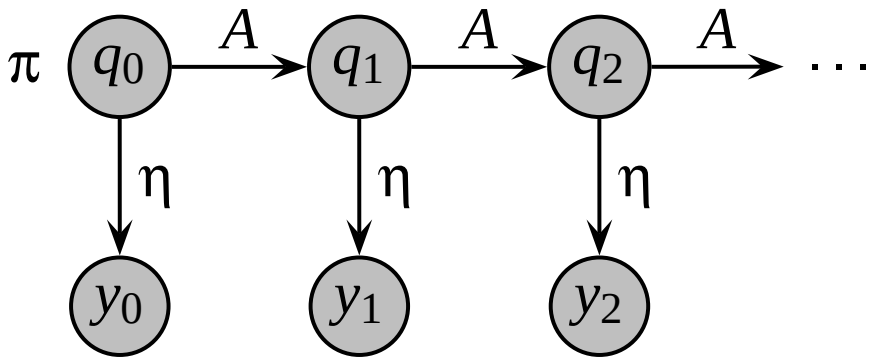
- Genomic sequence y observed, state sequence q unknown
- Viterbi algorithm computes state sequence q^*

  with highest probability given y:

$$q^* = \arg\max_q p(q|y) = \arg\max_q p(q, y)$$

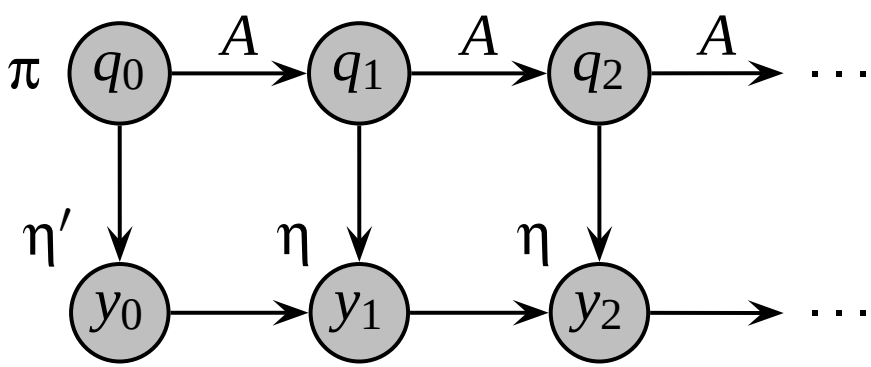
HMMs for gene finding: higher order states

Order 0:



η specifies $p(y_t|q_t)$

Order 1:



η specifies $p(y_t|q_t, y_{t-1})$

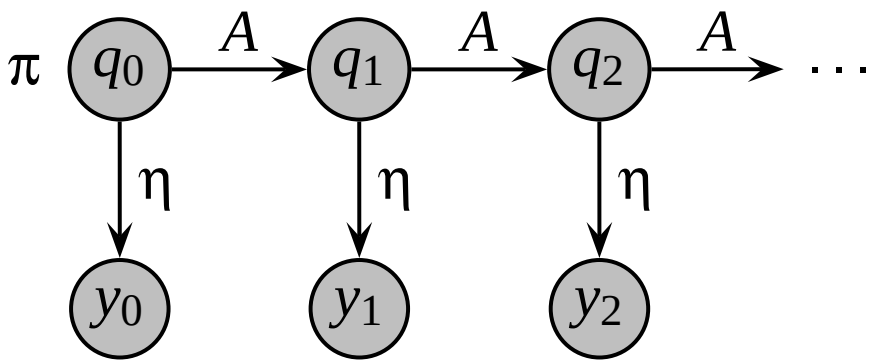
HMMs for gene finding: higher order states

η specifies $p(y_t|q_t, y_{t-1})$:

q_t	y_{t-1}	a	c	g	t
	a	0.24	0.23	0.34	0.19
	c	0.30	0.31	0.13	0.26
	g	0.27	0.28	0.28	0.17
	t	0.13	0.28	0.38	0.21
	a	0.30	0.18	0.27	0.25
	c	0.32	0.28	0.06	0.35
	g	0.27	0.22	0.27	0.24
	t	0.20	0.21	0.26	0.33

...

HMMs for gene finding: state transition diagrams



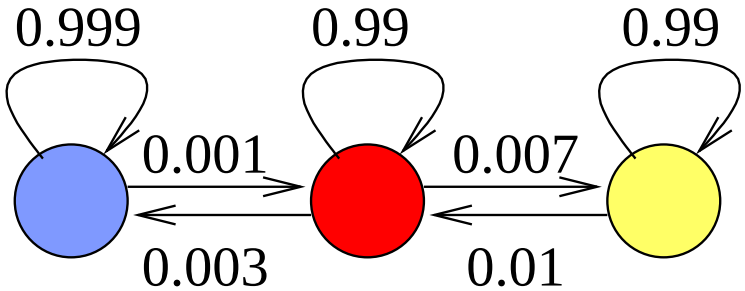
$q_t \in \{$ ■, ■, ■ $\}$

$y_t \in \{a, c, g, t\}$

A	■	■	■
■	0.99	0.007	0.003
■	0.01	0.99	0
■	0.001	0	0.999

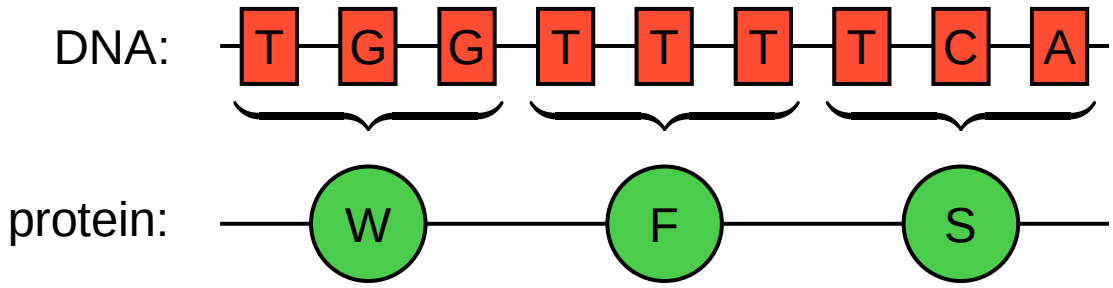
$p(q_t|q_{t-1})$

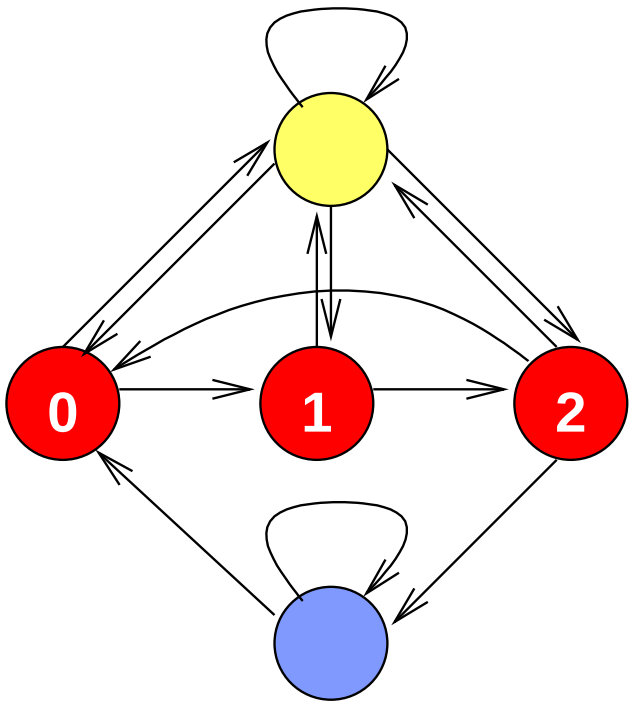
Graph representation of A:



More complex HMMs: change state transition diagrams

Recall: triple of letters in exon $\longrightarrow$ one amino acid

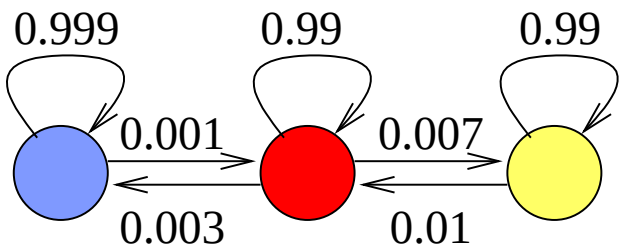


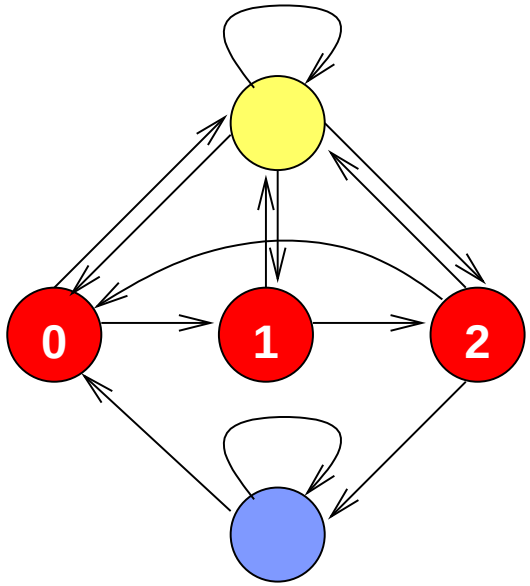


A	0	1	2		
0	0		0		0
1	0	0			0
2		0	0		
					0
		0	0		0

$p(q_t|q_{t-1})$

New states have separate emission probability



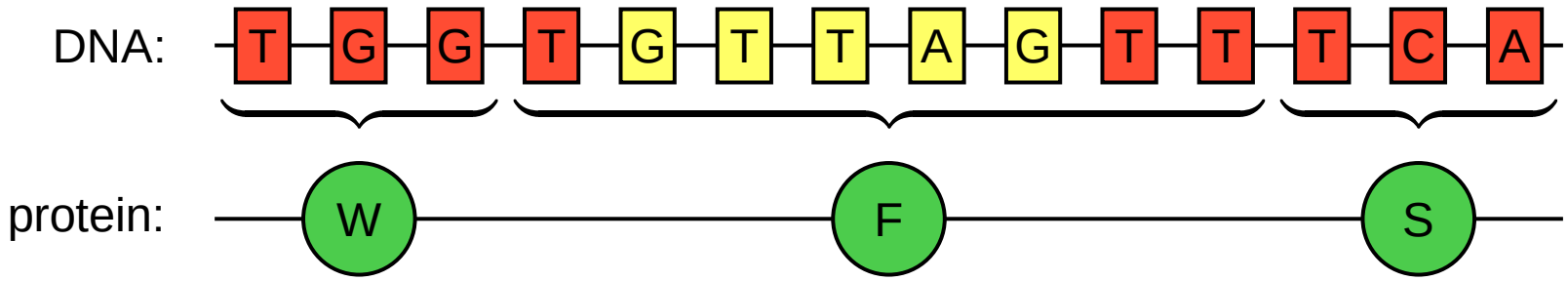


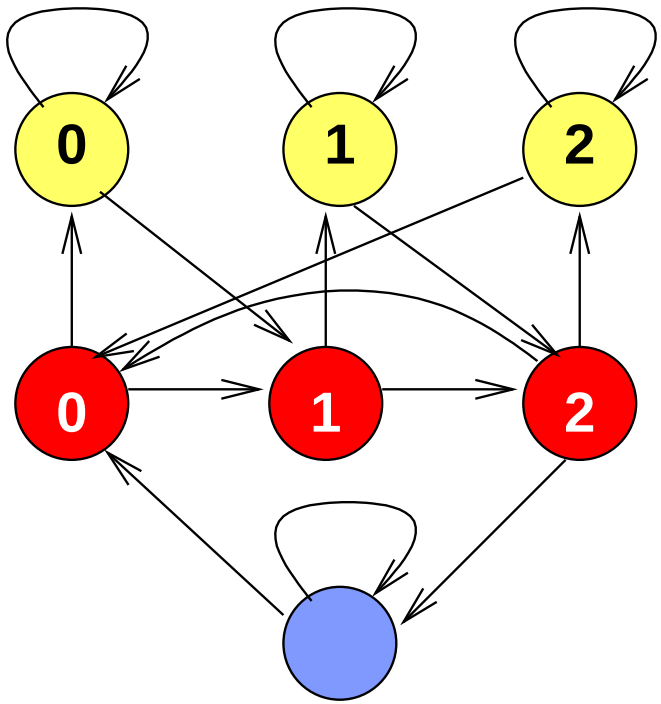
η	a	c	g	t
	0.24	0.27	0.28	0.21
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

η	a	c	g	t
0	0.26	0.26	0.32	0.16
1	0.30	0.24	0.20	0.26
2	0.17	0.32	0.31	0.20
	0.26	0.22	0.22	0.30
	0.27	0.23	0.23	0.27

More complex HMMs: change state transition diagrams

Keep consistent triples across introns:





More complex HMMs: remember those signals?

Exon | Intron

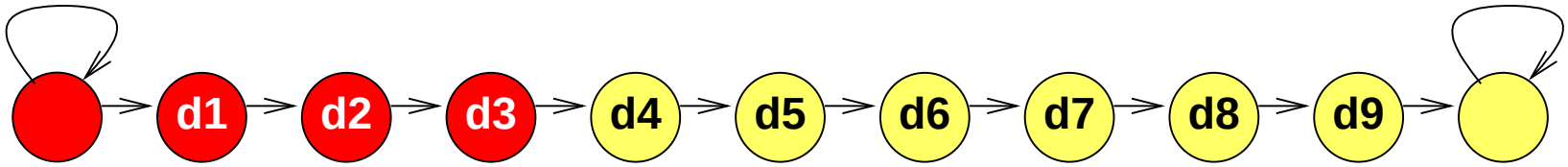
```
ccatcccctatatttatgg**caggtgagg**aaagggtgggggctgggg
attcatcatcatgggtgca**tcggtgagt**atctcccaggccccaatc
agaagatctaccccaccat**ctggtaagt**gtgtcccaccactgcccc
acagagtgagcccttcttc**aaggtgggt**ggtgtcagggcctccccc
acgagtcctgcatgagcca**gatgtaagg**cttgccgttgccctccct
tgcagaacctcatggtgct**gaggtgggg**ccaagcctgggccggggg
tcgatgaatttgggatcat**ccggtgaga**gctcttcctctctcctgg
agatgacgtccgtgatgag**aaggtagggg**ggtgcaccccagtcccca
gtggagaatgagaggtggg**atggtaggt**gatgccttcgaggcccag
tttcttgtggctattttaa**aaggtaatt**catggagaaatagaaaaa
tttcttgtggctattttaa**aaggtaatt**catggagaaatagaaaaa
tttgaagaaactccacgaa**gaggttgat**ggcagtgactttcggaaa
agtggatgcccttaaagga**accgtggag**taccaacccccctgcagt
cgacccgtgaccctcgtga**gaggtacga**agccccagcccggggctc
aaatgcagtggaagaggga**ctagtacgt**gagccatgctgggagtgt
catggcgggtgtgctgaag**aaggtgaga**cgaatggaggtcactgtt
tgtgtgcaatactcctcac**gaggtatgt**acctttgttctttcttcg
gaagctggctatggttaaa**gcggtaagt**agctaagtcagttttgtt
attagaagaggtgattctt**caggtaaag**aaaaagttgactatttag
```

More complex HMMs: remember those signals?

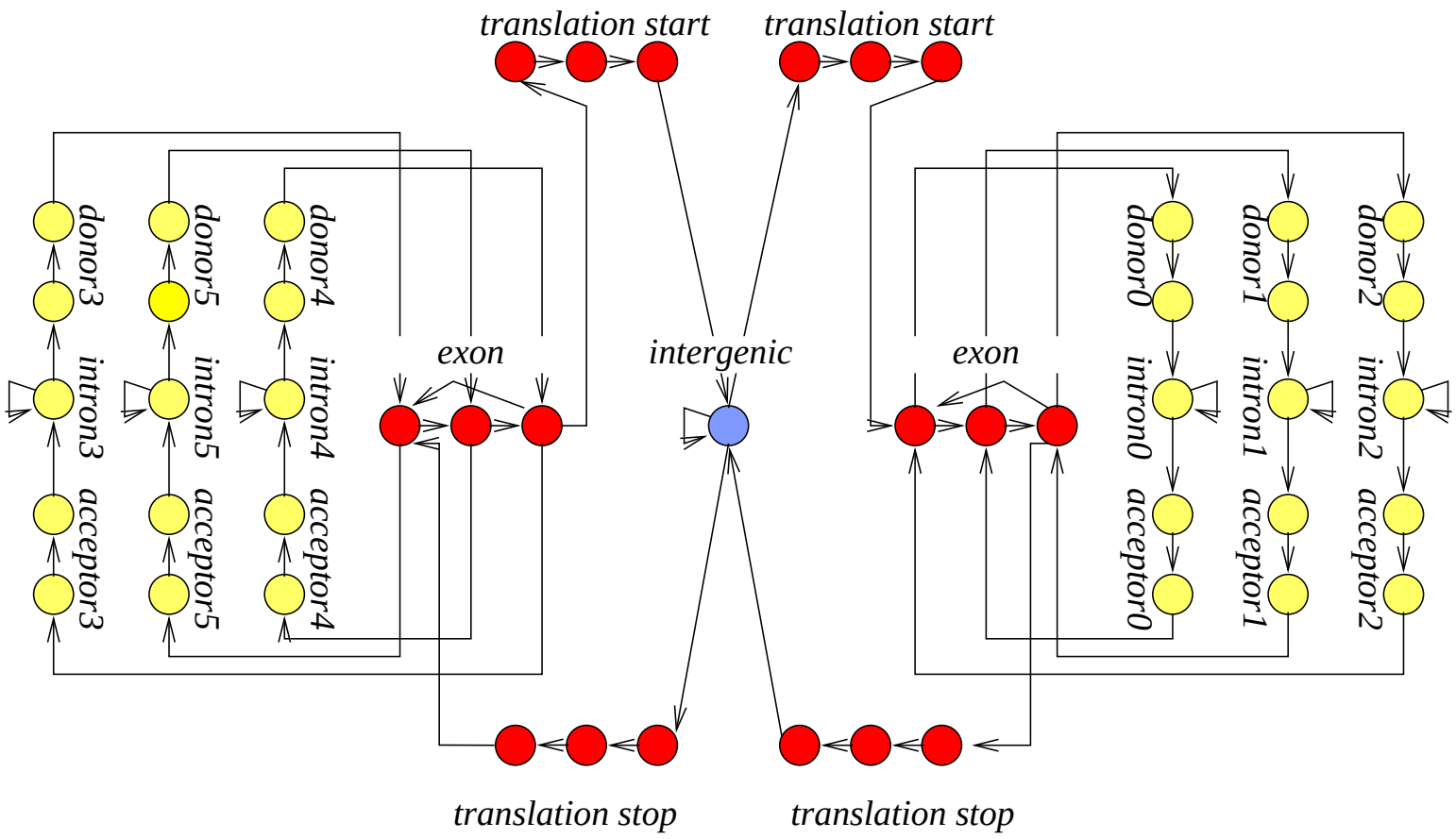
Exon | Intron

```
ccatcccctatatttatgg**caggtgagg**aaagggtgggggctgggg
attcatcatcatgggtgca**tcggtgagt**atctcccaggccccaatc
agaagatctaccccaccat**ctggtaagt**gtgtcccaccactgcccc
acagagtgagcccttcttc**aaggtgggt**ggtgtcagggcctccccc
```

Insert a series of states between exon and intron:



Example of an HMM for gene finding



Reverse strand ***Forward strand***

Summary

- HMMs are useful tool for gene finding
- Viterbi algorithm used for inference
- Higher order states model dependencies among adjacent letters
- Complex state transition diagrams incorporate biological knowledge

This is just a beginning

- More complex models used in practice: lengths of exons/introns, complex signal models, experimental information, multiple species,. . .
- HMMs for other bioinformatics tasks: protein secondary structure, protein families, segmenting genome,. . .