# Course Outline

Lecture 01.03

# Requirements

- Assignments – 30 %
- Midterm exams – 30 %
- Project – 30 %

- In-class quizzes – to monitor comprehension

# The topic

Computable problems in molecular biology and algorithmic solutions to these problems

# Your goal:

- To be familiar with the problems of modern molecular biology
- To be able to identify which of these problems are computable
- To use algorithmic tools to solve these problems

A side effect: understanding the ideas behind bioinformatics tools, i.e. *what* problem does the tool solve, and *how* it solves the problem.

# Background

- Main concepts of the molecular biology
- Algorithms, data structures, probability

# Molecular biology - definition

- Based on theories of living things in terms of chemical matter (molecules) and mechanisms
- Studies macromolecules – DNA, RNA, protein – and the mechanisms of their interaction

# Bioinformatics - definition

- Applies concepts of informatics and computer science to the field of molecular biology – to extract new knowledge from the information encoded in the genetic code

# Sample Bioinformatics problem

- Input:
  - Query: sequence of bases in a DNA molecule:
    - AACCCTTAG
  - The set of sequences of known genes:
    - ACCTAG
    - AGCCCGTA
    - AAGCCGCTTA
- Biological question: which is the most similar to the query sequence?

# Which is the most similar?

1

| |
|---|
| AACCCTTAG |
| ACCTAG |

2

| |
|---|
| AACCCTTAG |
| AGCCCGTA |

3

| |
|---|
| AACCCTTAG |
| AAGCCGCTTA |

# Protocol of solving a bioinformatics problem

1. Biological question (find *similar* sequences)
2. Formalization (how to measure *similarity*)
3. An *efficient* algorithm to solve the *formalized* problem
4. Model + learning – to learn the parameters of an algorithm from real data
5. Evaluation of results – distinguish (statistically) significant results from artifacts
6. Presentation of the results

# Another example

- Input: four DNA sequences taken from four species.

AAG

AAA

AGA

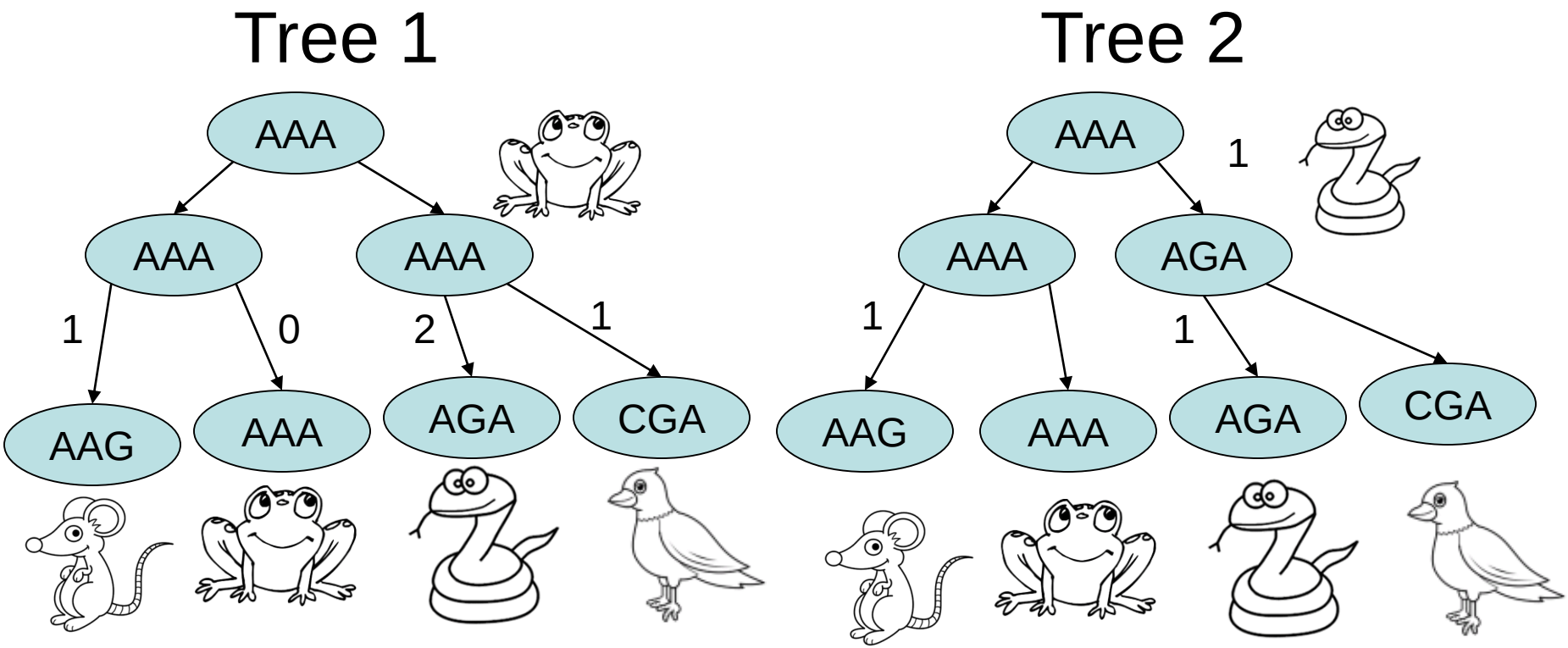
GGA

# Formalization

- 1. Biological question: which evolutionary tree *best* explains these sequences ?
- 2. Formalization: what is the measure for *the best* tree?

Let it be *the parsimony principle*: Pick a tree that has a minimum total number of substitutions of symbols between species and their originator in the evolutionary tree.

# Many possible trees




What tree is better by the parsimony principle?

# Next steps

3. Efficient algorithm: how can we compute the best tree efficiently ?
4. Adjusting parameters from the data: A is more likely to be replaced by G or by T?
5. Significance: is the best tree found significantly (statistically) better than others ?
6. Present results as a tree

The main question: does the tree make biological sense ?

# Molecular biology problems we are going to look at in this course

- Sequence comparison
- Gene finding
- Sequence-based evolution

# Algorithmic Tools outline

- Discrete algorithms:
  - Combinatorial pattern matching
  - String indexing
  - Dynamic programming
- Probabilistic models:
  - Hidden Markov Models
  - Maximum likelihood
  - Bayesian inference
- Hard problems:
  - Heuristics
  - Approximation algorithms

# Familiarize yourself with the object of our study – molecules of life

- DNA
- RNA
- Proteins