Bioinformatics (MVE360)

Course Organiser: Graham Kemp

http://www.cse.chalmers.se/edu/year/2013/course/MVE360/

Assessment

- Grades will be determined by a written exam at the end of the course.
- But in order to pass the course you must also submit solutions to specified exercises.

Learning outcome/goals

- understand the use of bioinformatics in addressing a range of biological questions
- describe how bioinformatics methods can be used to relate sequence, structure and function
- discuss the technologies for modern high-throughput DNA sequencing and their applications
- use and understand some central bioinformatics data and information resources
- know principles and algorithms of pairwise and multiple alignments, and sequence database searching
- perform pattern matching in biomolecular sequences
- describe how evolutionary relationships can be inferred from sequences (phylogenetics)
- understand the most important principles in gene prediction methods
- know basic principles of hidden Markov models and their application in sequence analysis
- understand and implement solutions to basic bioinformatics problems

Introduction to bioinformatics

> "*Bioinformatics:* Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."
>
> "Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful."

Working definition by the NIH Biomedical Information Science and Technology Initiative Consortium, 2000
http://www.bisti.nih.gov/docs/CompuBioDef.pdf

Content

The course covers basic methods used in sequence analysis such as pairwise and multiple alignment, searching databases for sequence similarity, profiles, pattern matching, hidden Markov models, RNA bioinformatics, gene prediction methods and principles for molecular phylogeny.

The course includes modern high-throughput sequencing techniques and their applications, as well as molecular biology databases and different systems to query such databases.

The course considers theoretical principles as well as how existing programs are being used by bioinformaticians.

Sequence alignment

Comparison of macromolecular sequences.

Nucleic acids (DNA, RNA) or proteins.

Assignment of nucleotide-nucleotide or residue-residue correspondences.

Suggest evolutionary, structural and functional relationships.

Rigorous algorithms for global and local alignment.

Heuristic algorithms for practical database searching.

Dotplots

A pictorial representation of the similarity between two sequences.

Compare a sequence with itself:
- Repeats
- Palindromic sequences

Compare two sequences:
- Any path from upper left to lower right represents an alignment.
- Horizontal or vertical moves correspond to gaps in one of the sequences.
- Path with highest score corresponds to an optimal alignment.

Measures of sequence similarity

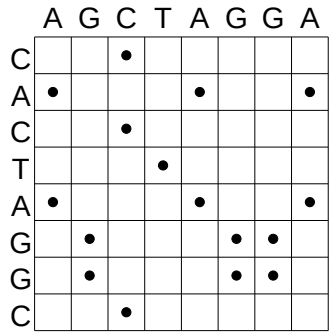
Hamming distance:
- Number of positions with mismatching characters.
- Defined for two strings of equal length.

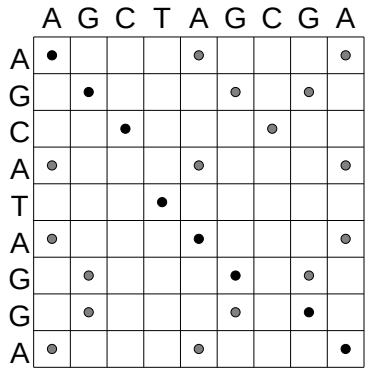
```
agtc
cgta
```

Levenshtein distance:
- Minimum number of edit operations (delete, insert, change a single character) needed to change one sequence into another.

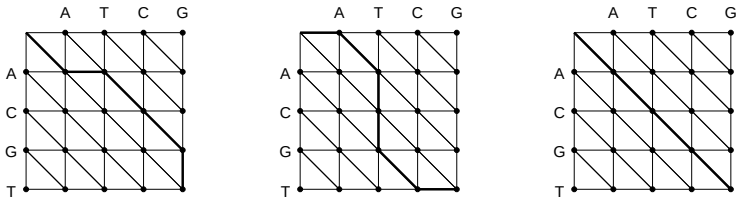
```
agtcc
cgctca
```


Dotplots





Each path represents an alignment




```
A-CGT
| ||
ATCG-
```



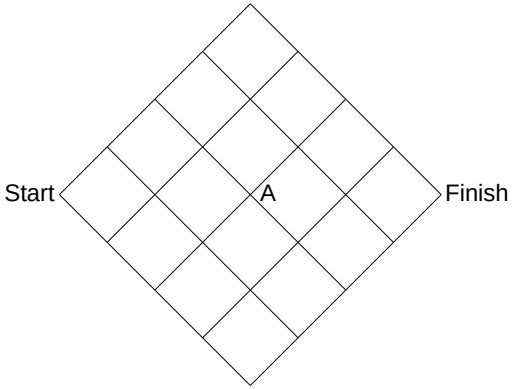
```
-ACGT-

AT--CG
```



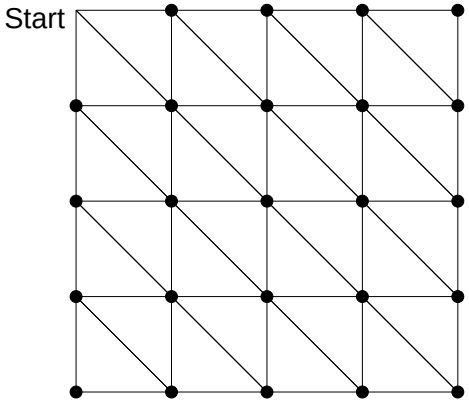
```
ACGT
|
ATCG
```

- Vertical steps add a gap to the horizontal sequence
- Horizontal steps add a gap to the vertical sequence

Do we have to enumerate all paths?



How many paths?



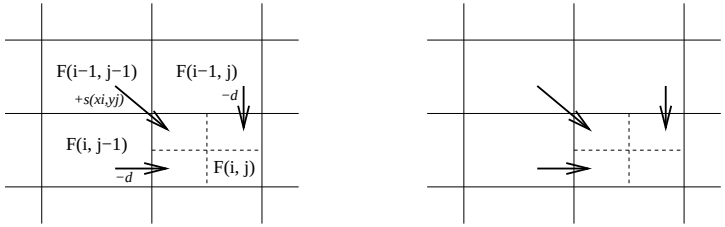
Pairwise global alignment (Needleman-Wunsch algorithm)

Rigorous algorithms use dynamic programming to find an optimal alignment.

- match score
- mismatch score
- gap penalty

$$F(i,\,j) = \max \begin{cases} F(i-1,\,j-1) + s(x_i, y_j) \\ F(i-1,\,j) - d \\ F(i,\,j-1) - d \end{cases}$$

Dynamic programming



Percent identity

Having obtained an alignment, it is common to quantify the similarity between a pair of sequences by stating the percent identity.

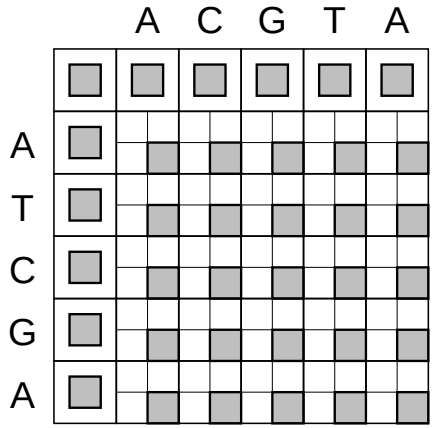
Count the number of alignment positions with matching characters and divide by ... *what*?

- the length of the shortest sequence?
- the length of the alignment?
- the average length of the sequences?
- the number of non-gap positions?
- the number of equivalenced positions excluding overhangs?

Sequences are either homologous (i.e. they share a common evolutionary ancestor) or they are not.

The phrase “percent homology” is meaningless!

Score matrix



Pairwise local alignment (Smith-Waterman algorithm)

Local similarities may be masked by long unrelated regions.

A minor modification to the global alignment algorithm.

- If the score for a subalignment becomes negative, set the score to zero.

$$F(i,\,j) = \max \begin{cases} 0 \\ F(i-1,\,j-1) + s(x_i, y_j) \\ F(i-1,\,j) - d \\ F(i,\,j-1) - d \end{cases}$$

- Trace back from the position in the score matrix with the highest value.
- Stop at cell where score is zero.