Markov Models: Markov Chains

Michael Anděl

Department of ComputerScience, FEL ČVUT

Seminars – Timeline

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

Seminars – Timeline

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

Seminars – Timeline

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 Chains – A Short Quiz

- 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?

- What does the *observable* mean?

Markov Chains – A Short Quiz

- ¿ 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?
- ¿ What does the *observable* mean?

Markov Chains – A Short Quiz

- ¿ 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?

- ¿ What does the *observable* mean?

Markov Chains – A Short Quiz

- ¿ 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?

- ¿ What does the *observable* mean?

Markov Chains – A Short Quiz

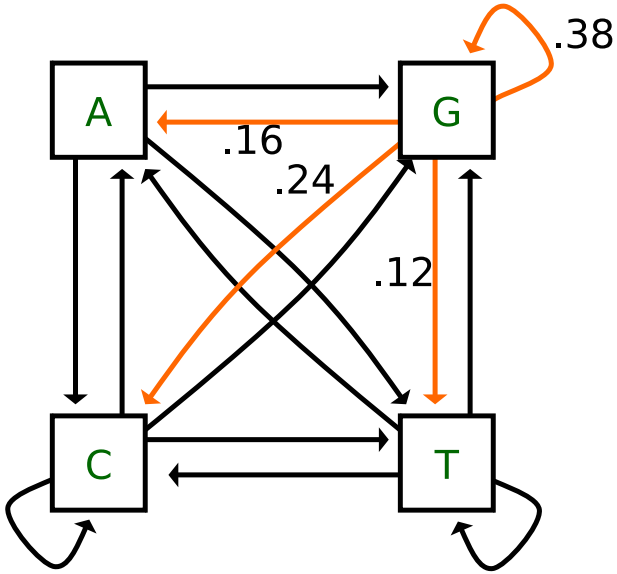
¿ What is the main advantage of Markov Model (MM)?

Markov Chains – A Short Quiz

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

Markov Chains – A Short Quiz

Observable MM, an example:



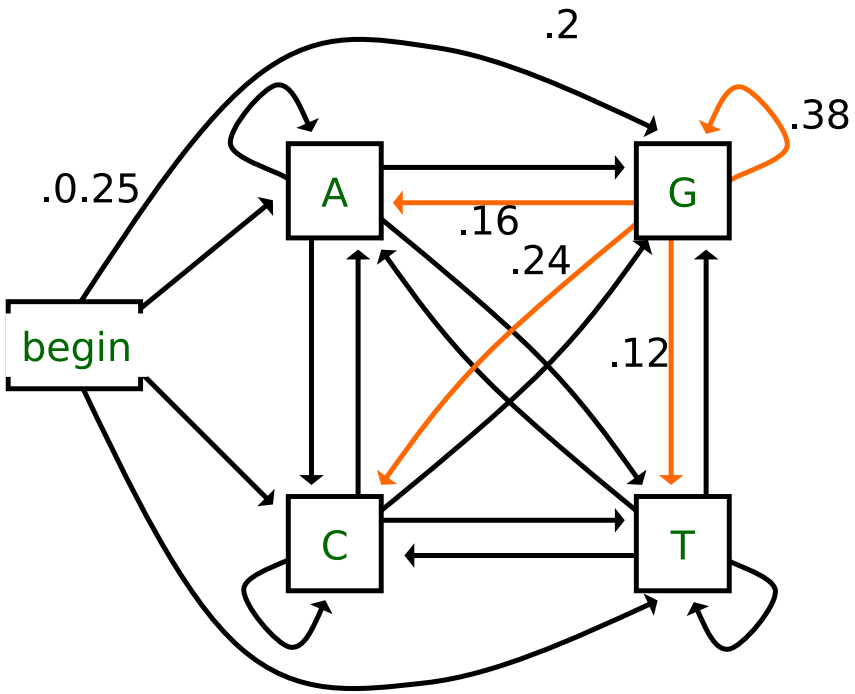
transition probabilities

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

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

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:



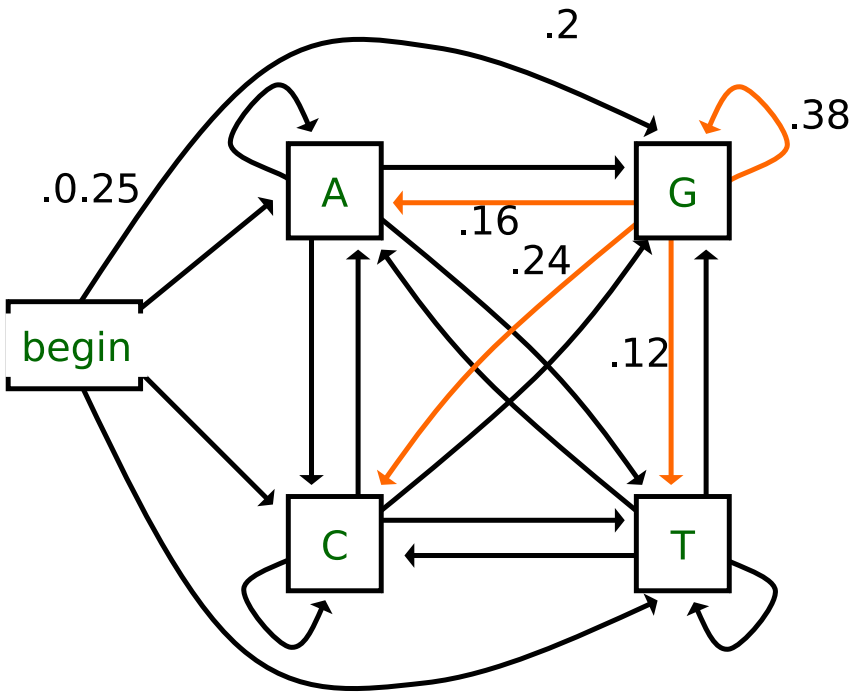
transition probabilities

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

- How to adjust the MM formalism?
- How long can be the sequences generated?

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:



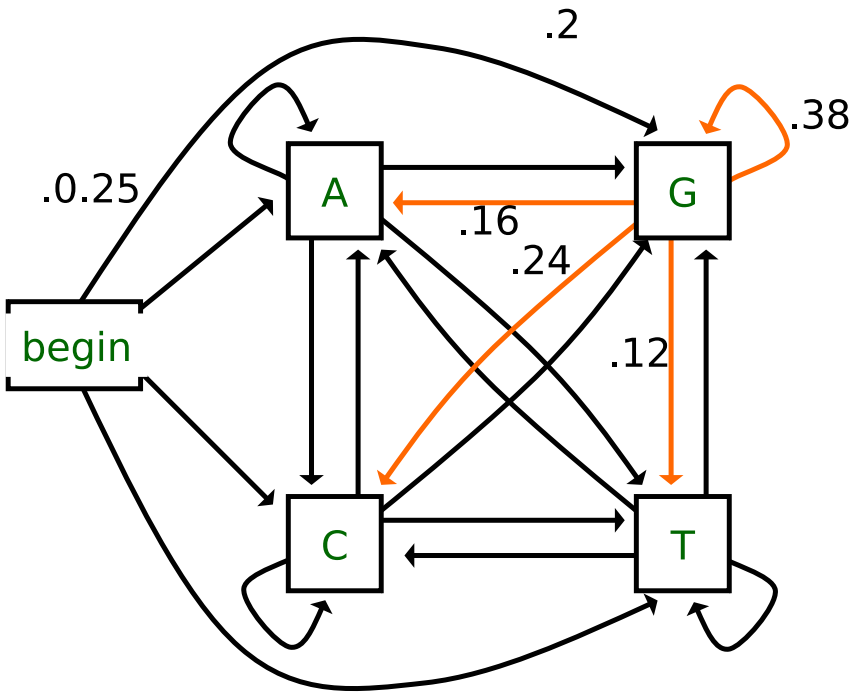
transition probabilities

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

- How to adjust the MM formalism?
- How long can be the sequences generated?

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:



transition probabilities

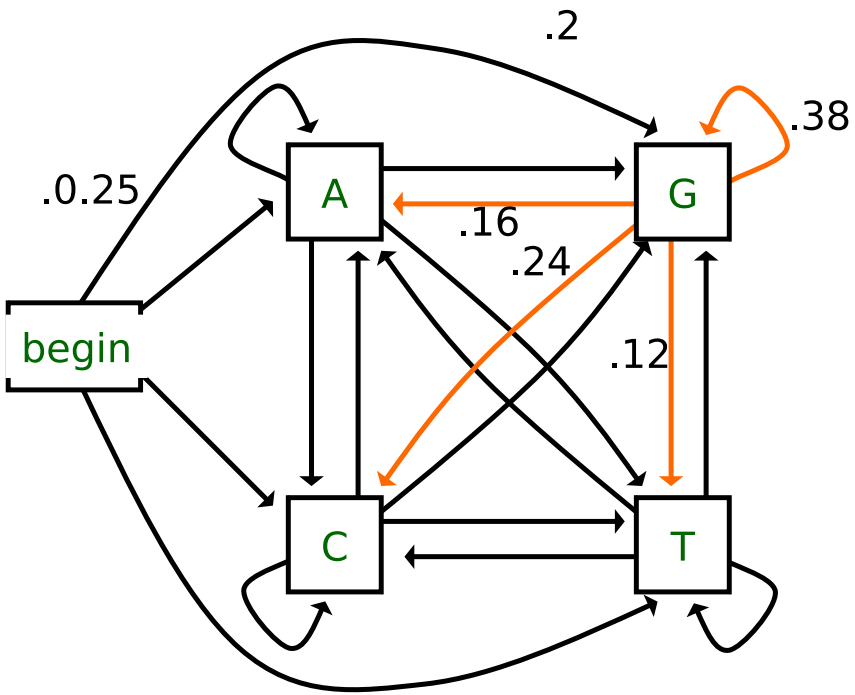
$$P(x_i = a \mid x_{i-1} = g) = 0.16$$
$$P(x_i = c \mid x_{i-1} = g) = 0.34$$
$$P(x_i = g \mid x_{i-1} = g) = 0.38$$
$$P(x_i = t \mid x_{i-1} = g) = 0.12$$

¿ How to adjust the MM formalism?

¿ How long can be the sequences generated?

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:



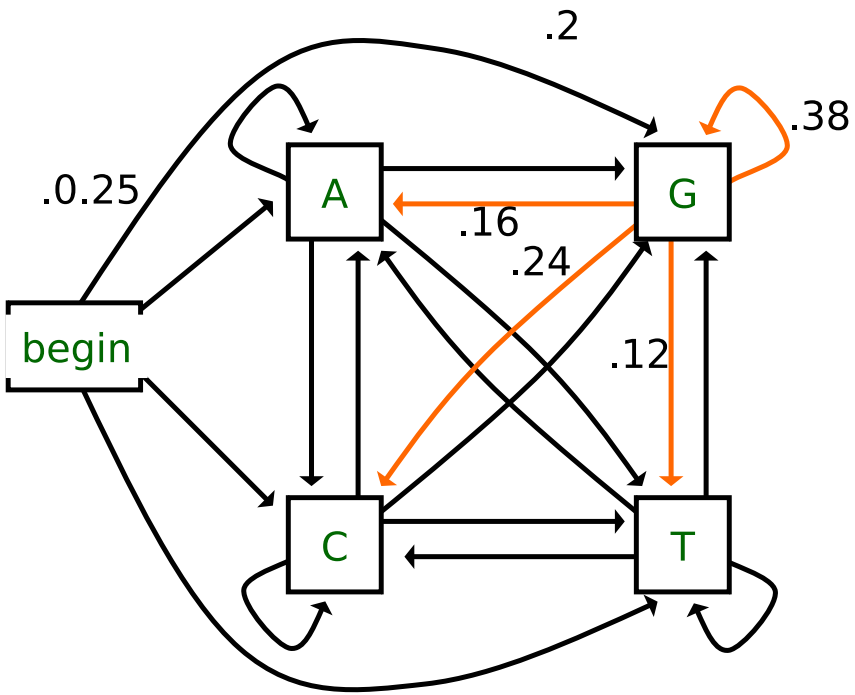
transition probabilities

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

¿ How to adjust the MM formalism?

Markov Chains – A Short Quiz

Adding a *silent* `BEGIN` state:



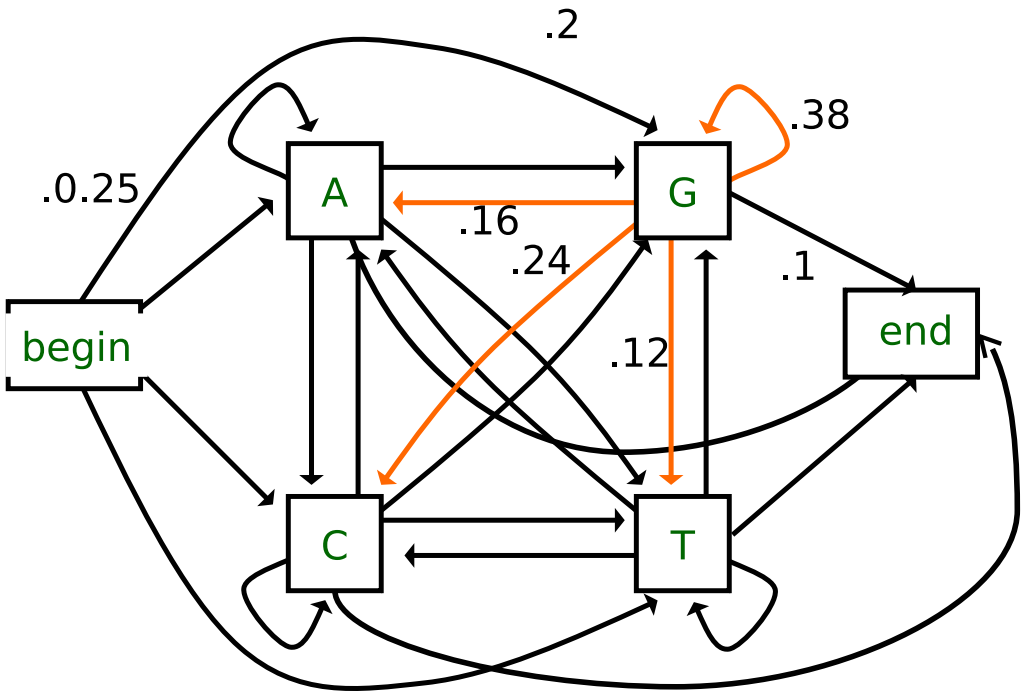
transition probabilities

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

¿ How to adjust the MM formalism?

Markov Chains – A Short Quiz

Adding a *silent* `END` state:



source: Mark Craven

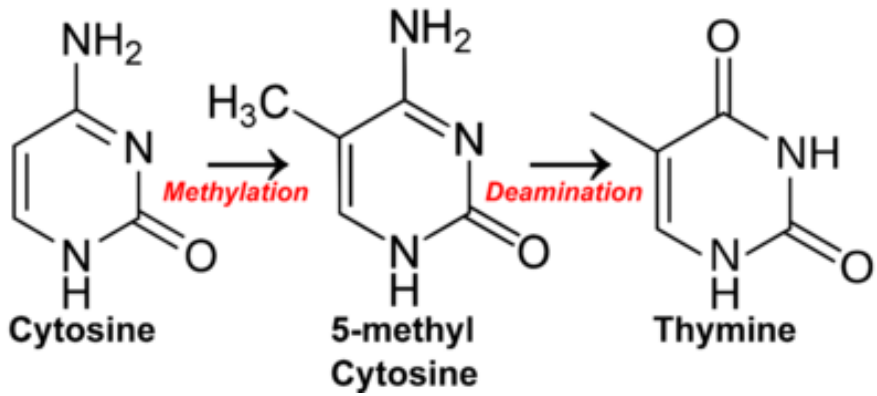
Markov Chains: Learning

Simply, learning the probabilities:

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

Motivation: CpG islands

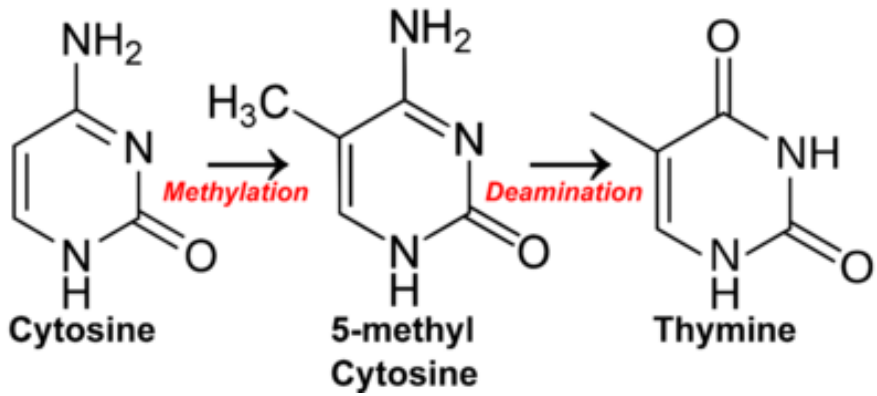
- What are the CpG islands?
- Why do we call them 'CpG'?
- What is *CG content*?
- Given that the CG content in the human genome is 41%, what CpG frequency would we expect?



source: wikipedia.org

Motivation: CpG islands

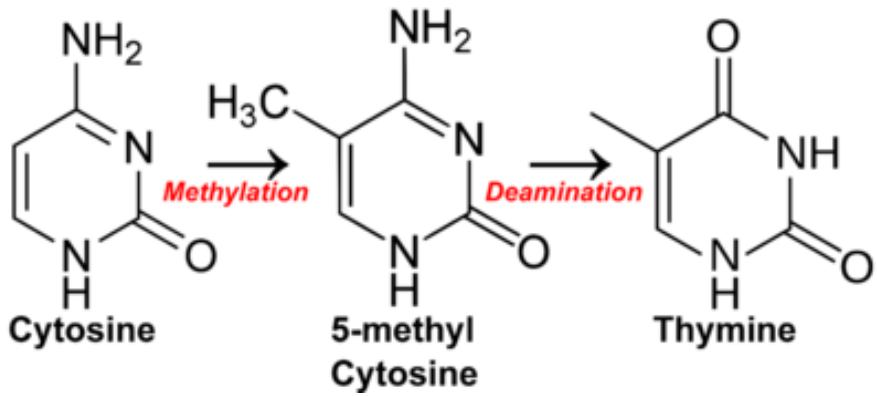
- What are the CpG islands?
- Why do we call them 'CpG'?
- What is *CG content*?
- Given that the CG content in the human genome is 41%, what CpG frequency would we expect?



source: wikipedia.org

Motivation: CpG islands

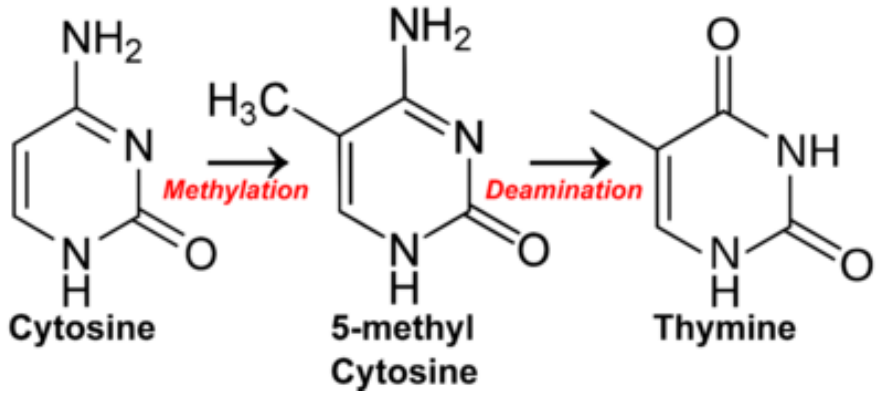
- What are the CpG islands?
- Why do we call them 'CpG'?
- What is *CG content*?
- Given that the CG content in the human genome is 41%, what CpG frequency would we expect?



source: wikipedia.org

Motivation: CpG islands

- 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

Motivation: CpG islands

General Classification Task on MM:

- Given two sets of sequences $\{\mathbf{x}_i\}_{i=1|class}^{N}$ originated from two different *families* (e.g. the CpG regions and rest of the genome)
- Learn two Markov models approximating these two *distribution* $P(\mathbf{x}|class)$
- Decide for an unseen $\mathbf{x}_{\text{new}}$ sequence its belonging:

$$\text{IF } P(class_1|\mathbf{x}_{\text{new}}) > P(class_2|\mathbf{x}_{\text{new}}) \text{ THEN } class1 \text{ ELSE } class_2$$

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)