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

Bioinformatics in TMS145

Focus on two important kinds of biological data:

- biological sequences
- macromolecular structures

Using data to understand evolutionary relationships

- Lab on sequence alignment; comparing related biological sequence
- Lab on protein structure comparison; examining similarities and differences in a family of related protein stuctures

Computational biology

> "*Computational Biology:* The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."
>
> "Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology."

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

Some questions of interest

Do two biological sequences have significant similarity?

Given a newly discovered biological sequence, is it like anything that has been seen before?

What can be learned from a multiple protein sequence alignment?

How are protein sequence similarity and structural similarity related?

How can comparative modelling be used to construct models of proteins that have a common ancestor?

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

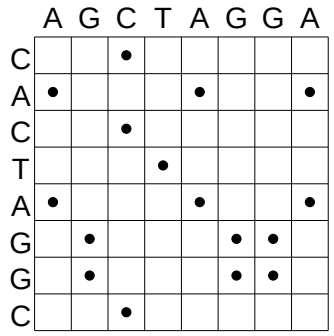
Hamming distance:
- Number of positions with mismatching characters.
- Defined for two strings of equal length.

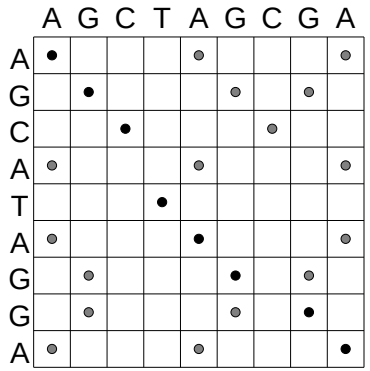
```
agtc
cgta
```

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

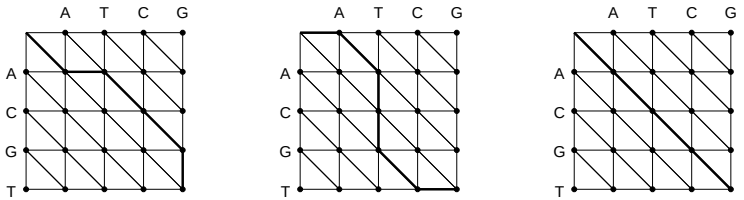
```
agtcc
cgctca
```


Dotplots





Each path represents an alignment




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



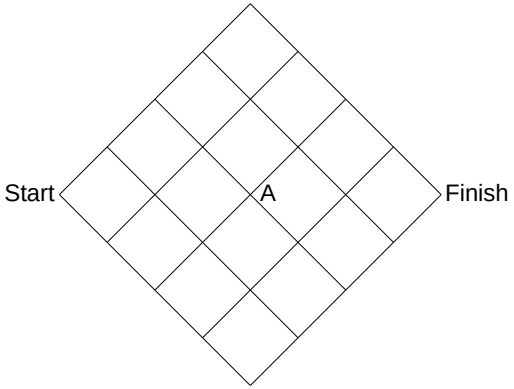
```
-ACGT-

AT--CG
```



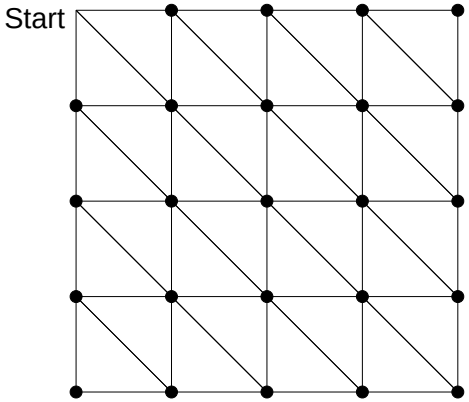
```
ACGT
|
ATCG
```

- 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?



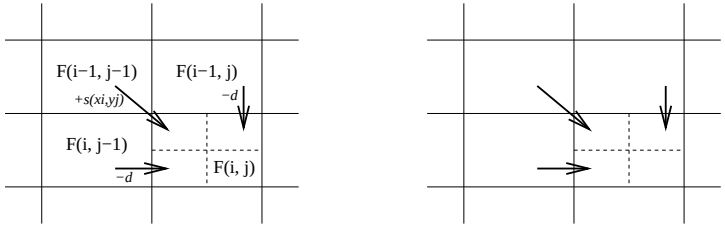
Pairwise global alignment (Needleman-Wunsch algorithm)

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.

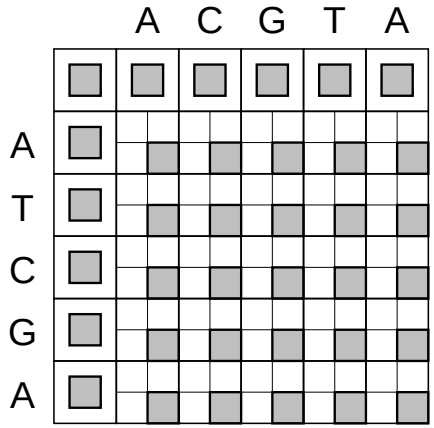
Count the number of alignment positions with matching characters and divide by ... *what*?

- 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.