

Hidden Markov Models

Rune Lyngsø

2nd of November 2011

Outline

Hidden Markov Models

Viterbi Algorithm

Full Likelihood Computation

Posterior Probabilities

Applications

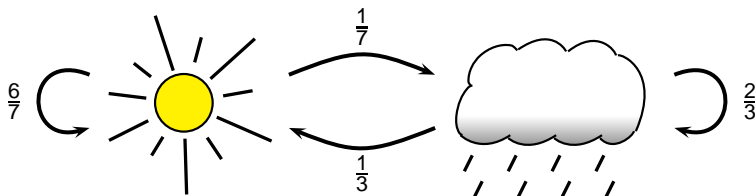
Parameter Estimation

Advanced Topics

Markov Models

Empirical Fact

Weather tomorrow is most likely to be the same as today

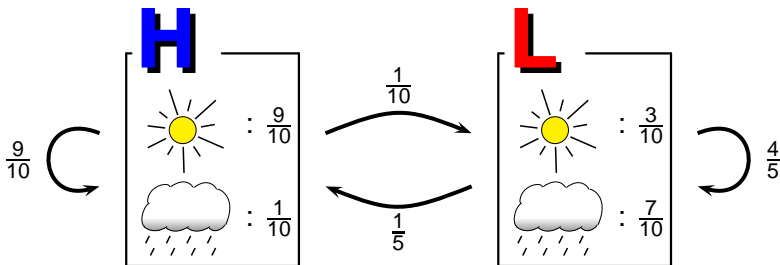


This is typical example of Markov behaviour

Hidden Markov Models

More realism

Underlying unobserved phenomena may be the conserved factor

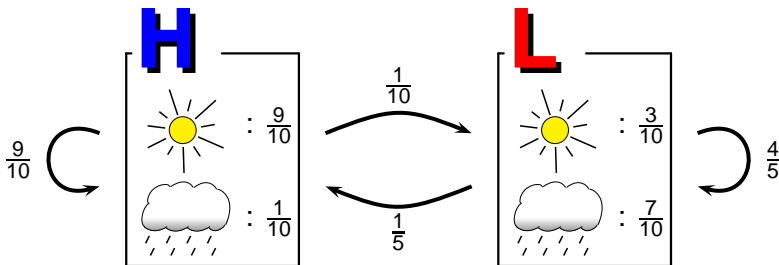


States of the Markov chain are not observed directly – observed characters are emitted from the hidden chain of states

Hidden Markov Models

More realism

Underlying unobserved phenomena may be the conserved factor



States of the Markov chain are not observed directly – observed characters are emitted from the hidden chain of states

Other types of states

Silent states have no emissions, and special start and end states results in distribution over *finite* observed sequences

Common HMM Uses

Annotation



Common HMM Uses

Annotation

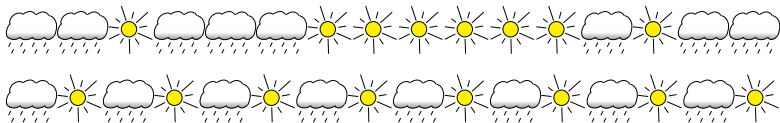


Common HMM Uses

Annotation



Classification

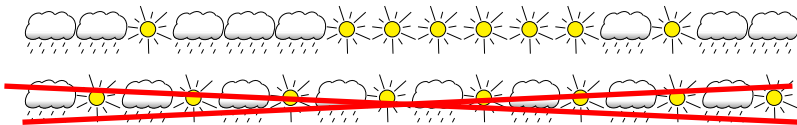


Common HMM Uses

Annotation



Classification



Naïve Approach

Observed ☁☁☀ – enumerate all possibilities:

Path	$Pr(\text{Path})$	$Pr(\text{Data} \mid \text{Path})$	$Pr(\text{Data})$
	$\pi_L a_{LL} a_{LL}$	$e_{L,\text{☁}} e_{L,\text{☁}} e_{L,\text{☀}}$	0.04704
	$\pi_L a_{LL} a_{LH}$	$e_{L,\text{☁}} e_{L,\text{☁}} e_{H,\text{☀}}$	0.03528
	$\pi_L a_{LH} a_{HL}$	$e_{L,\text{☁}} e_{H,\text{☁}} e_{L,\text{☀}}$	0.00021
	$\pi_L a_{LH} a_{HH}$	$e_{L,\text{☁}} e_{H,\text{☁}} e_{H,\text{☀}}$	0.00567
	$\pi_H a_{HL} a_{LL}$	$e_{H,\text{☁}} e_{L,\text{☁}} e_{L,\text{☀}}$	0.00084
	$\pi_H a_{HL} a_{LH}$	$e_{H,\text{☁}} e_{L,\text{☁}} e_{H,\text{☀}}$	0.00063
	$\pi_H a_{HL} a_{HL}$	$e_{H,\text{☁}} e_{H,\text{☁}} e_{L,\text{☀}}$	0.00014
	$\pi_H a_{HL} a_{HH}$	$e_{H,\text{☁}} e_{H,\text{☁}} e_{H,\text{☀}}$	0.00365

Number of paths grows exponentially!

Naïve Approach

Observed ☁☁☀ – enumerate all possibilities:

Path	$Pr(\text{Path})$	$Pr(\text{Data} \mid \text{Path})$	$Pr(\text{Data})$
L L L	$\pi_L a_{LL} a_{LL} a_{LL}$	$e_{L,\text{☁}} e_{L,\text{☁}} e_{L,\text{☁}}$	0.04704
L H L	$\pi_L a_{LH} a_{HL} a_{LL}$	$e_{L,\text{☁}} e_{H,\text{☁}} e_{L,\text{☁}}$	0.00021
H L L	$\pi_H a_{HL} a_{LL} a_{LL}$	$e_{H,\text{☁}} e_{L,\text{☁}} e_{L,\text{☁}}$	0.00084
H H L	$\pi_H a_{HH} a_{HL} a_{LL}$	$e_{H,\text{☁}} e_{H,\text{☁}} e_{L,\text{☁}}$	0.00014
L L H	$\pi_L a_{LL} a_{LL} a_{LH}$	$e_{L,\text{☁}} e_{L,\text{☁}} e_{L,\text{☀}}$	0.03528
L H H	$\pi_L a_{LH} a_{HH} a_{HH}$	$e_{L,\text{☁}} e_{H,\text{☁}} e_{H,\text{☀}}$	0.00567
H L H	$\pi_H a_{HL} a_{LL} a_{LH}$	$e_{H,\text{☁}} e_{L,\text{☁}} e_{L,\text{☀}}$	0.00063
H H H	$\pi_H a_{HH} a_{HH} a_{HH}$	$e_{H,\text{☁}} e_{H,\text{☁}} e_{H,\text{☁}}$	0.00365

Number of paths grows exponentially!

Naïve Approach

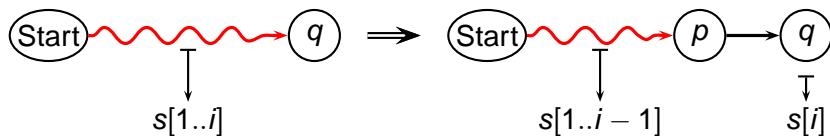
Observed ☁☁☀ – enumerate all possibilities:

Path	$Pr(\text{Path})$	$Pr(\text{Data} \mid \text{Path})$	$Pr(\text{Data})$
L L L	$\pi_L a_{LL} a_{LL}$	$e_{L,\text{☁}} e_{L,\text{☁}}$	0.04704
H L L	$\pi_H a_{HL} a_{LL}$	$e_{H,\text{☁}} e_{L,\text{☁}}$	0.00084
L H L	$\pi_L a_{LH} a_{HL}$	$e_{L,\text{☁}} e_{H,\text{☁}}$	0.00021
H H L	$\pi_H a_{HH} a_{HL}$	$e_{H,\text{☁}} e_{H,\text{☁}}$	0.00014
L L H	$\pi_L a_{LL} a_{LH}$	$e_{L,\text{☁}} e_{L,\text{☀}}$	0.03528
H L H	$\pi_H a_{HL} a_{LH}$	$e_{H,\text{☁}} e_{L,\text{☀}}$	0.00063
L H H	$\pi_L a_{LH} a_{HH}$	$e_{L,\text{☁}} e_{H,\text{☀}}$	0.00567
H H H	$\pi_H a_{HH} a_{HH}$	$e_{H,\text{☁}} e_{H,\text{☀}}$	0.00365

Number of paths grows exponentially!

Recursion...

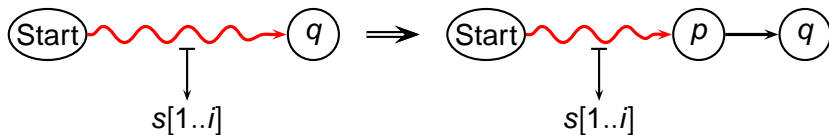
Last state is non-silent



$$V(q, i) = \max_p \{ V(p, i-1) \cdot Pr(p \rightarrow q) \cdot Pr(s[i] | q) \}$$

Recursion...

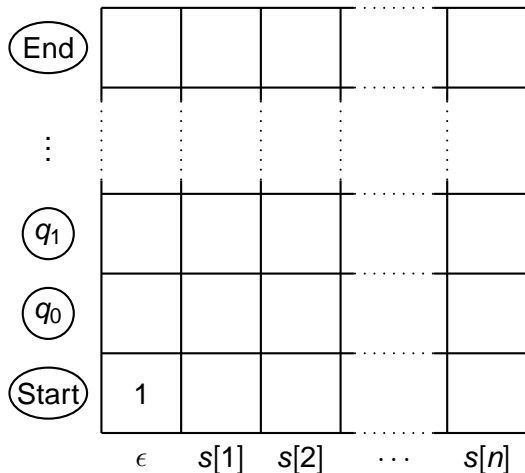
Last state is silent



$$V(q, i) = \max_p \{ V(p, i) \cdot Pr(p \rightarrow q) \}$$

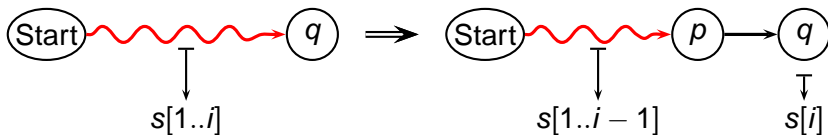
... and a Table

Can compute probabilities efficiently by dynamic programming



Forward Algorithm

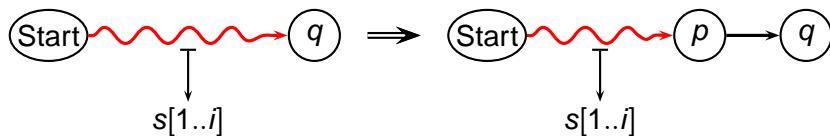
Last state is non-silent



$$F(q, i) = \sum_p F(p, i-1) \cdot P(p \rightarrow q) \cdot P(s[i] | q)$$

Forward Algorithm

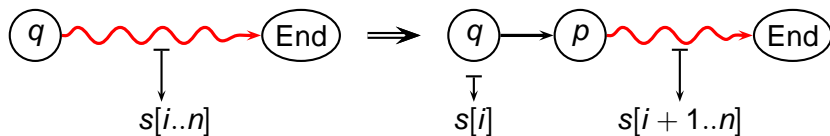
Last state is silent



$$F(q, i) = \sum_p V(p, i) \cdot P(p \rightarrow q)$$

Backward Algorithm

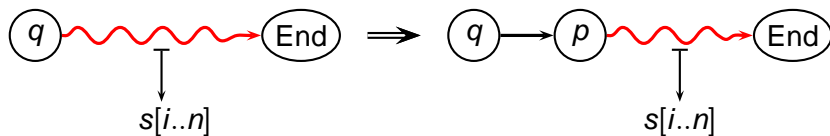
First state is non-silent



$$B(q, i) = \sum_p V(p, i+1) \cdot P(q \rightarrow p) \cdot P(s[i] | q)$$

Backward Algorithm

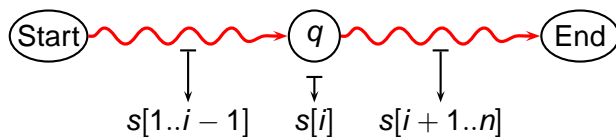
First state is non-silent



$$B(q, i) = \sum_p V(p, i) \cdot P(q \rightarrow p)$$

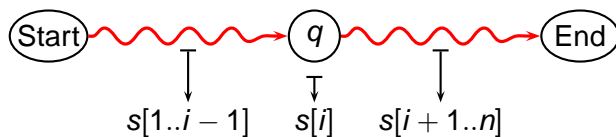
Maximum A Posteriori Decoding

Forward and backward algorithms give $P(s[i] | q)$:



Maximum A Posteriori Decoding

Forward and backward algorithms give $P(s[i] | q)$:



MAP Decoding Path

$$A(q, i) = \max_{p: P(p \rightarrow q) > 0} \{A(p, i-1) + P(s[i] | q)\}$$

HMM Applications

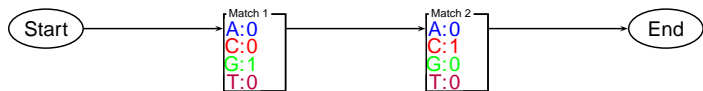
Bioinformatics

- Sequence alignment
- Gene finding
- Homology modelling
- Protein secondary structure prediction
- Genetic variation
- Inference of genealogical histories
- Nucleosome positioning

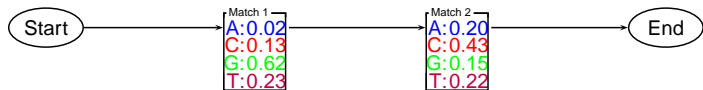
Speech Recognition

Signal Processing

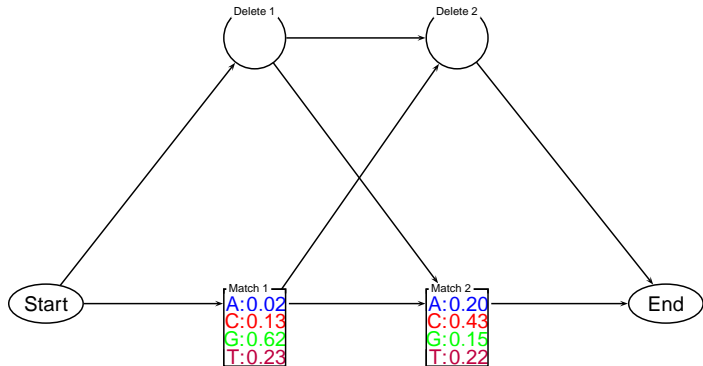
Sequence Alignment



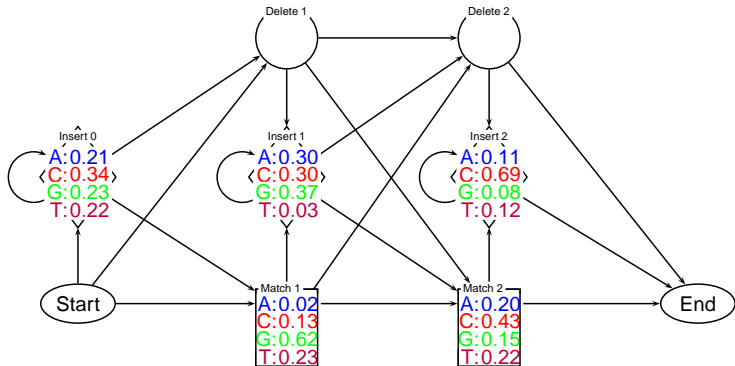
Sequence Alignment



Sequence Alignment



Sequence Alignment



Gene Finding

Elements of protein coding gene in eukaryotes includes

- Codon based structure of amino acid encoding in blocks of three nucleotides

Gene Finding

Elements of protein coding gene in eukaryotes includes

- Codon based structure of amino acid encoding in blocks of three nucleotides
- Special start and end codons

Gene Finding

Elements of protein coding gene in eukaryotes includes

- Codon based structure of amino acid encoding in blocks of three nucleotides
- Special start and end codons
- Exon/intron structure of gene

Gene Finding

Elements of protein coding gene in eukaryotes includes

- Codon based structure of amino acid encoding in blocks of three nucleotides
- Special start and end codons
- Exon/intron structure of gene
- Special motifs/compositions upstream and downstream of codign part

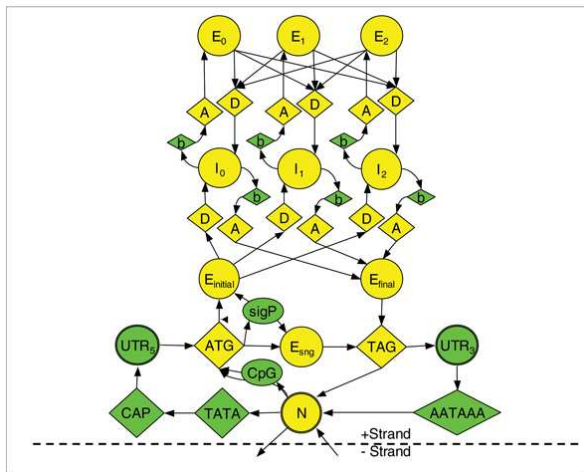
Gene Finding

Elements of protein coding gene in eukaryotes includes

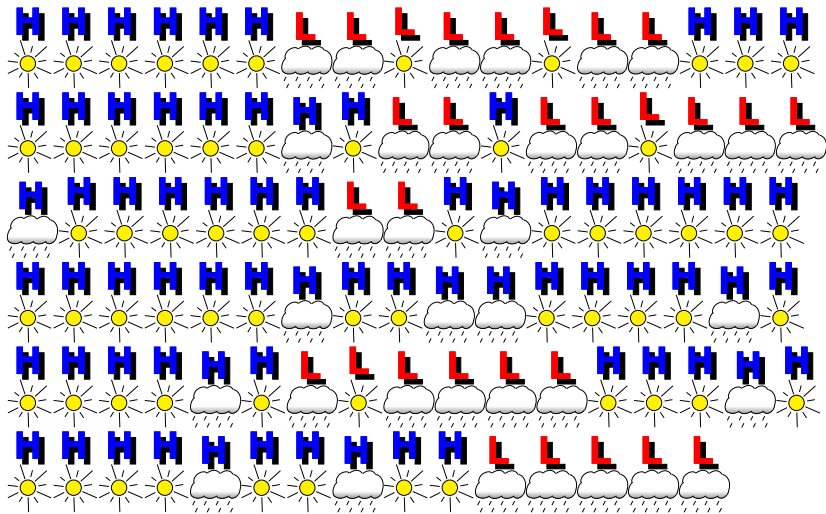
- Codon based structure of amino acid encoding in blocks of three nucleotides
- Special start and end codons
- Exon/intron structure of gene
- Special motifs/compositions upstream and downstream of codign part
- can be encoded on either strand of the chromosome

Gene Finding

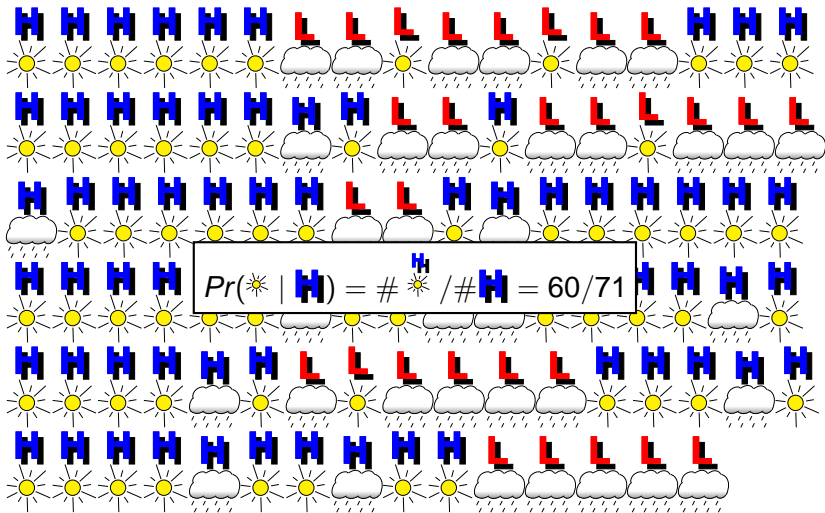
Elements of protein coding gene in eukaryotes includes



Annotated Data



Annotated Data



Unannotated Data

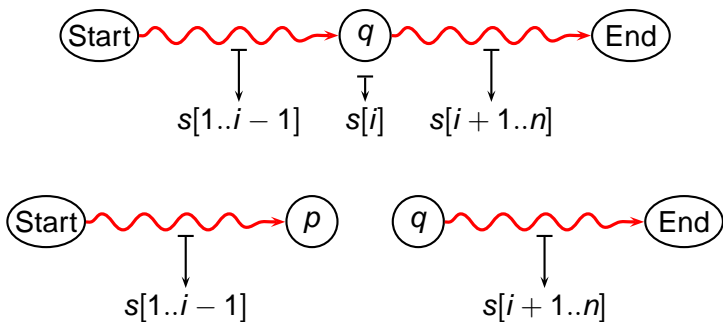
Use expectations instead of observed frequencies:

$$Pr(\odot | \mathbf{H}) = \mathbb{E} \left[\# \odot \right] / \mathbb{E} [\# \mathbf{H}]$$

Unannotated Data

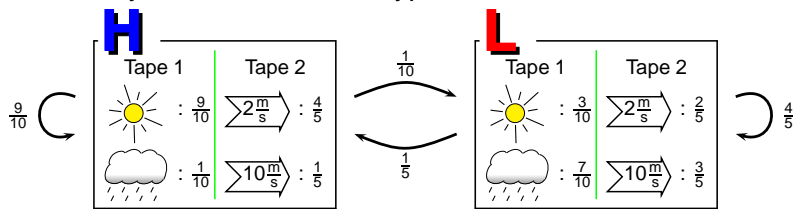
Use expectations instead of observed frequencies:

$$Pr(\odot | \mathbf{H}) = \mathbb{E} \left[\# \odot \right] / \mathbb{E} \left[\# \mathbf{H} \right]$$



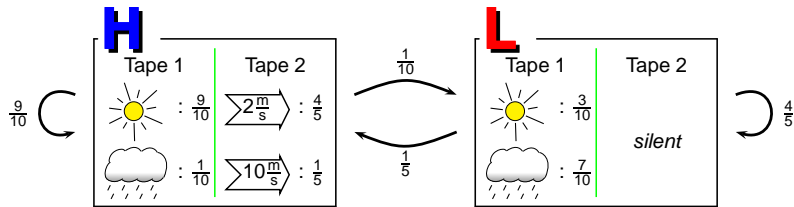
Generating Several Outputs

There may be more than one type of observables for each state



Generating Several Outputs

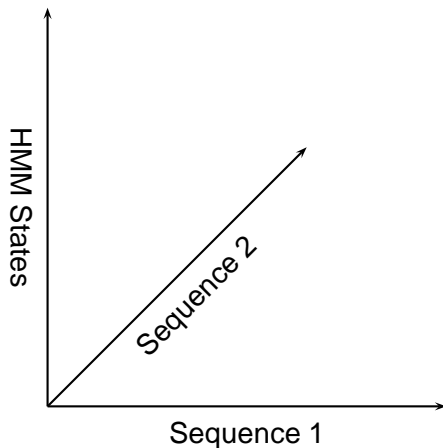
There may be more than one type of observables for each state



When states emit only some types of observables, the computing probabilities become more complex

Algorithms for Multitape HMMs

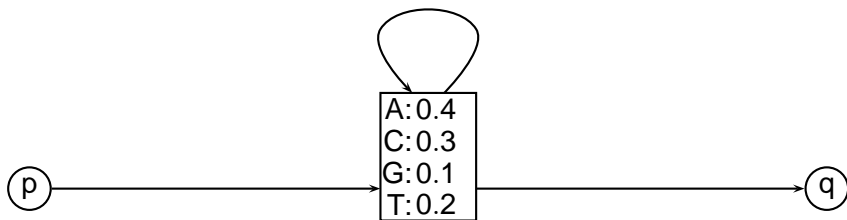
Same as for normal HMMs, just with more dimensions:



$$V(q, i, j) = \max_p \{ V(p, i-1, j-1) \cdot P(p \rightarrow q) P(s[i], t[j] | q) \}$$

Non-geometric Waiting Times

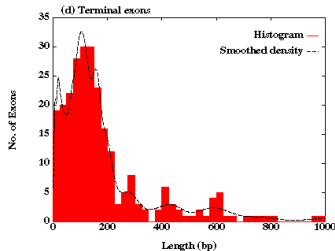
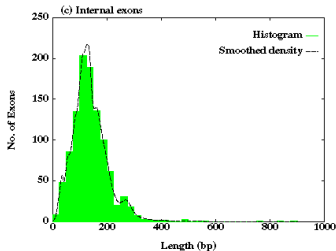
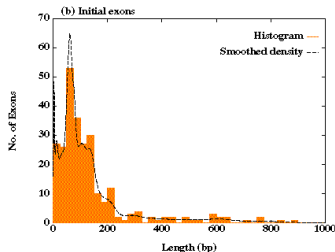
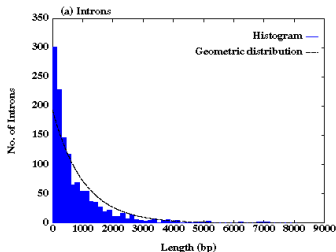
Waiting times in states are geometric



Non-geometric Waiting Times

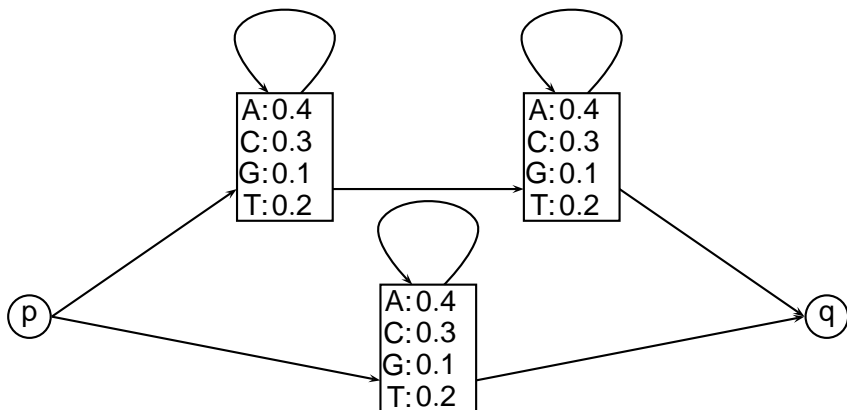
Waiting times in states are geometric

Length distributions of human introns and initial, internal and terminal exons



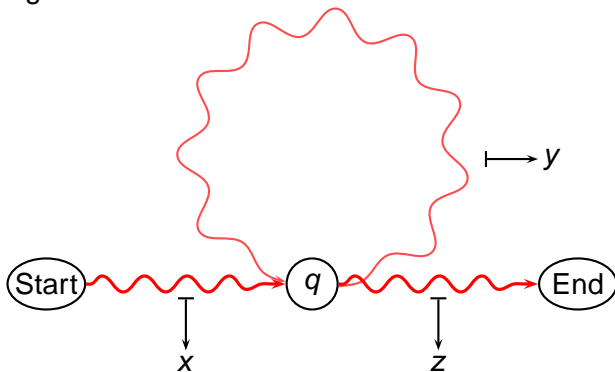
Non-geometric Waiting Times

Waiting times in states are geometric, but can be made more general by replicating states:



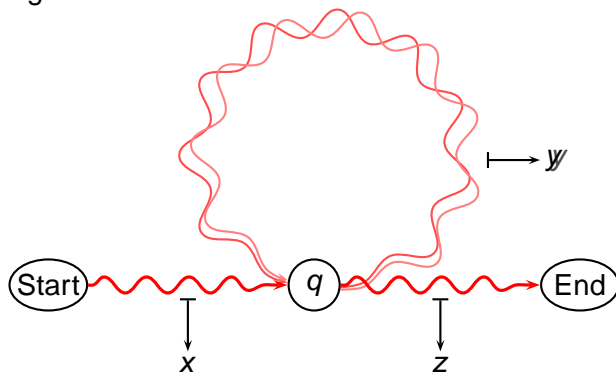
Limitations of Hidden Markov Models

For sufficiently long sequences, there will be a cycle in any path generating it from a fixed HMM



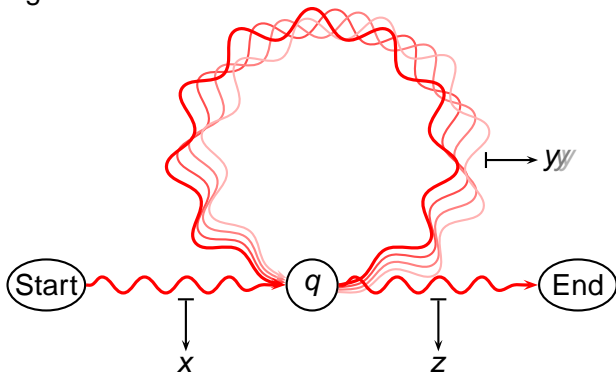
Limitations of Hidden Markov Models

For sufficiently long sequences, there will be a cycle in any path generating it from a fixed HMM



Limitations of Hidden Markov Models

For sufficiently long sequences, there will be a cycle in any path generating it from a fixed HMM



Consequence: Cannot generate e.g. palindromes and

$\{\text{☀}^i \text{☁}^i \mid i > 0\}$