Sequence alignment

Comparison of macromolecular sequences.

Nucleic acids (DNA, RNA) or proteins.

Assignment of nucleotide-nucleotide or protein-protein 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.

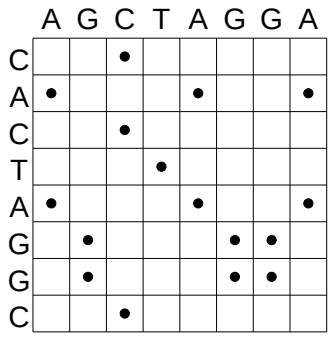
```
agtc
cgta
```

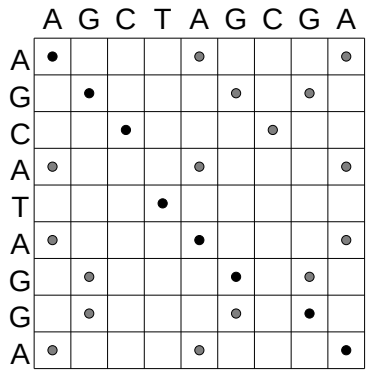
Levenshtein distance:

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

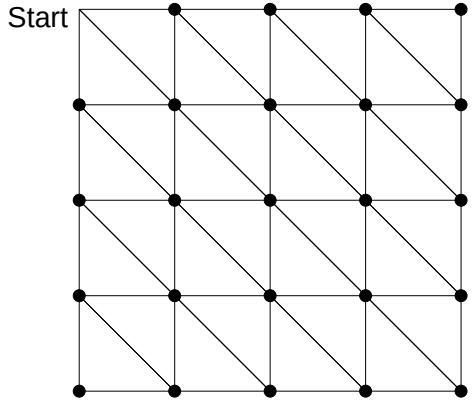
```
agtcc
cgctca
```


Dotplots





How many paths?



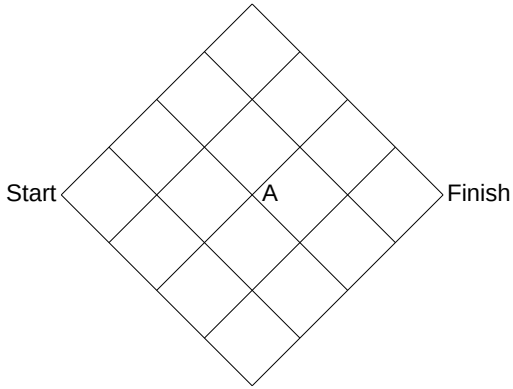
Pairwise global alignment (Needleman-Wunsch algorithm)

Rigorous algorithms use dynamic programming to find an optimal alignment.

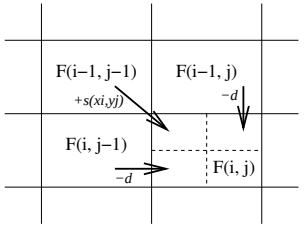
- match score
- mismatch score
- gap penalty

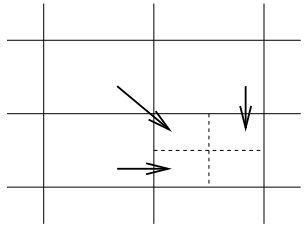
$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Do we have to enumerate all paths?

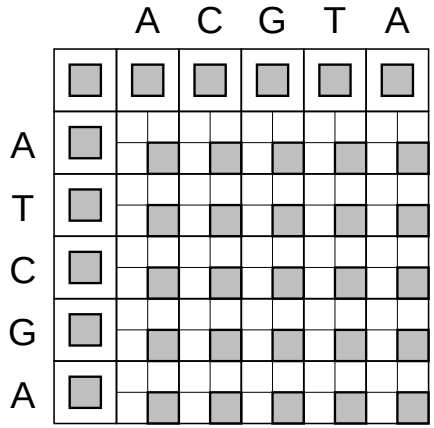


Dynamic programming





Score matrix



Percent identity

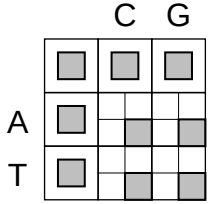
Having obtained an alignment, it is common to quantify the similarity between a pair of sequences by stating the percent identity.

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!

How many ways can “AT” be aligned with “CG”?



Two diagonal moves:

AT
CG

One diagonal move:

AT−	A−T
C−G	CG−
AT−	−AT
−CG	CG−
A−T	−AT
−CG	C−G

No diagonal moves:

AT−−	A−T−	A−−T
−−CG	−C−G	−CG−
−−AT	−A−T	−AT−
CG−−	C−G−	C−−G

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.