Genomes and Algorithms

Computer analysis of genetic information

François Rechenmann

Genomes and Algorithms

1. Genomic texts
2. Genes and proteins
3. Gene prediction
4. **Sequence comparison**
5. Phylogenic tree construction

François Rechenmann

4. Sequence comparison

- **How to predict gene/protein functions?**
- Why gene/protein sequences may be similar?
- Measuring sequence similarity
- Aligning sequences is an optimization problem
- A sequence alignment as a path
- A path is optimal if all its sub-paths are optimal
- Alignment costs
- A recursive algorithm
- Recursion can be avoided: an iterative version
- How efficient is this algorithm?

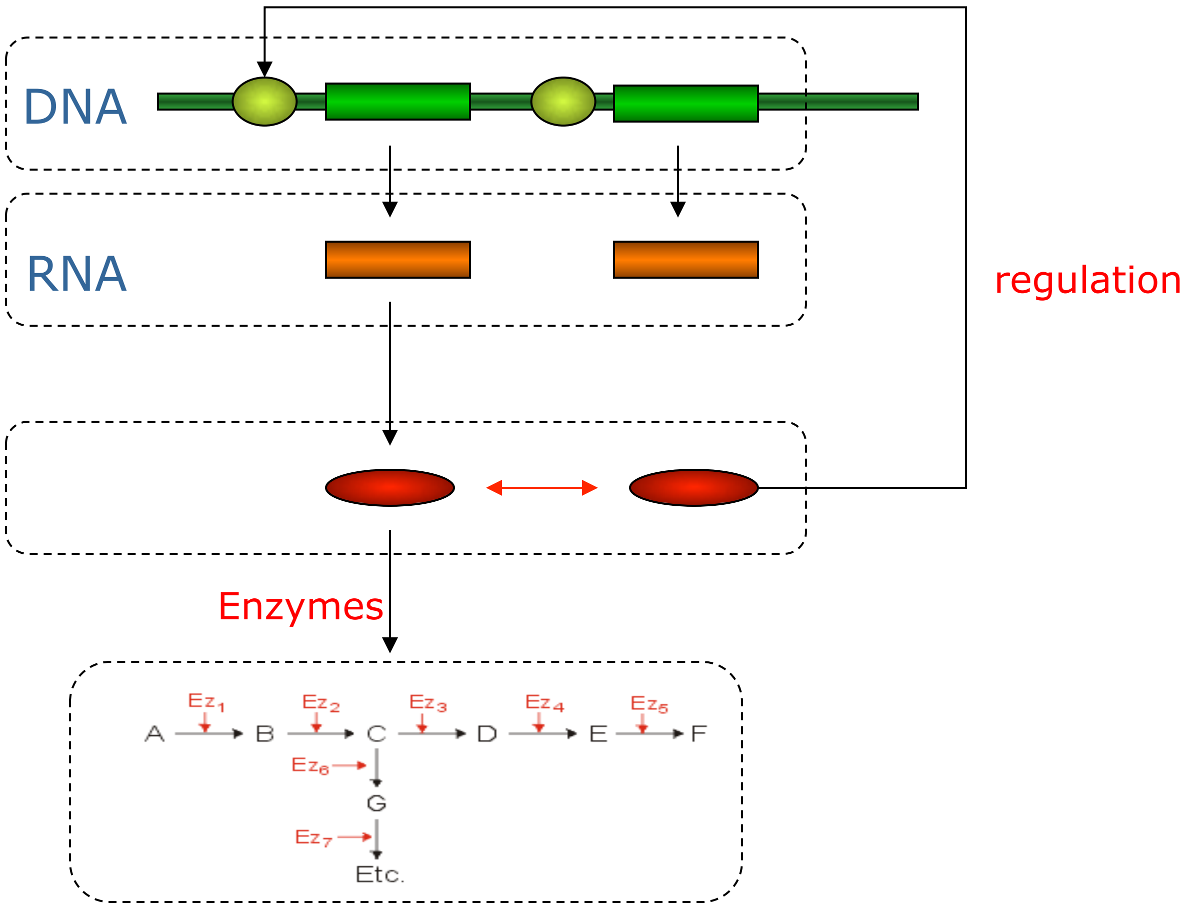
Gene/protein databases

- GeneBank, UniProt,…
- Sequence of a gene/protein is associated with several types of information
- Information on the functions
  - Free text
  - Keywords
  - Enzymatic classification entries

Genes

Proteins

Metabolic reactions and pathways



Gene/protein databases

- Public databases: GeneBank, UniProt,…
- Sequence of a gene/protein is associated with several types of information
- Information on the functions
  - Free text
  - Keywords
  - Enzymatic classification entries
    - ✓ tripeptide aminopeptidases: EC 3.4.11.4

Tables extracted from databases

Sequence	Organism	Function (EC number)
ACCGTTACG…	E.coli	3.4.11.4
ACTTTTGCC…	B. subtilis	2.3.4
TGGTATGCT…	H. influenzae	4.1.1.3

Function prediction

- Start from a gene/protein sequence
- Search in the first column of the file for a similar sequence
- When a similar sequence is found, record the associated information
- Continue the search until end of file

Function prediction

- Start from a gene/protein sequence
- Search in the first column of the file for a similar sequence
- When a similar sequence is found, record the associated information
- Continue the search until end of file

How can the similarity between two sequences be measured?