Pair-wise Algorithm

Protein Sequence Alignment and Database Searching

- **Alignment of Two Sequences (Pair-wise Alignment)**
  - The Scoring Schemes or Weight Matrices
  - Techniques of Alignments
  - DOTPLOT
- **Multiple Sequence Alignment (Alignment of > 2 Sequences)**
  - Extending Dynamic Programming to more sequences
  - Progressive Alignment (Tree or Hierarchical Methods)
  - Iterative Techniques
    - Stochastic Algorithms (SA, GA, HMM)
    - Non Stochastic Algorithms
- **Database Scanning**
  - FASTA, BLAST, PSIBLAST, ISS
- **Alignment of Whole Genomes**
  - MUMmer (Maximal Unique Match)

Pair-Wise Sequence Alignment

Scoring Schemes or Weight Matrices

- Identity Scoring
- Genetic Code Scoring
- Chemical Similarity Scoring
- Observed Substitution or PAM Matrices
- PEP91: An Update Dayhoff Matrix
- BLOSUM: Matrix Derived from Ungapped Alignment
- Matrices Derived from Structure

Techniques of Alignment

- Simple Alignment, Alignment with Gaps
- Application of DOTPLOT (Repeats, Inverse Repeats, Alignment)
- Dynamic Programming (DP) for Global Alignment
- Local Alignment (Smith-Waterman algorithm)

Important Terms

- Gap Penalty (Opening, Extended)
- PID, Similarity/Dissimilarity Score
- Significance Score (e.g. Z & E)

Aligning biological sequences

- Nucleic acid (4 letter alphabet + gap)

  `TT-GCAC`
  `TTTACAC`

- Proteins (20 letter alphabet + gap)

  `RKVA--GMAKPNM`
  `RKIAVAAASKPAV`

Problem

- Any two sequences can always be aligned
- There are many possible alignments
- Sequence alignment needs to be scored to find the „optimal“ alignment
- In many cases there will be several solutions with the same score

```
 ACGTACGTACGTACGTACGTACGTACGT
 |   |   |   |   |   |   |
GATCGATCGATCGATCGATCGATCGATC
```

```
ACGTACGTACGTACGTACGTACGTACGT
  |   |   |   |   |   |   |
 GATCGATCGATCGATCGATCGATCGATC
```

```
   ACGTACGTACGTACGTACGTACGTACGT
      |   |   |   |   |   |   |
    GATCGATCGATCGATCGATCGATCGATC
```

```
ACGTACGTACGTACGTACGTACGTACGT
  |   |   |   |   |   |   |
 GATCGATCGATCGATCGATCGATCGATC
```

```
ACGTACGTACGTACGTACGTACGTACGT
     |   |   |   |   |   |
  GATCGATCGATCGATCGATCGATCGATC
```

```
ACCGGTACGTTACGATACGTAACGTTACTGTACTGT
   | |   |       |   |  | |
  GATCGATCGATCGATCGATCGATCGATC
```

Question:
what is „similar“
enough to be relevant ?

What is sequence alignment

Given two sentences of letters (strings), and a scoring scheme for evaluating matching letters, find the optimal pairing of letters from one sequence to letters of the other sequence

```
Align:
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
THIS IS A SHORT SENTENCE
```

```
THIS IS A RATHER LONGER - SENTENCE THAN THE NEXT
|||| || | --*|-- -|---| - |||||||| ---- --- ----
THIS IS A --SH-- -O---R T SENTENCE ---- ---  ----
```

or

```
THIS IS A RATHER LONGER SENTENCE THAN THE NEXT
|||| || | ------ ------ |||||||| ---- --- ----
THIS IS A SHORT- ------ SENTENCE ---- --- ----
```

Dynamic Programming

- Dynamic Programming allow Optimal Alignment between two sequences
- Allow Insertion and Deletion or Alignment with gaps
- Needlman and Wunsh Algorithm (1970) for global alignment
- Smith & Waterman Algorithm (1981) for local alignment
- Important Steps
  - Create DOTPLOT between two sequences
  - Compute SUM matrix
  - Trace Optimal Path