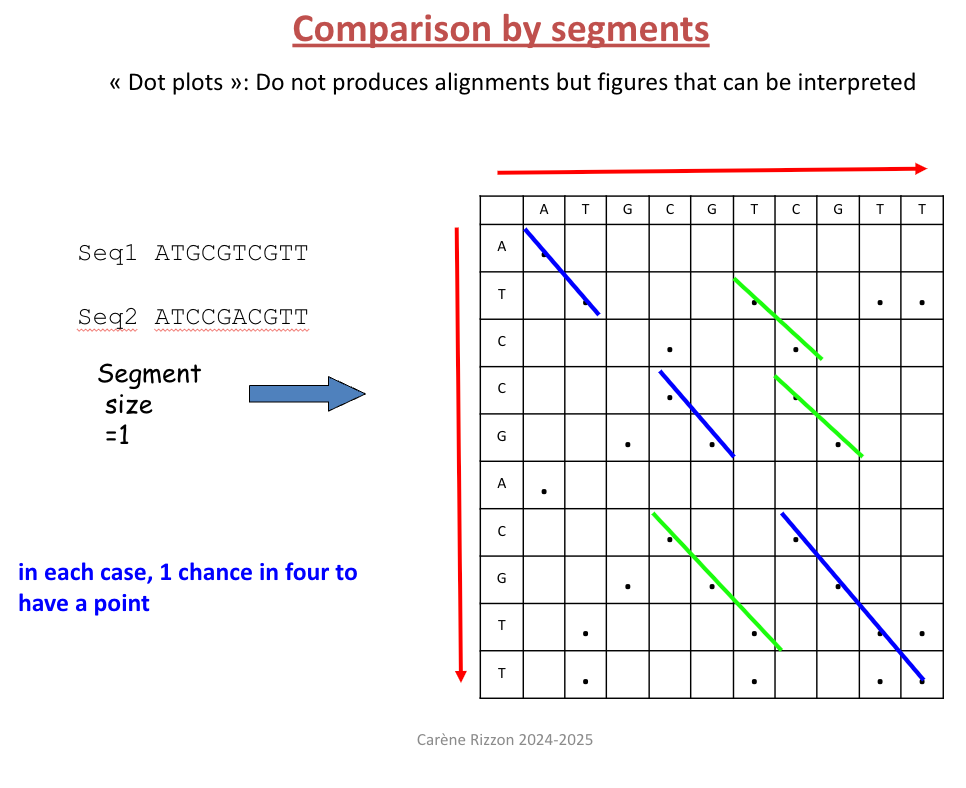
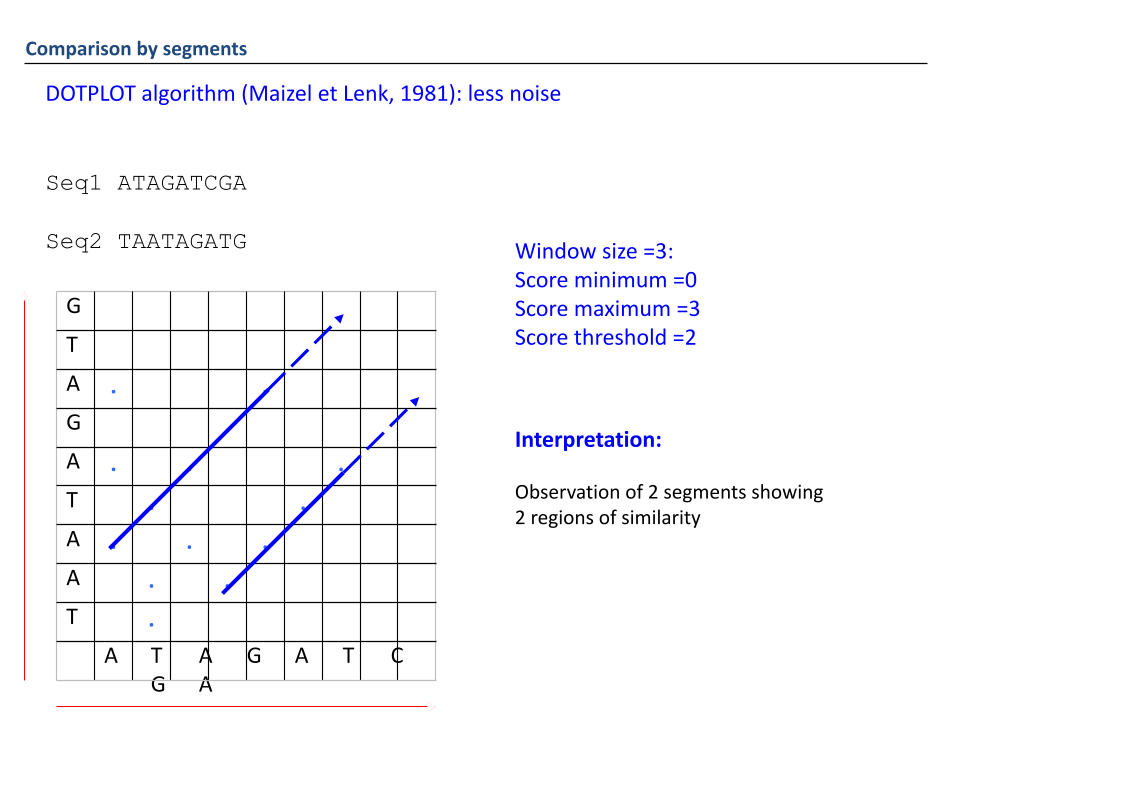
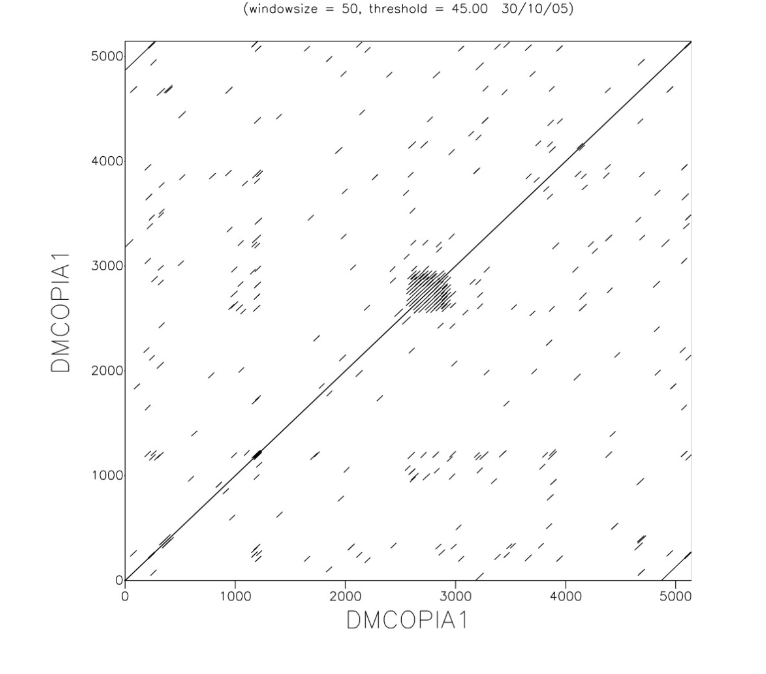
* **Parsimony**: The simplest algorithm to identify phylogenetic relationships between different genes. BLAST uses this **paradigm** of inference.
* The hypothesis for inference when comparing genetic sequences is the theory of evolution.
* What can cause differences in genetic material?
  + **Mutation**: SE.
  + **Selection**: Those who survive with advantageous mutations etc.
  + **Genetic drift**: Population induced shift in genetic profiles, often induced by a reduction in genetic variation, can be caused by migration; no **adaptive** changes, just proportional shift in **random mutations**, **deleterious or additive mutations**.
  + **Migration**: Genetic adaptation to new environments.
* In aligning sequences, one can benefit by knowing, or even hypothesising, that the sequences could be homologous as/if there is/isn’t proof of their common ancestry. In order to ensure that the sequences are actually homologous, one must identify conserved regions at the exact loci of the homologous genes.
* The **base mismatches** in an alignment are the **observed differences**.
  + Using these observe differences and their **frequency**, **location**, among others, **algorithms** are used to infer phylogeny, infer the un-observed differences (possibly over time; **evolutionary**), which led to the gene being how it is at the moment of sequencing.
* Remember the term models, referring to the fact that the sequence printed in a file could contain notations other than “ATCG”, owing to the model which the sequence is trying to fit into.
* BLAST result targets:
  + **Homology**: SE
  + **Orthology**: To identify the phylogeny of species, particularly, **speciation events**.
  + **Paralogy**: SE
  + **Duplication**: SE
* **PPI Data: Protein to Protein Interaction**.
* **Gene Ontology**: Database for gene annotations.





**Window size**: the number of nucleotides you will try and align in the dot plot.

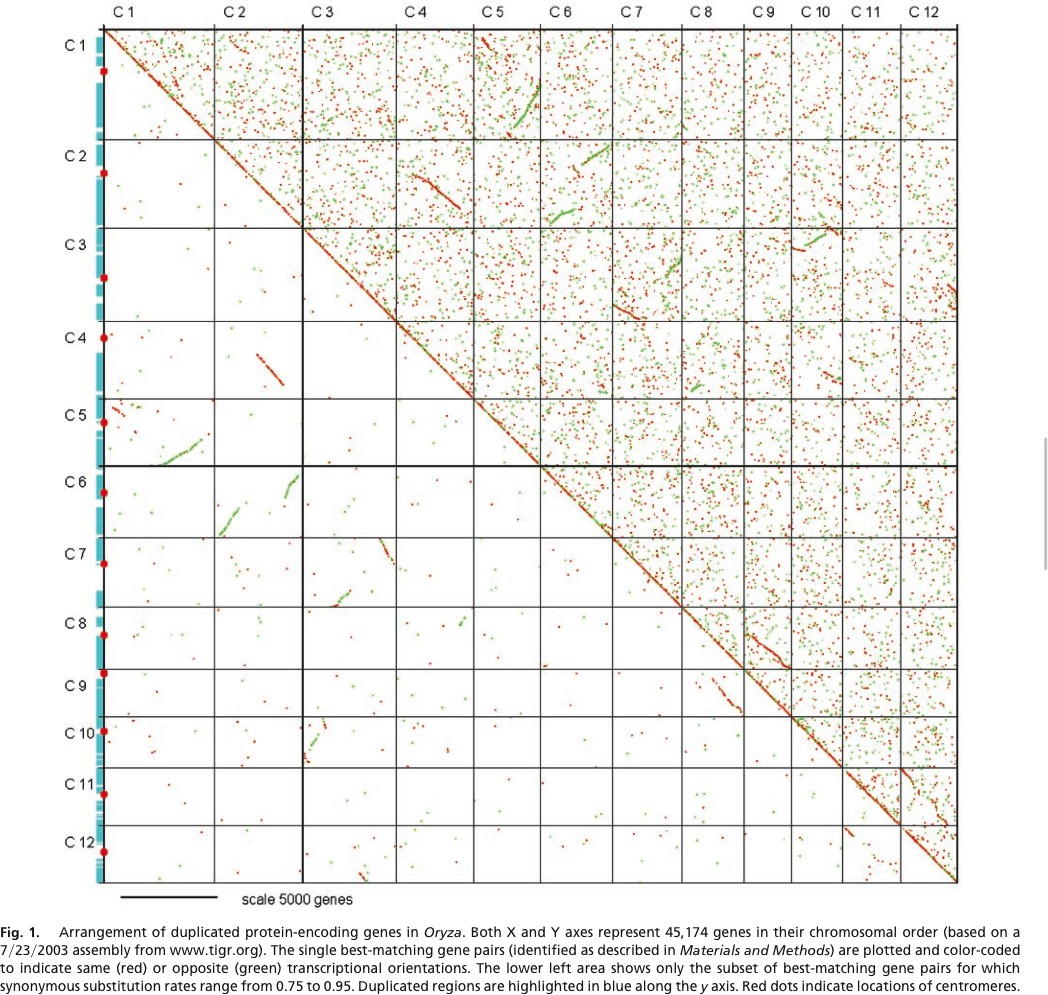
**Threshold**: the number of boxes in the dot plot which have a match, **the minimum number of matches** in percentage.



Here you are comparing a sequence against itself. The smaller lines are noise. With a specific threshold, instead of having a **black box**, you have a noise.

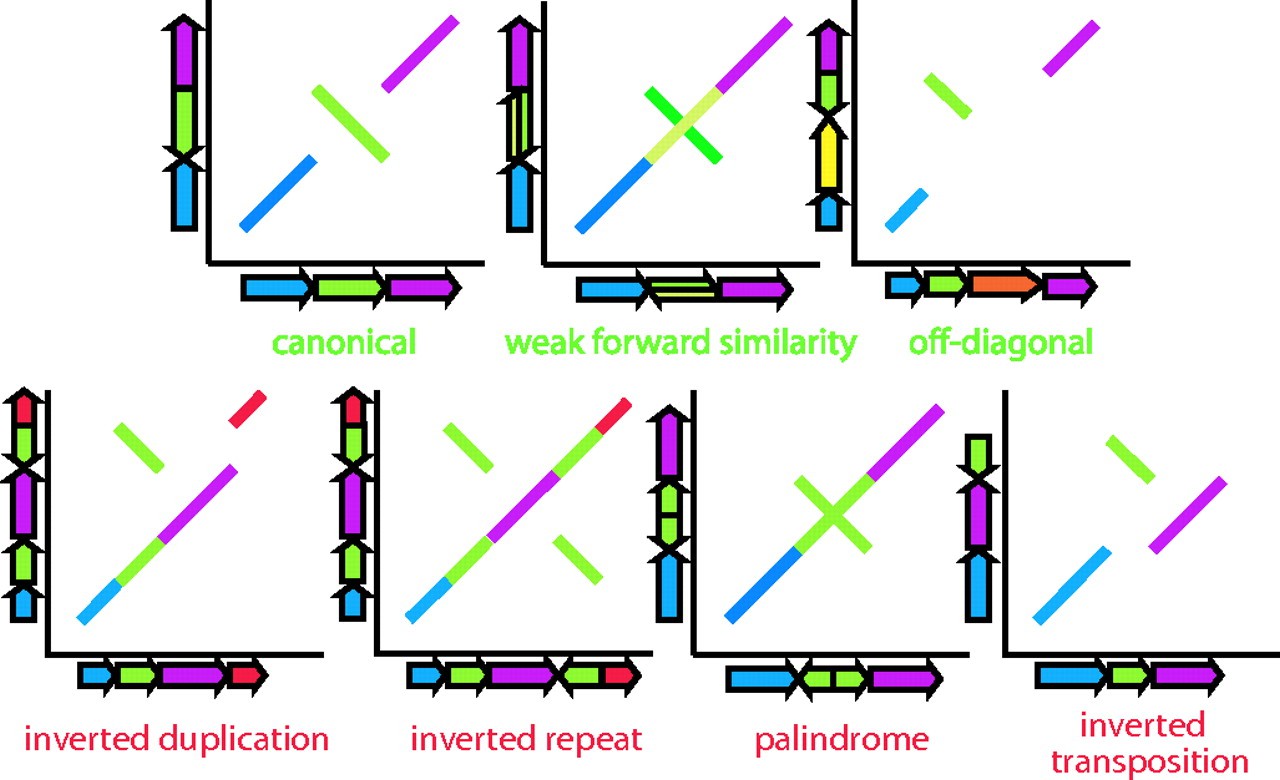
The only observable trait of this graph is the cluster in the middle, **long terminal repeats (LTR), or microsatellites**.

Now you can also see that there is a **line at the beginning and end of the sequence**.



In this dot plot of a genome, you can see **small lines**, which are **remnants of genome duplication**, which are present in **chromosomes** and some in fact, are pointing to the left, instead of the right, **so inversions**.

Or, **only some of the chromosomes** got duplicated.

DOT PLOT GUIDE

* **Transition** mutations are when **A has been changed to T** for example.
* **Transversion mutations** are when **A has been changed to G** for example. **When purines have been converted to pyrimidines.**
* Remember that transitions should receive a higher similarity score than transversions.
* You can also either have **score maximums** or **score minimums**, so the threshold is set by either having a small score or a maximum score.