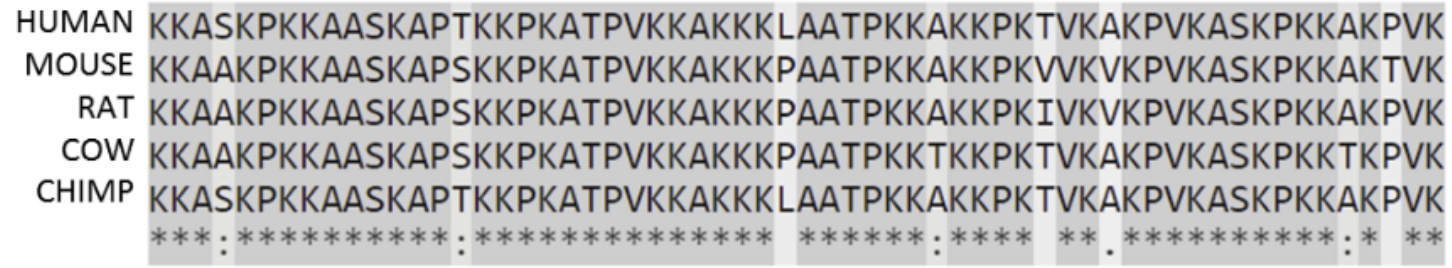
BIO 285/CSCI 285/MATH 285
Bioinformatics
Programming Lecture 7
Pairwise Sequence Alignment

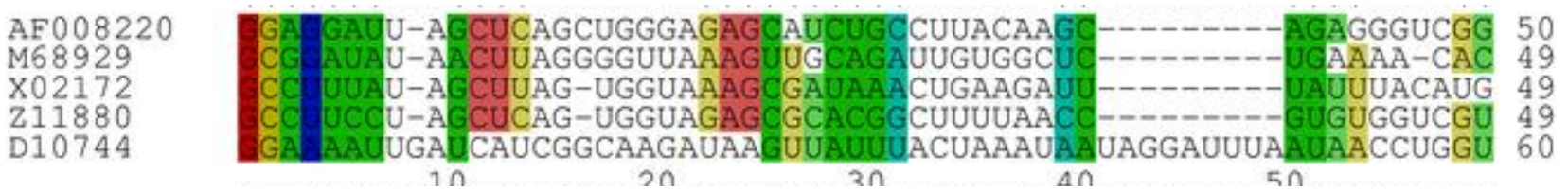
Instructor: Lei Qian
Fisk University

Sequence Alignment

Sequence Alignment

A sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.





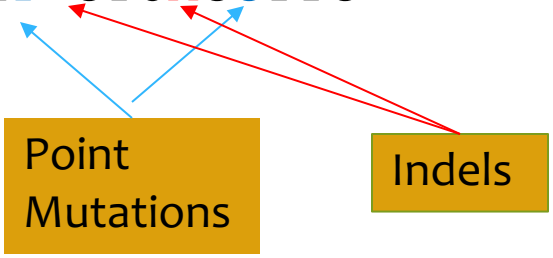
Sequence Alignment

Why do we need sequence alignment?

If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another.

```
AACGTCGCTTG
ATGTCAGGTTG

AACGTC-GCTTG
AT-GTCAGGTTG
```



Alignment

- Measure their similarity
- Infer evolutionary relationships:
  - finding homologues
- More precise tools are needed to analyze the sequences in detail including
  - Dot plots for graphic analysis
  - Local or global alignments for residue/residue analysis
- The alignment procedure comparing two biological sequences (could be DNA, RNA or protein) is called a *pairwise* sequence alignment.
- The alignment procedure comparing three or more biological sequences is called a *multiple* sequence alignment.

PSL vs MSA

Pairwise Alignment

- Can be categorized as global and local alignment
- Comparatively simple algorithms
- a) Find out conserved regions between the two sequences
  b)Similarity searches in a database

Example Tools:
LAIGN, BLAST, EMBOSS Needle, EMBOSS Water.

Multiple Sequence Alignment

- Generally a global alignment.
- Complex sophisticated algorithm
- a) To detect regions of variability or conservation in a family of proteins.
  b) Phylogenetic analysis.
  c)Detection of homology between a newly sequenced gene and an existing gene family prediction of protein structure.
  d) Demonstration of homology in multigene families.

EXAMPLE Tools:
MULSCLE, T-Coffee, MAFFT, CLUSTALW

PSL vs MSA

Local Alignment

Pairwise Sequence Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
              |||| |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

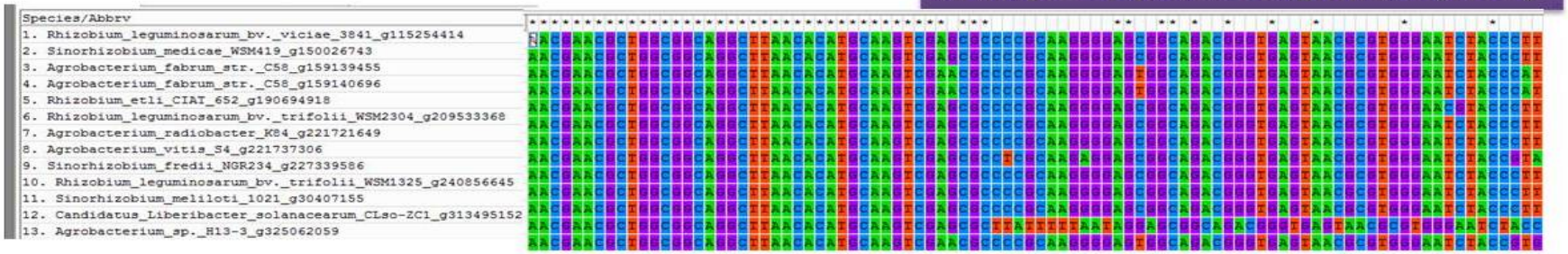
Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

Multiple Sequence Alignment (MSA)



Alignment

Different types of pairwise comparisons	
Method name	***Situation***
Dot plot	**General exploration of your sequence** Discovering repeats Finding long insertions and deletions Extracting portions of sequences to make a multiple alignment
Local alignments	**Comparing sequences with partial homology** Making high-quality alignments Making residue-per-residue analysis
Global alignments	**Comparing two sequences over their entire length** Identifying long insertions and deletions Checking the quality of your data Identifying every mutation in your sequences

Alignment

Dot Plot Method:

- dot plot is a graphic representation of pairwise similarity
- The simplest method for identifying similarities between two sequence
- Ideal for looking for features that may come in different orders
- Reveal complex patterns
- Benefit from the most sophisticated statistical analysis tool -- your brain
- Uses a 2-dimensional table
  - one of the sequences labels the rows
  - the other sequence labels the columns
  - mark a ● in each cell that has matching (row, column) labels

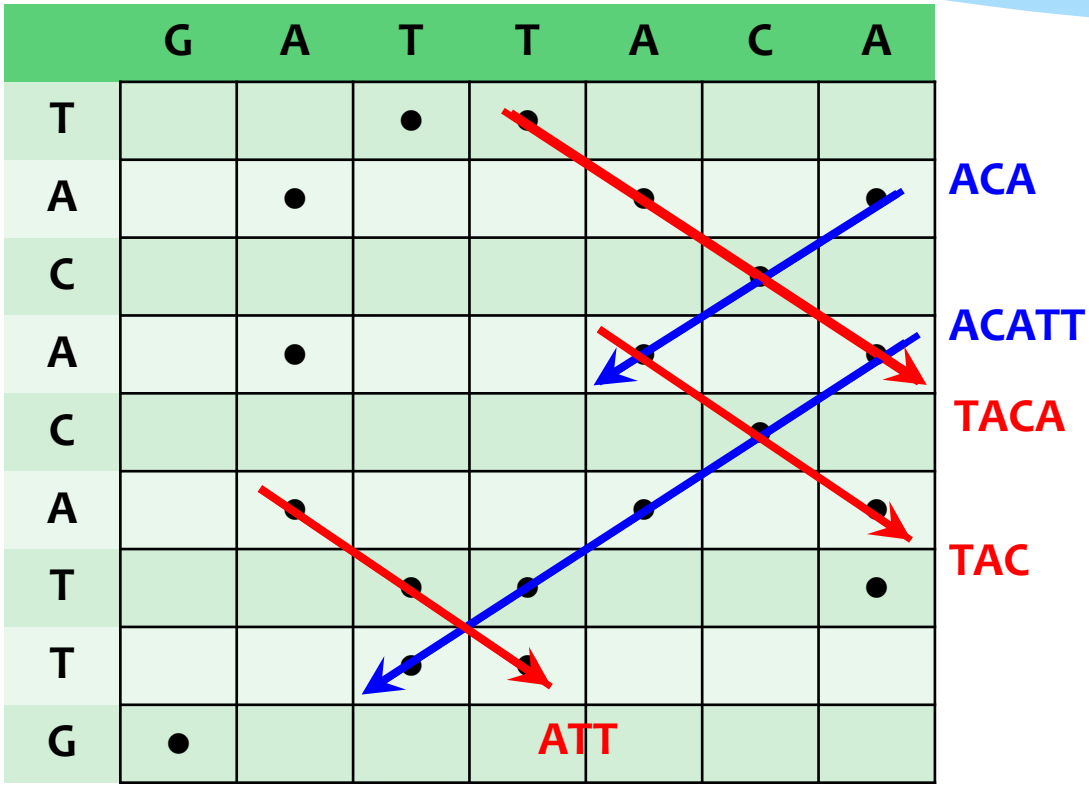
Alignment

Example: Dot plot for “GATTACA” and “TACACATTG”:
Step 1: Mark similarities

	G	A	T	T	A	C	A
T	?	?	●	?	?	?	?
A	?	?	?	?	?	?	?
C	?						
A							
C							
A							
T							
T							
G							

Alignment

Example: Dot plot for “GATTACA” and “TACACATTG”:
Step 2: Find long matching diagonal lines.



Alignment

Dot Plots:

- Diagonal lines indicate regions of similarities between two sequences.
  - SE slope – similarity along the direction of the sequences.
  - SW slope – similarity along one sequence in reverse.
- Susceptible to noise – especially with DNA/RNA since only 4 possible symbols there will be a lot of random hits.

Alignment

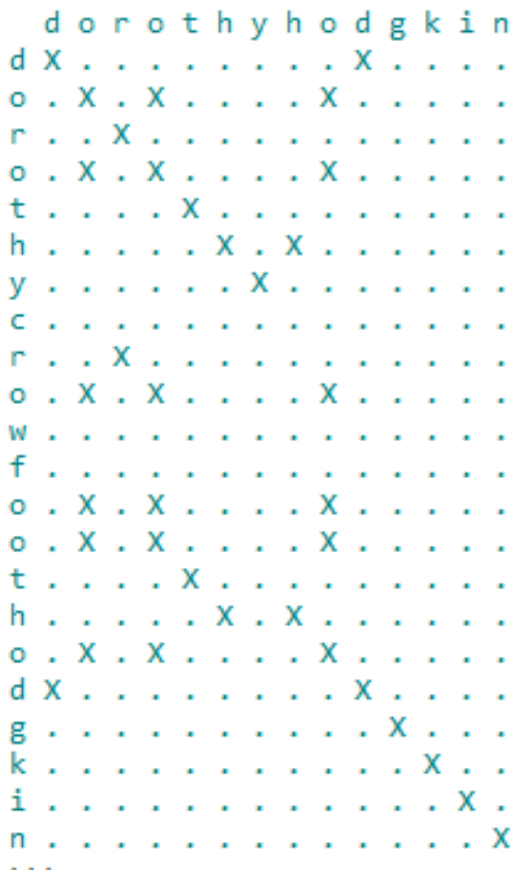
A simple Python dot plot program

```
#A simple Dot Plot Porgram. By L Qian
def simpleDotPlot(s1, s2):
   #Compare String s1 and String s2
   print ' ',
   #print the first row
   for c1 in s1:
      print c1,
   print  #start from a new line
   #print dots
   for c2 in s2: #for each character in s2, print a row
      print c2,
      for c1 in s1:  #for each character in s1, compare to the char in s2 and print X or .
         if c1==c2:
            print 'X',
         else:
            print '.',
      print

simpleDotPlot("dorothyhodgkin", "dorothycrowfoothodgkin")
```

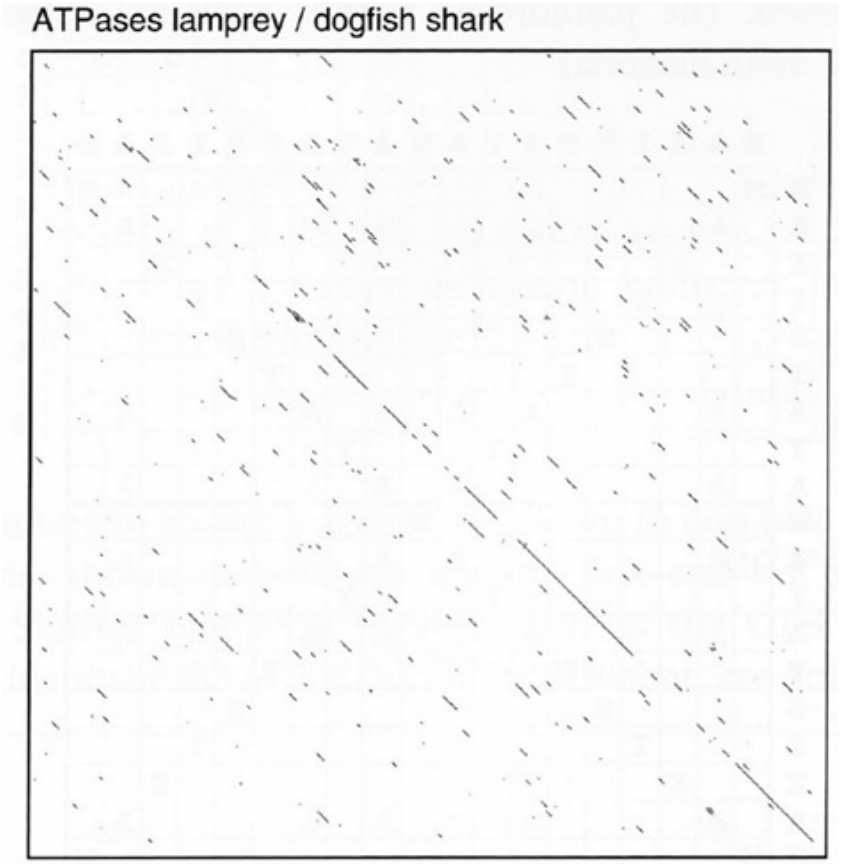
Alignment

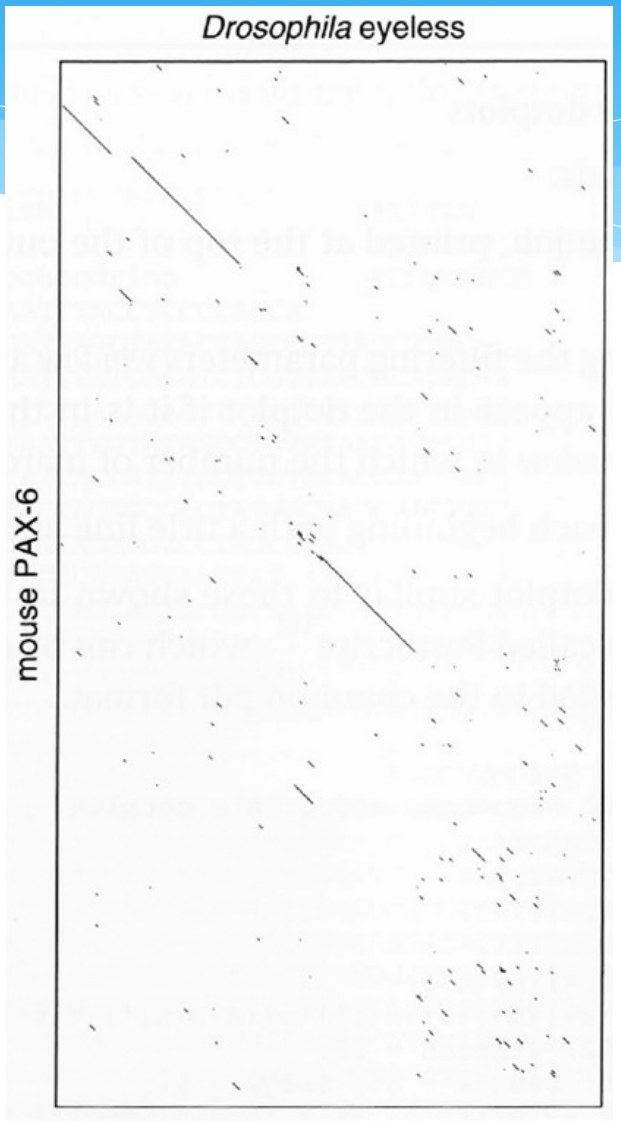
A simple Python dot plot program



Alignment

Examples of dot plots:

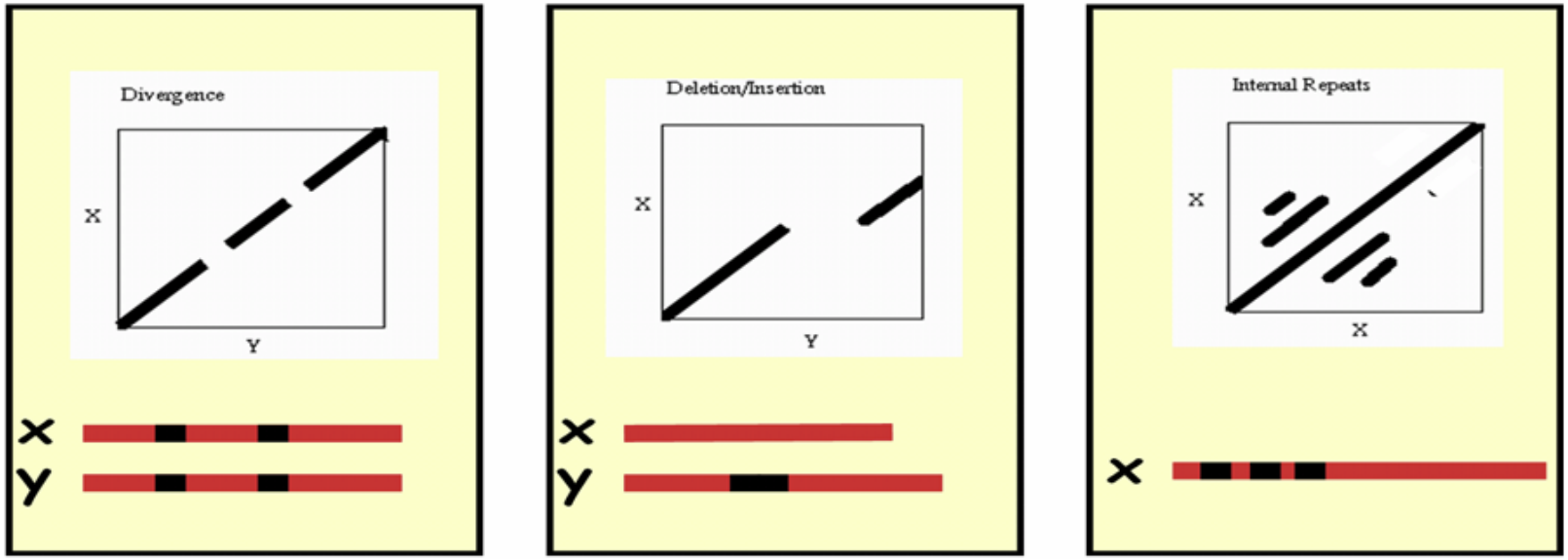




Alignment

Some typical dot plot comparisons

- Divergent sequences where only a segment is homologous
- Long insertions and deletions
- Tandem repeats: The square shape of the pattern is characteristic of these repeats



Alignment

Dot Plots with Sliding Window:

- Noise can be eased using a sliding window
  - *consider fragments of length W in the two sequences*
  - *place ● in each cell that is the "origin" of the sliding window*

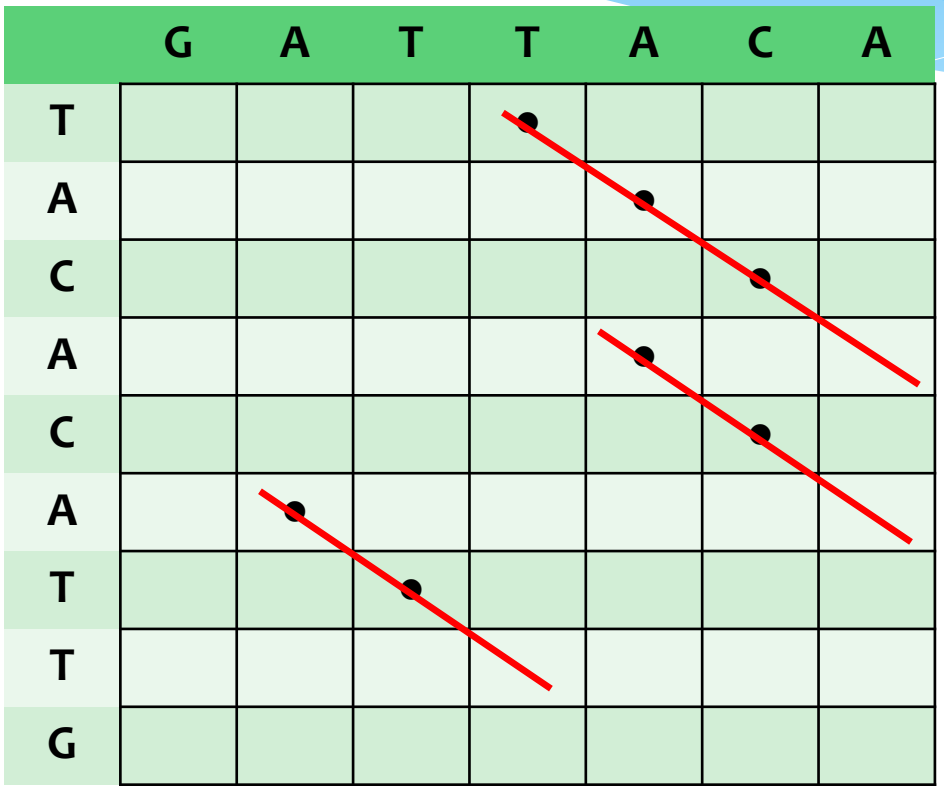
Alignment

Dot Plots with Sliding Window (W=2)

	G	A	T	T	A	C	A
T	?	?	?	?	?	?	
A	?	?	?	?	?	?	
C							
A							
C							
A							
T							
T							
G							

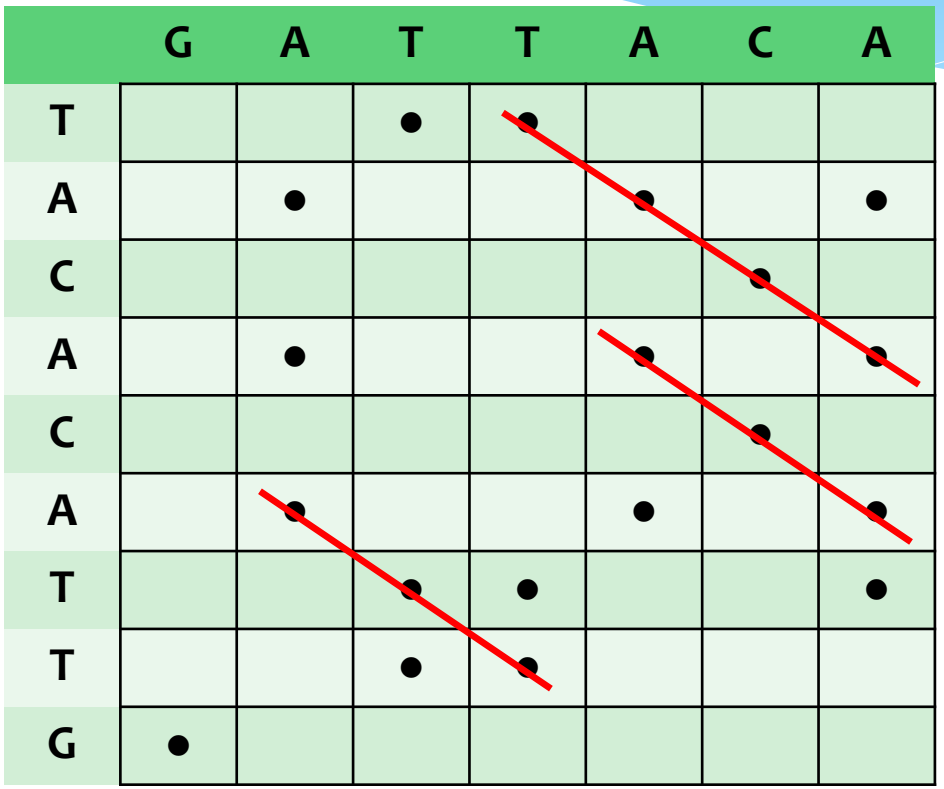
Alignment

Dot Plots with Sliding Window (W=2)



Alignment

Dot Plots with Sliding Window (W=1)



Alignment

Dot Plots with Sliding Window size W:

Compare with next slide with W = 1

- *noise has disappeared*
- *one fewer dots per matching region*
- *in general if N matches per region, #dots = N – (W-1)*

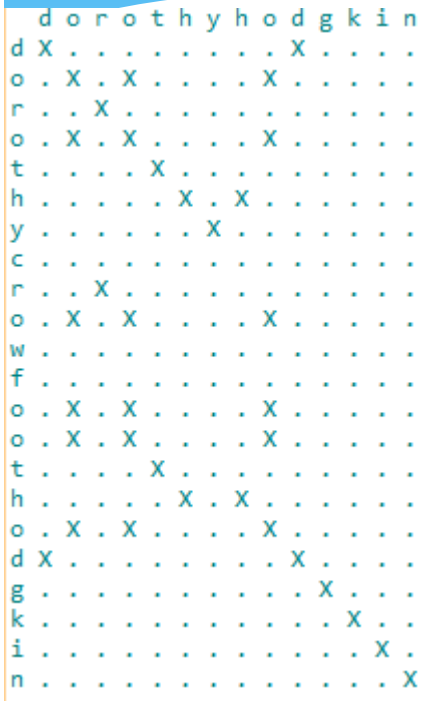
Alignment

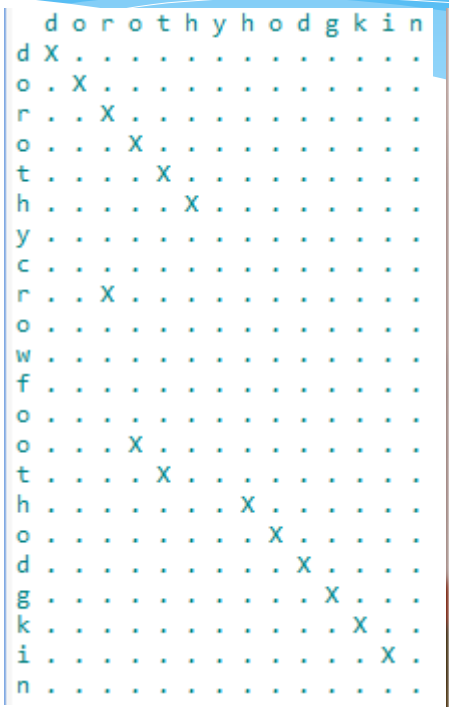
A simple Python dot plot program with sliding window:

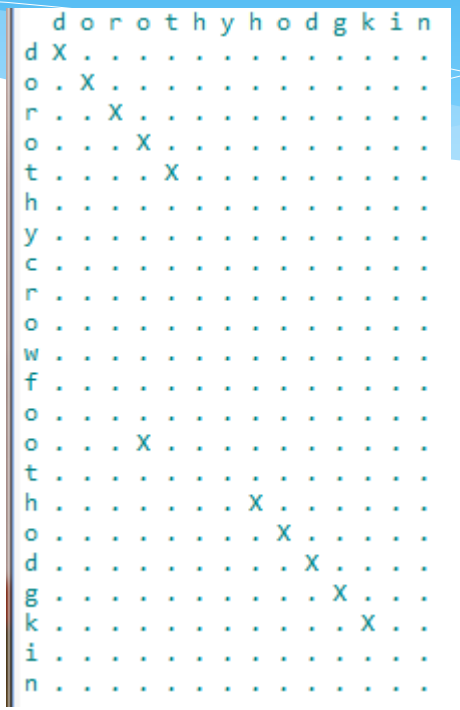
```
def simpleDotPlotW(s1, s2, wsize):
……..        # code to print the first row
row = 0
   vWindow = []  #vertical windows
   for c2 in s2:
      vWindow.append(c2)
      if row>=wsize-1:
         if row>=wsize:
            vWindow.pop(0) #remove the first item.
         print vWindow[0],
         hWindow = []
         col = 0
         for c1 in s1:
            hWindow.append(c1)
            if col>=wsize - 1:
               if col>=wsize:
                  hWindow.pop(0)
               if hWindow==vWindow:
                  print 'X',
               else:
                  print ".",
            col += 1
         for c in range(wsize-1):
            print '.',
         print
      row +=1
```

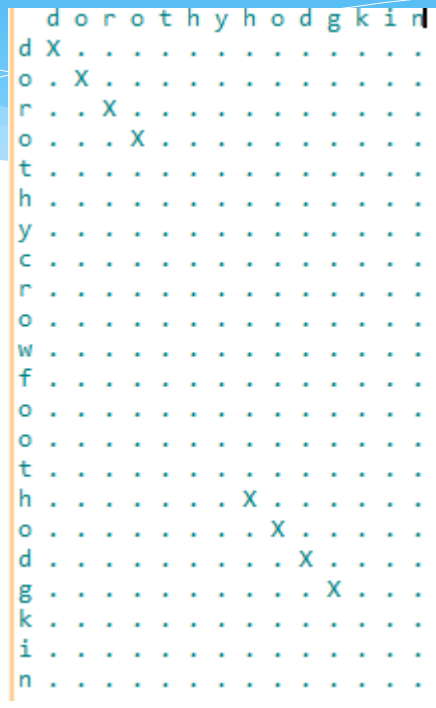
Alignment

Output for windows = 1, 2, 3 and 4.









Alignment

Self Alignment:

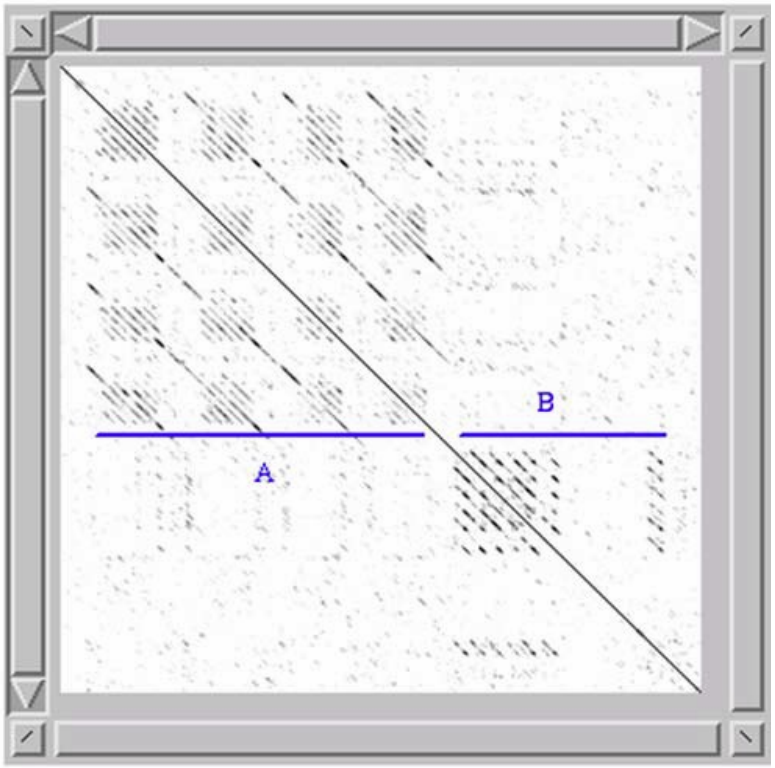
Comparing a sequence with itself
We can discover

- Repeated domains
- Motifs repeated many times (low complexity)
- Mirror regions (palindromes) in nucleic acids

Alignment

Self Alignment:

- The square shape is typical of tandem repeats.
- The repeats are not perfect because the sequences have diverged after their duplication.



Alignment

Dot Plot tools online:

Dotplot Program by Sonnhammer:
http://sonnhammer.sbc.su.se/Dotter.html

Web based Dot Plot:
http://myhits.isb-sib.ch/cgi-bin/dotlet