# Sequence Alignment 

Mark Voorhies

4/24/2012

## Exercise: Transforming sequences

(1) Write a function to return the antisense strand of a DNA sequence in 3'! 5' orientation.
(2) Write a function to return the complement of a DNA sequence in 5'! 3' orientation.
(3) Write a function to translate a DNA sequence

## Why compare sequences?

## Why compare sequences?

- To find genes with a common ancestor
- To infer conserved molecular mechanism and biological function
- To find short functional motifs
- To find repetitive elements within a sequence
- To predict cross-hybridizing sequences (e.g. in microarray design)
- To predict nucleotide secondary structure


## Whiteboard Image



## Nomenclature

Homologs heritable elements with a common evolutionary origin.

## Nomenclature

Homologs heritable elements with a common evolutionary origin.
Orthologs homologs arising from speciation.
Paralogs homologs arising from duplication and divergence within a single genome.

## Nomenclature

Homologs heritable elements with a common evolutionary origin.
Orthologs homologs arising from speciation.
Paralogs homologs arising from duplication and divergence within a single genome.
Xenologs homologs arising from horizontal transfer.
Onologs homologs arising from whole genome duplication.

## Dotplots


(1) Unbiased view of all ungapped alignments of two sequences

## Dotplots


(1) Unbiased view of all ungapped alignments of two sequences
(2) Noise can be filtered by applying a smoothing window to the diagonals.

## Types of alignments

Global Alignment Each letter of each sequence is aligned to a letter or a gap (e.g., Needleman-Wunsch)
Local Alignment An optimal pair of subsequences is taken from the two sequences and globally aligned (eg., Smith-Waterman)

## Exercise: Scoring an ungapped alignment

$$
\begin{aligned}
& \text { s =f"A":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "T":f"A": 1.0,"T": 1.0,"G": 1.0,"C": } 1.0 \mathrm{~g} \text {, } \\
& \text { "G":f"A": 1.0,"T": 1.0,"G": } 1.0 \text {,"C": } 1.0 \mathrm{~g}, \\
& \text { "C":f"A": 1.0,"T": 1.0,"G": 1.0,"C": } 1.0 \mathrm{gg}
\end{aligned}
$$

## Exercise: Scoring an ungapped alignment

$$
\begin{aligned}
& s=f " A ": f " A ": 1.0, " T ": 1.0, " G ": 1.0, " C^{\prime \prime}: 1.0 \mathrm{~g}, \\
& \text { "T":f"A": 1.0,"T": } 1.0, " G ": 1.0, " C ": 1.0 \mathrm{~g}, \\
& \text { "G":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "C":f"A": 1.0,"T": 1.0,"G": 1.0,"C": } 1.0 \mathrm{gg} \\
& \mathrm{~S}(\mathrm{x}, \mathrm{y})=\sum_{i}^{N} \mathrm{~s}\left(\mathrm{x}_{i}, \mathrm{y}_{i}\right)
\end{aligned}
$$

## Exercise: Scoring an ungapped alignment

$$
\begin{aligned}
& s=f " A ": f " A ": 1.0, " T ": 1.0, " G ": 1.0, " C^{\prime \prime}: 1.0 \mathrm{~g}, \\
& \text { "T":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "G":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "C":f"A": 1.0,"T": 1.0,"G": 1.0,"C": } 1.0 \mathrm{gg} \\
& \mathrm{~S}(\mathrm{x}, \mathrm{y})=\sum_{i}^{N} \mathrm{~s}\left(\mathrm{x}_{i}, \mathrm{y}_{i}\right)
\end{aligned}
$$

(1) Given two equal length sequences and a scoring matrix, return the alignment score for a full length, ungapped alignment.

## Exercise: Scoring an ungapped alignment

$$
\begin{aligned}
& \text { s =f"A":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "T":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "G":f"A": 1.0,"T": 1.0,"G": 1.0,"C": 1.0g, } \\
& \text { "C":f"A": 1.0,"T": 1.0,"G": 1.0,"C": } 1.0 \mathrm{gg} \\
& \mathrm{~S}(\mathrm{x}, \mathrm{y})=\sum_{i}^{N} \mathrm{~s}\left(\mathrm{x}_{i}, \mathrm{y}_{i}\right)
\end{aligned}
$$

(1) Given two equal length sequences and a scoring matrix, return the alignment score for a full length, ungapped alignment.
(2) Given two sequences and a scoring matrix, find the offset that yields the best scoring ungapped alignment.

## Exercise: Scoring a gapped alignment

(1) Given two equal length gapped sequences (where "-" represents a gap) and a scoring matrix, calculate an alignment score with a -1 penalty for each base aligned to a gap.

## Exercise: Scoring a gapped alignment

(1) Given two equal length gapped sequences (where "-" represents a gap) and a scoring matrix, calculate an alignment score with a -1 penalty for each base aligned to a gap.
(2) Write a new scoring function with separate penalties for opening a zero length gap (e.g., $G=-11$ ) and extending an open gap by one base (eg., $\mathrm{E}=-1$ ).

$$
\mathrm{S}_{\text {gapped }}(\mathrm{x}, \mathrm{y})=\mathrm{S}(\mathrm{x}, \mathrm{y})+\sum_{i}^{\text {gaps }}(\mathrm{G}+\mathrm{E} \quad \text { len }(\mathrm{i}))
$$

## Homework

(1) Read chapter 3 of the BLAST book (Sequence Alignment).
(2) Try initializing and filling in a dynamic programming matrix by hand (e.g, try reproducing one of the examples from the BLAST book on paper).

