Markov Models: Markov Chains

Michael Aněl

Department of ComputerScience, FEL ČVUT

Markov Models:

- Observable Markov Models
- Simple assignment *CpG-islands recognition* (5 pt.)
  - Motivation
  - Preparation for the main assignment

} 1 seminar

- Hidden Markov Models
- Basic algorithms
- Main assignment: *Gene finding* (15 pt.)

} 2 seminars

Gene Expression:

- Assignment: *Gene expression data analysis* (10 pt.)
- Modern approaches: Deep learning, sequencing...

} 2 seminars

Advanced Bioinformatics:

- Higher-order structures, gene-networks modelling...
- Voluntary assignment

} 2 – 3 seminars

Markov Models:

- Observable Markov Models
- Simple assignment *CpG-islands recognition* (5 pt.)
  - Motivation
  - Preparation for the main assignment

}1 seminar

- Hidden Markov Models
- Basic algorithms
- Main assignment: *Gene finding* (15 pt.)

}2 seminars

Gene Expression:

- Assignment: *Gene expression data analysis* (10 pt.)
- Modern approaches: Deep learning, sequencing...

}2 seminars

Advanced Bioinformatics:

- Higher-order structures, gene-networks modelling...
- Voluntary assignment

}2 – 3 seminars

Markov Models:

- Observable Markov Models
- Simple assignment *CpG-islands recognition* (5 pt.)
  - Motivation
  - Preparation for the main assignment

  } 1 seminar

- Hidden Markov Models
- Basic algorithms
- Main assignment: *Gene finding* (15 pt.)

  } 2 seminars

Gene Expression:

- Assignment: *Gene expression data analysis* (10 pt.)
- Modern approaches: Deep learning, sequencing...

  } 2 seminars

Advanced Bioinformatics:

- Higher-order structures, gene-networks modelling...
- Voluntary assignment

  } 2 – 3 seminars

¿ What are the specifics of sequence-like data?

¿ Is it optimal to employ relational paradigm for
   a) data storage,
   b) data mining?

¿ What is the Markov Model – Markov Chain?
   $\mathcal{M} = (\mathcal{A}, \mathcal{S}, P_t, P_{init})$, where:
   $\mathcal{A}$ ... alphabet, here $\mathcal{A} = \{a, c, t, g\}$
   $\mathcal{S}$ ... state space, here $\mathcal{S} = \mathcal{A}$
   $P_{init} : \mathcal{S} \rightarrow [0, 1]$ ... initial probabilities
   $P_t : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$ ... transition probabilities

¿ What does the *observable* mean?

¿ What are the specifics of sequence-like data?

¿ Is it optimal to employ relational paradigm for
  a) data storage,
  b) data mining?

¿ What is the Markov Model – Markov Chain?
  $\mathcal{M} = (\mathcal{A}, \mathcal{S}, P_t, P_{init})$, where:
  $\mathcal{A}$ ... alphabet, here $\mathcal{A} = \{a, c, t, g\}$
  $\mathcal{S}$ ... state space, here $\mathcal{S} = \mathcal{A}$
  $P_{init} : \mathcal{S} \to [0, 1]$ ... initial probabilities
  $P_t : \mathcal{S} \times \mathcal{S} \to [0, 1]$ ... transition probabilities

¿ What does the *observable* mean?

¿ What are the specifics of sequence-like data?

¿ Is it optimal to employ relational paradigm for
  a) data storage,
  b) data mining?

¿ What is the Markov Model – Markov Chain?

$\mathcal{M} = (\mathcal{A}, \mathcal{S}, P_t, P_{init})$, where:

$\mathcal{A}$ ... alphabet, here $\mathcal{A} = \{a, c, t, g\}$

$\mathcal{S}$ ... state space, here $\mathcal{S} = \mathcal{A}$

$P_{init} : \mathcal{S} \rightarrow [0, 1]$ ... initial probabilities

$P_t : \mathcal{S} \times \mathcal{S} \rightarrow [0, 1]$ ... transition probabilities

¿ What does the *observable* mean?

¿ What are the specifics of sequence-like data?

¿ Is it optimal to employ relational paradigm for
   a) data storage,
   b) data mining?

¿ What is the Markov Model – Markov Chain?
$\mathcal{M} = (\mathcal{A}, \mathcal{S}, P_t, P_{init})$, where:
$\mathcal{A}$ ... alphabet, here $\mathcal{A} = \{a, c, t, g\}$
$\mathcal{S}$ ... state space, here $\mathcal{S} = \mathcal{A}$
$P_{init} : \mathcal{S} \to [0, 1]$ ... initial probabilities
$P_t : \mathcal{S} \times \mathcal{S} \to [0, 1]$ ... transition probabilities

¿ What does the *observable* mean?

¿ What is the main advantage of Markov Model (MM)?

let $\mathbf{x} = <x_1, x_2, \ldots, x_L> \in A^L$ be a sequence made of $A \in \mathcal{A}$

then $P(\mathbf{x}) = P(x_1, x_2, \ldots, x_L) = P(x_1)P(x_2|x_1)\ldots P(x_L|x_{L-1})$

¿ What is the main advantage of Markov Model (MM)?
let $\mathbf{x} = <x_1, x_2, \ldots, x_L> \in A^L$ be a sequence made of $A \in \mathcal{A}$
then $P(\mathbf{x}) = P(x_1, x_2, \ldots, x_L) = P(x_1)P(x_2|x_1)\ldots P(x_L|x_{L-1})$

Observable MM, an example:
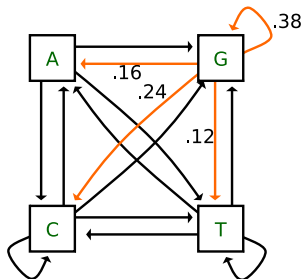
transition probabilities

$$P(x_i = a \mid x_{i-1} = g) = 0.16$$
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$
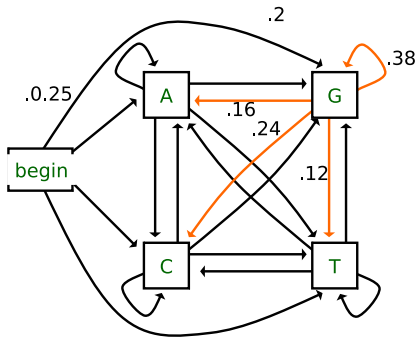$$P(x_i = g \mid x_{i-1} = g) = 0.38$$
$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

¿ What do you miss to compute the probability of a sequence?

Markov Chains – A Short Quiz
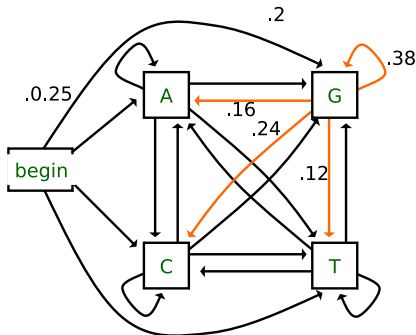
Adding a *silent* BEGIN state:

transition probabilities
$$P(x_i = a \mid x_{i-1} = g) = 0.16$$
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$
$$P(x_i = g \mid x_{i-1} = g) = 0.38$$
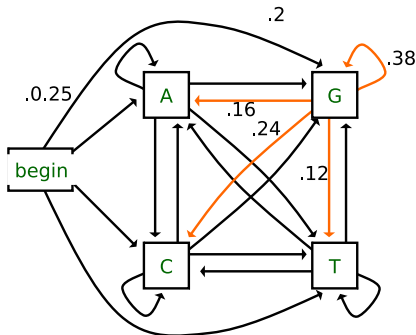$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

¿ How to adjust the MM formalism?

$\mathcal{M} = (\mathcal{A}, \mathcal{S}_{\text{init}}, \mathcal{S}, P_{\text{t}})$, where $\mathcal{S} = \mathcal{A} \cup \mathcal{S}_{\text{init}}$

¿ How long can be the sequences generated?

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:

transition probabilities

$P(x_i = a \mid x_{i-1} = g) = 0.16$

$P(x_i = c \mid x_{i-1} = g) = 0.34$

$P(x_i = g \mid x_{i-1} = g) = 0.38$

$P(x_i = t \mid x_{i-1} = g) = 0.12$

¿ How to adjust the MM formalism?
$\mathcal{M} = (\mathcal{A}, S_{\mathsf{init}}, \mathcal{S}, P_{\mathsf{t}})$, where $\mathcal{S} = \mathcal{A} \cup S_{\mathsf{init}}$

¿ How long can be the sequences generated?

Adding a *silent* BEGIN state:

transition probabilities
$$P(x_i = a \mid x_{i-1} = g) = 0.16$$
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$
$$P(x_i = g \mid x_{i-1} = g) = 0.38$$
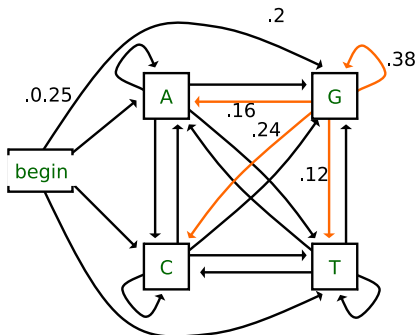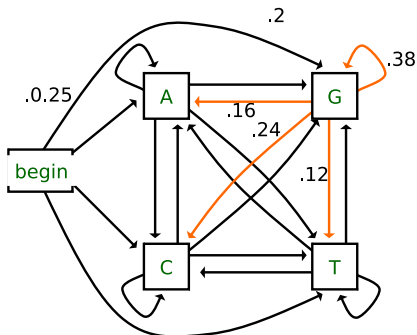$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

¿ How to adjust the MM formalism?
$\mathcal{M} = (\mathcal{A}, S_{\text{init}}, \mathcal{S}, P_{\text{t}})$, where $\mathcal{S} = \mathcal{A} \cup S_{\text{init}}$

¿ How long can be the sequences generated?

Adding a *silent* BEGIN state:

transition probabilities
$$P(x_i = a \mid x_{i-1} = g) = 0.16$$
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$
$$P(x_i = g \mid x_{i-1} = g) = 0.38$$
$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

¿ How to adjust the MM formalism?
$\mathcal{M} = (\mathcal{A}, \mathcal{S}_{\text{init}}, \mathcal{S}, P_{\text{t}})$, where $\mathcal{S} = \mathcal{A} \cup \mathcal{S}_{\text{init}}$

Adding a *silent* BEGIN state:

transition probabilities

$P(x_i = a \mid x_{i-1} = g) = 0.16$

$P(x_i = c \mid x_{i-1} = g) = 0.34$
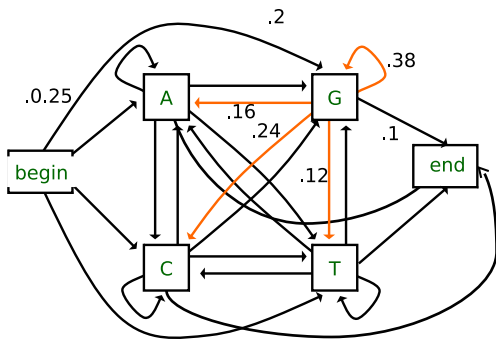
$P(x_i = g \mid x_{i-1} = g) = 0.38$

$P(x_i = t \mid x_{i-1} = g) = 0.12$

¿ How to adjust the MM formalism?

$\mathcal{M} = (\mathcal{A}, S_{\text{init}}, \mathcal{S}, P_{\text{t}})$, where $\mathcal{S} = \mathcal{A} \cup S_{\text{init}}$

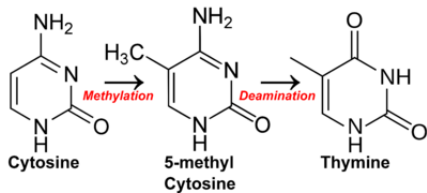Adding a *silent* END state:

source: Mark Craven

Markov Chains: Learning

Simply, learning the probabilities:

- $P(a) = \frac{\#('a')+1}{\#('*')+5}$

- $P(a|c) = \frac{\#('ca')+1}{\#('c*')+5}$

- $P(\text{end}|c) = \frac{\#('c\backslash n')+1}{\#('c*')+5}$

¿ What are the CpG islands?

¿ Why do we call them 'CpG'?

¿ What is *CG content*?

¿ Given that the CG content in the human genome is 41%, what CpG frequency would we expect?
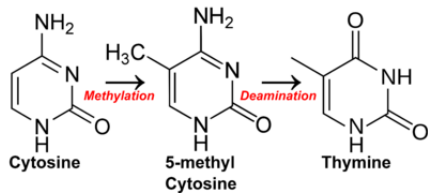

source: wikipedia.org

¿ What are the CpG islands?

¿ Why do we call them 'CpG'?

¿ What is *CG content*?

¿ Given that the CG content in the human genome is 41%, what CpG frequency would we expect?

source: wikipedia.org
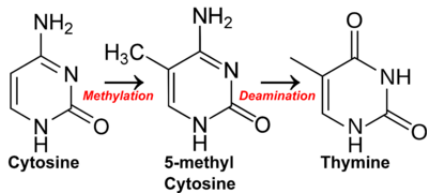
¿ What are the CpG islands?

¿ Why do we call them 'CpG'?

¿ What is *CG content*?

¿ Given that the CG content in the human genome is 41%, what CpG frequency would we expect?

source: wikipedia.org
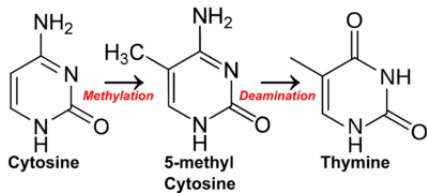
¿ What are the CpG islands?

¿ Why do we call them 'CpG'?

¿ What is *CG content*?

¿ Given that the CG content in the human genome is 41%, what CpG frequency would we expect?

source: wikipedia.org

General Classification Task on MM:

- Given two sets of sequences $\{\mathbf{x}_i \in \mathcal{A}^*\}_{i=1|class}^{N}$ originated from two different *classes* (e.g. *class* $\in \{\mathrm{CpG}, \mathrm{null}\}$ for the CpG regions and rest of the genome, respectively)

- Learn two Markov models approximating these *distribution* $P(\mathbf{x}|class)$:
  e.g.: $P(\mathbf{x}|\mathrm{CpG}) = P(x_1, x_2, \ldots, x_L|\mathrm{CpG}) =$
  $= P_{\mathrm{CpG}}(x_1)P_{\mathrm{CpG}}(x_2|x_1)P_{\mathrm{CpG}}(x_3|x_2)\ldots P_{\mathrm{CpG}}(x_L|x_{L-1})$

- Decide for an unseen $\mathbf{x}_{\mathrm{new}}$ sequence its belonging:

  IF $P(class_1|\mathbf{x}_{\mathrm{new}}) > P(class_2|\mathbf{x}_{\mathrm{new}})$ THEN *class*1 ELSE *class*$_2$

- $P(c_1|\mathbf{x}) > P(c_2|\mathbf{x}) \iff P(c_1)P(\mathbf{x}|c_1) > P(c_2)P(\mathbf{x}|c_2)$

Assignment: CpG-islands Recognition

1. Implement a function which learns a MM based on a set of training sequences.
2. Learn the two models on the sequences form `cpg_train.txt` and `null_train.txt`
3. Enumerate the accuracy of your classifier (models) according to the test sequences `seqs_test.txt` and appropriate labels `classes_test.txt` ('1' stands for CpG, '0' for the rest)