

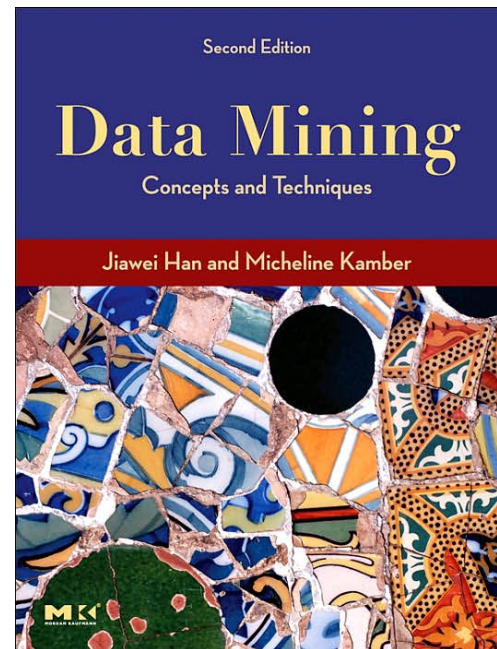


# Data Mining for Biological Data Analysis (Part II)

Data Mining and Text Mining (UIC 583 @ Politecnico di Milano)

# References

- Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems (Second Edition)
  - ▶ Chapter 8



# Biological Sequence Alignment

# Alignment of biological sequence (1)

- ❑ Given two or more input biological sequences, **identify similar sequences with long conserved subsequences**
- ❑ Sequences can be either nucleotides (DNA/RNA) or amino acids (proteins)
  - ▶ Nucleotides align if they are identical
  - ▶ Amino acids align if identical or if one can be derived from the other
- ❑ **Tasks**
  - ▶ Pairwise sequence alignment
  - ▶ Multiple sequence alignment
- ❑ **Applications**
  - ▶ Discovering phylogenetic trees
  - ▶ Similarity searches

## Alignment of biological sequence (2)

- ❑ Substitution matrix is used to define
  - ▶ cost of substitutions
  - ▶ cost of insertions and deletions
- ❑ Cost is inversely proportional to the probability that a substitution/insertion/deletion occurred
- ❑ Gaps (“—”) can be used to indicate positions where it is preferable not to align two symbols
- ❑ The introduction of a gap (“—”) is usually associated to a negative cost (**penalty**)

# Example

Align the following sequences:

HEAGAWGHEE  
PAWHEAE

Evaluate the following alignments according to the substitution matrix provided and the a gap penalty of -8

	A	E	G	H	W
A	5	-1	0	-2	-3
E	-1	6	-3	0	-3
H	-2	0	-2	10	-3
P	-1	-1	-2	-2	-4
W	-3	-3	-3	-3	15

<i>H</i>	<i>E</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>W</i>	<i>G</i>	<i>H</i>	<i>E</i>	-	<i>E</i>
<i>P</i>	-	<i>A</i>	-	-	<i>W</i>	-	<i>H</i>	<i>E</i>	<i>A</i>	<i>E</i>
-2	-8	+5	-8	-8	+15	-8	+10	+6	-8	+6 = 0
<i>H</i>	<i>E</i>	<i>A</i>	<i>G</i>	<i>A</i>	<i>W</i>	<i>G</i>	<i>H</i>	<i>E</i>	-	<i>E</i>
-	-	<i>P</i>	-	<i>A</i>	<i>W</i>	-	<i>H</i>	<i>E</i>	<i>A</i>	<i>E</i>
-8	-8	-1	-8	+5	+15	-8	+10	+6	-8	+6 = +1

# Pairwise sequence alignment

- ❑ Two major approaches
  - ▶ **Local alignment**, works on segments and merge them
  - ▶ **Global alignment**, works on entire sequence
- ❑ Global alignment approaches search for the optimal alignment starting from optimal subsequences
- ❑ *Needleman-Wunsch* and *Smith-Waterman* algorithms exploit **dynamic programming** to find the optimal solution
- ❑ Both these algorithm have a **computational complexity** that is **quadratic w.r.t. sequences length!**
- ❑ Local alignment approaches (e.g. BLAST and FASTA) may be not able to find the best alignment but are more suitable to deal with long sequences

# BLAST

- ❑ BLAST breaks the sequences in small fragments called **words**
- ❑ A word is a **k-tuple** of elements (typically 11 nucleotides or 3 amino acids)
- ❑ BLAST first builds an **hash tables** of **neighborhood** words, that are closely matching
- ❑ Starting from a fragment, the alignment is extended in both the direction by choosing the best scoring matches
- ❑ BLAST has computationally complexity linear w.r.t. to the sequence length
- ❑ Several specialized versions of BLAST have been introduced
  - ▶ Protein similarity searches (BLASTP)
  - ▶ Variable word size (BLASTN)
  - ▶ Not contiguous alignments (MEGABLAST)

- ❑ Is important both in phylogenetic analysis and in the discovery of protein structures
- ❑ Multiple alignment is computationally more challenging
- ❑ **Freng-Doolittle alignment**
  - ▶ Performs the pairwise alignments
  - ▶ Merge them following a guide tree generated with a hierarchical clustering approach
- ❑ **Hidden Markov Models**
  - ▶ More sophisticated probabilistic approach to represent statistical regularities in the sequences