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?**

The number of comparisons is quadratic

- Needleman and Wunsch algorithm (1970)

- For two sequences of length N and M, the number of comparisons is O (N*M)
- The algorithmic complexity is said to be quadratic, $O(N^2)$
- If the lengths of the sequences double, the computation time is expected to increase 4 times
  - $N \times M \rightarrow (2\times N) \times (2\times M) = 4 \times (N\times M)$

- The same type of algorithm on 3 sequences: O (N^3)!

- Dedicated algorithms
  for multiple alignment (more than 2 sequences)

Blast

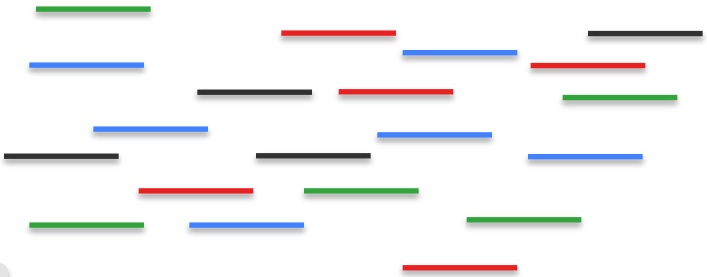
- The most frequently used bioinformatics method
- Relies on non-exact matches
- Search for similar sequences in databases

- Metagenomics analyses
  - Sequencing all the DNA in a sample

Blast

- The most frequently used bioinformatics method
- Relies on non-exact matches
- Search for similar sequences in databases

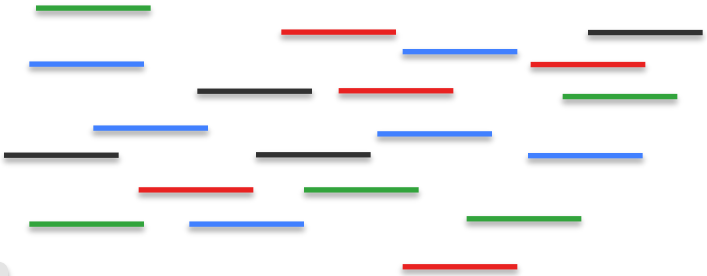
- Metagenomics analyses
  - Sequencing all the DNA in a sample



Blast

- The most frequently used bioinformatics method
- Relies on non-exact matches
- Search for similar sequences in databases

- Metagenomics analyses
  - Sequencing all the DNA in a sample
  - “Blast” each read, to identify the species of the sample



Blast

- The most frequently used bioinformatics method
- Relies on non-exact matches
- Search for similar sequences in databases

- Metagenomics analyses
  - Sequencing all the DNA in a sample
  - “Blast” each read, to identify the species of the sample