Algorithms in Molecular Biology

Course Outline

Information

- Lectures:

TWF 10:30 - 11:20 DSB C126

- Lecturers: Marina Barsky / Ulrike Stege
- Office hours:

TW, 2.00-3.30 PM ECS 617

- E-mail: mgbarsky@csc.uvic.ca
  ustege@csc.uvic.ca

- Site: https://connex.csc.uvic.ca/portal/site/a31ada3d-7a2e-4a0c-936e-a93f672be3ce

Requirements

- Assignments – 40 %
- Midterm exams – 30 %
- Project – 30 %

- Quzzes – 0.5% bonus each

Background

- Main concepts of molecular biology
- Algorithms, data structures
- Probability theory

The topic

Algorithms in molecular biology.

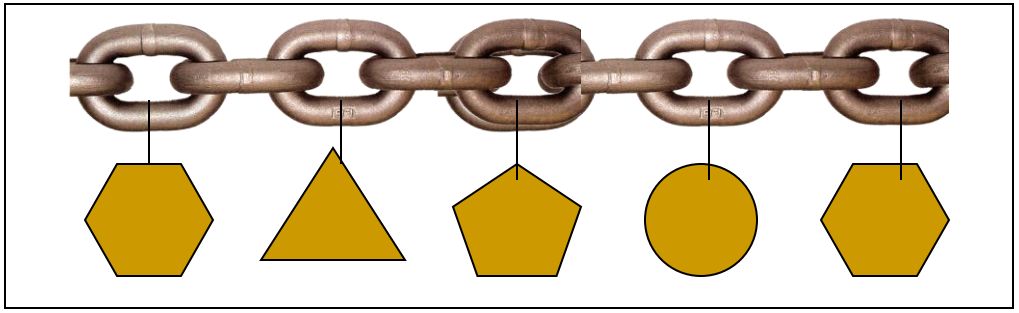
Bioinformatics algorithms

Molecular biology - definition

- Molecular biology considers living things in terms of chemical matter (molecules) and mechanisms

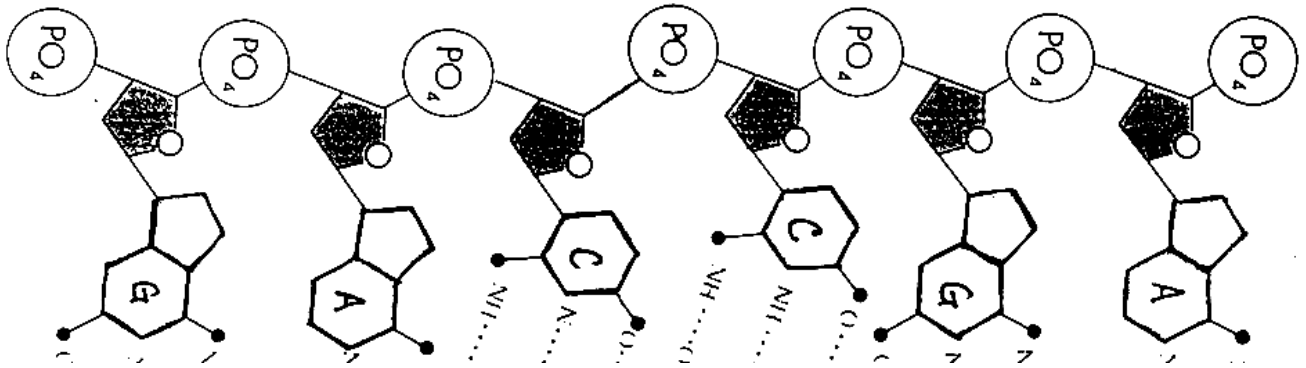
Macromolecules

- The main molecules are DNA, RNA, and protein



Bioinformatics - definition

- Applies concepts of *informatics* and *computer science* to the field of molecular biology – to extract new *knowledge* from the *information* encoded in *biosequences*
- *Biosequence* is an abstraction of ordered information encoded in macromolecules (nucleic acids and proteins)



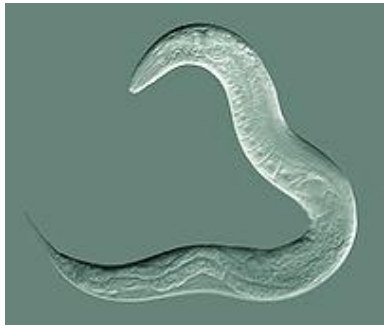
The objectives

- To be familiar with the problems of the 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.

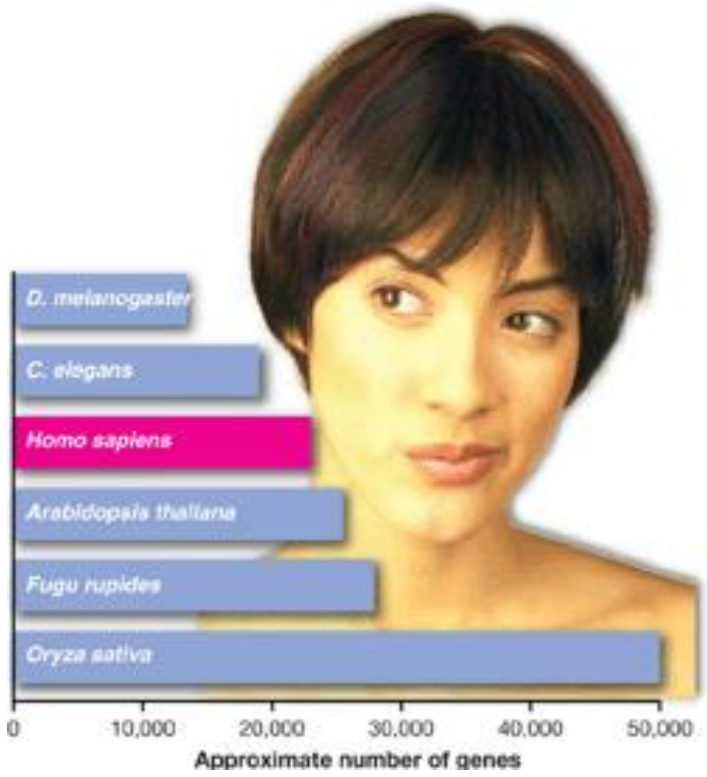
Example: how many genes are there?

- Estimated in 2001 count of human genes – 20,000 – 25,000 (the simple roundworm Caenorhabditis elegans has 20,000 genes, rice has 43,000 genes)



Example: how many genes are there?

- Estimated in 2001 count of human genes – 20,000 – 25,000
- Others estimate as 65,000 - 75,000 [ref]
  - Overestimate - prediction based on the sequence itself (*ab initio*)
  - Underestimate - based on the comparison with known genes
- Why human gene count is so low?



Sample Bioinformatics problem

- Input:
  - Query DNA sequence of an unknown gene:
    - AACCCTTAG
  - The sequences of known genes:
    - ACCTAG
    - AGCCCGTA
    - AAGCCGCTTA
- Biological question: find among these sequences the most similar to the query sequence

What pair is the most similar?

1	A	A	C	C	C	T	T	A	G	
	A	C	C	T	A	G				

2	A	A	C	C	C	T	T	A	G	
	A	G	C	C	C	G	T	A		

3	A	A	C	C	C	T	T	A	G	
	A	A	G	C	C	G	C	T	T	A

What pair is the most similar?

Local similarity

1	A	A	C	C	C	T	T	A	G	
		A	C	C			T	A	G	

Overall (global) similarity

2	A	A	C	C	C	T	T	A	G	
	A	G	C	C	C	G	T	A		

Does the deletion of the symbols matter?

3	A	A		C	C		C	T	T	A	G
	A	A	G	C	C	G	C	T	T	A	

What pair is the most similar?

What if only ACC and TAG determine the shape

(therefore, the functionality) of the encoded protein?

1	A	A	C	C	C	T	T	A	G	
		A	C	C			T	A	G	

2	A	A	C	C	C	T	T	A	G	
	A	G	C	C	C	G	T	A		

3	A	A		C	C		C	T	T	A	G
	A	A	G	C	C	G	C	T	T	A	

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 significant (statistically) results from artifacts
6. Presentation of the results

Another example

- Input: four DNA sequences taken from four species.

AAG

AAA

AGA

GGA

Formalization

AAG

AAA

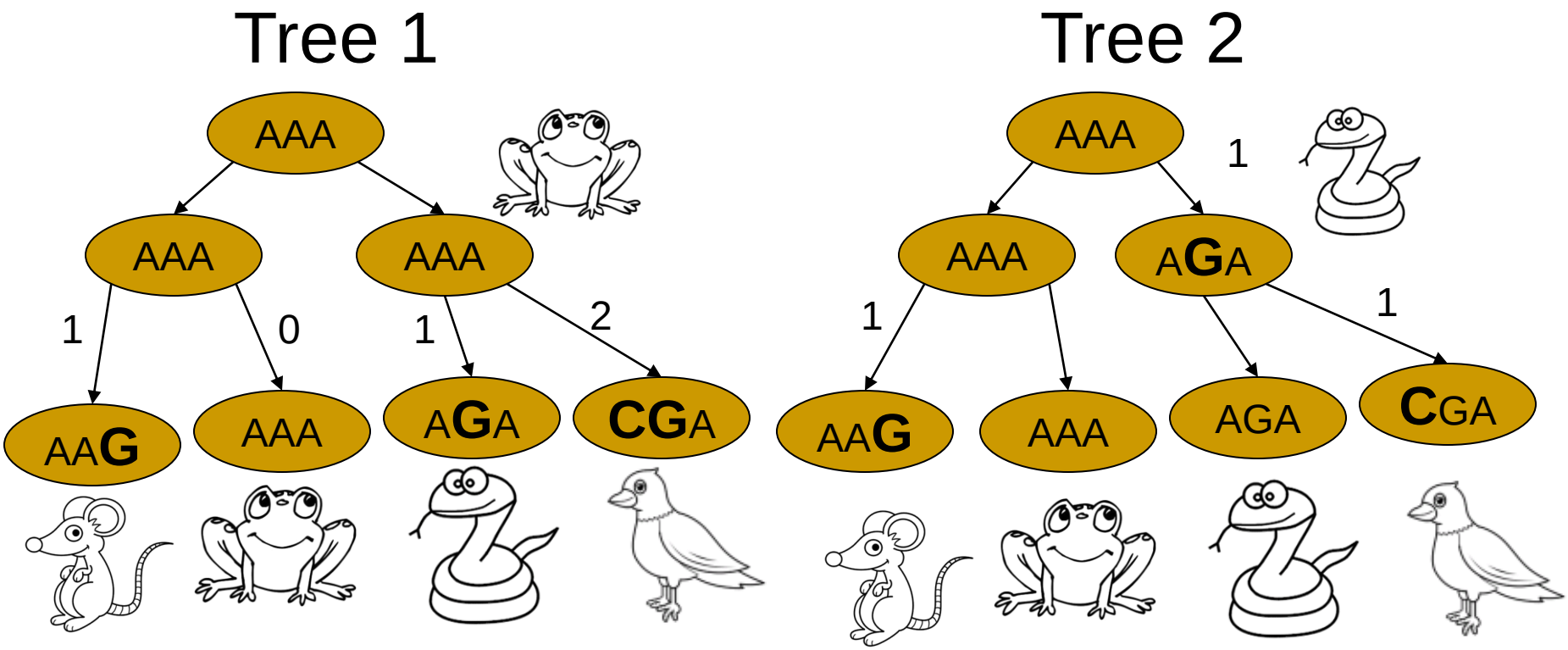
AGA

GGA

- 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 symbol changes between species and their originator in the evolutionary tree.

Many possible trees



Which 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 ?

The scope of the problems

- Sequence comparison
- Gene finding
- Sequence-based evolution
- Sequence folding
- Gene expression profiles

The scope of algorithms

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

The closer look at the object of our study – molecules of life

- DNA
- RNA
- Proteins

More reading

- **Molecular Biology (Stanford Encyclopedia of Philosophy)**

  http://plato.stanford.edu/entries/molecular-biology/
- **Beginner's Guide to Molecular Biology**

  http://www.rothamsted.bbsrc.ac.uk/notebook/courses/guide/