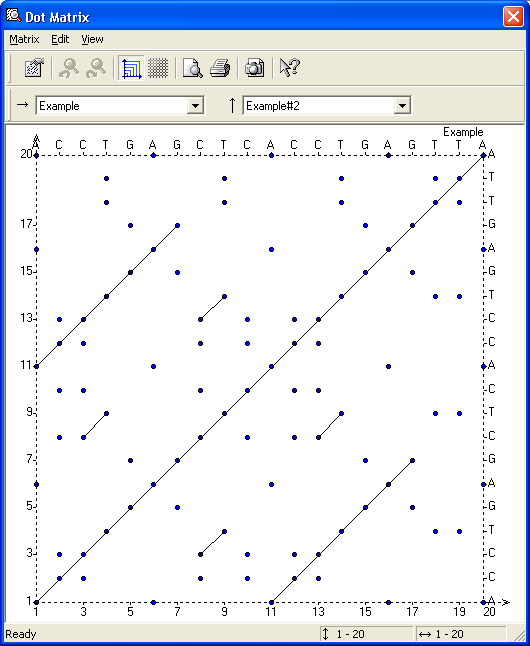
**Course: Advanced Bioinformatics**

**Module title: Dot Plots**

**Module no. : 29**

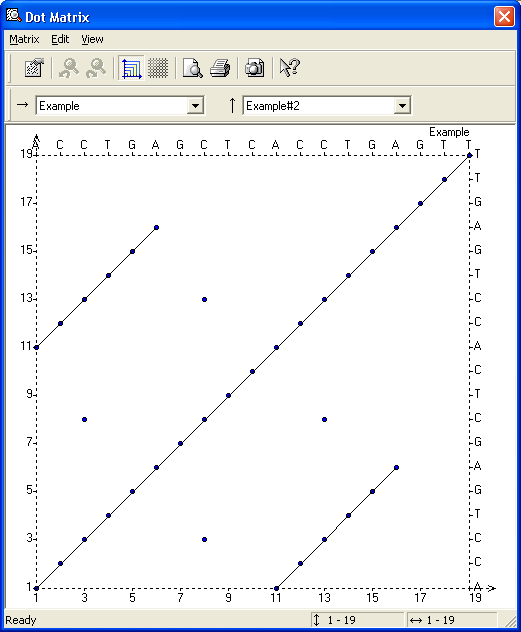
One of the more basic, yet important techniques for determining the alignment between two sequences is by using a visual alignment known as ***dot plots***. Dot plots of sequence similarity are created using a matrix where the rows in the matrix correspond to the characters in the first sequence and the columns in the matrix correspond to the characters in the second sequence. The dot plot is created as follows: loop through each row. For the current row, take the character in that row and compare it to the character in each column. If they are equal, place a dot in the matrix. Continue until all nodes in the matrix have been considered.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **A** | **C** | **C** | **T** | **G** | **A** | **G** | **C** | **T** | **C** | **A** | **C** | **C** | **T** | **G** | **A** | **G** | **T** | **T** | **A** |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **C** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **G** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **T** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **A** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |



Results for aligning ACCTGAGCTCACCTGAGTTA to itself using the Dot Matrix option of the AlignX feature of Informax’s Vector NTI program.

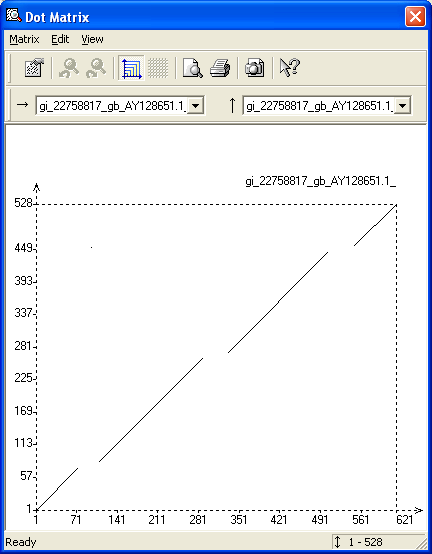
When a dotplot is created to compare nucleic acids, there will be a lot of noise, since one out of every four positions will match at random, if there are an equal number of A, C, G, and T in the sequence. Therefore, dot plots can be filtered for stringency requiring that a certain percentage of nucleotides match in a given window size. With the example above, if we filter the sequences to only show matches of two or more consecutive nucleotides, the dot plot now looks as the following:



**Information within Dot Plots**

Dot plots are useful as a first-level filter for determining an alignment between two sequences. Regions of similarity will show up as diagonals within the dot plot matrix.

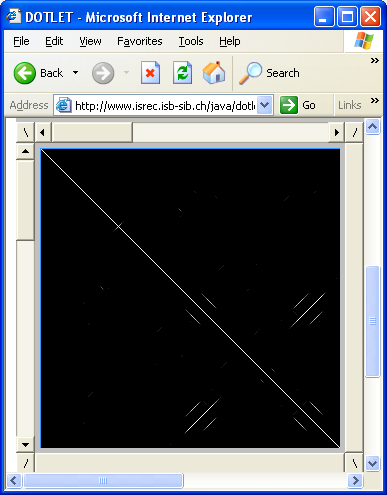
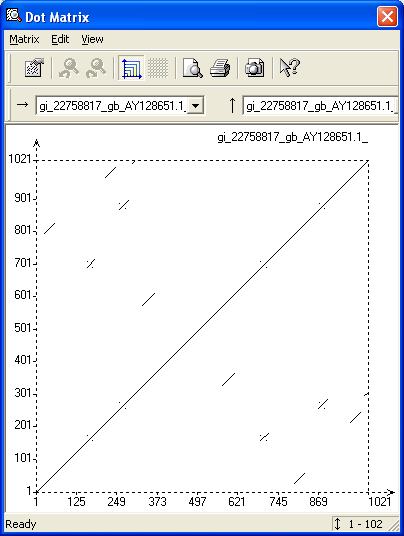
Regions containing insertions/deletions can be readily determined. One potential application is to determine the number of coding regions (exons) contained within a processed mRNA.



Example of a Dot Plot showing insertion/deletion events.

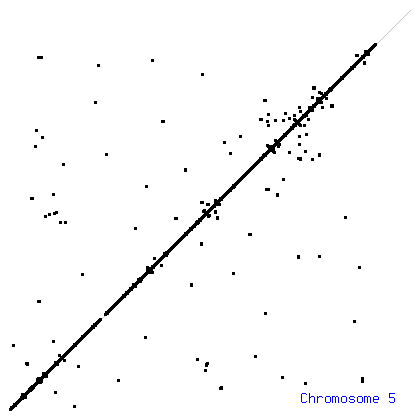
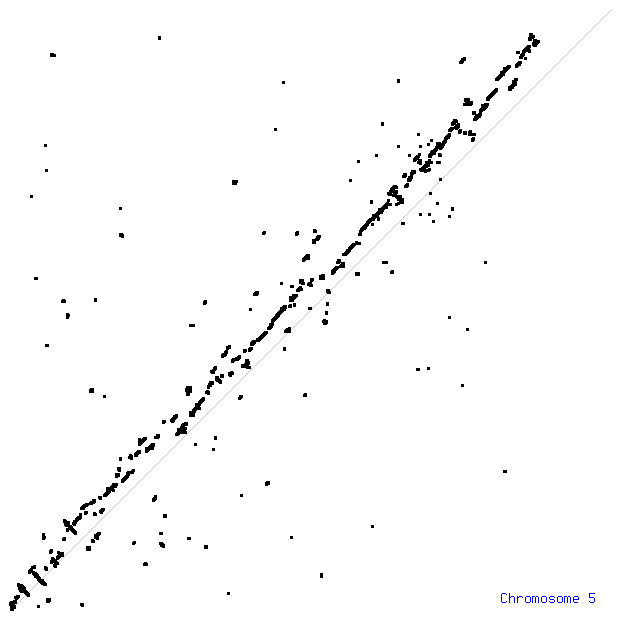
Regions of genomic DNA can contain repetitive regions. For instance, approximately 50 percent of the human genome is composed of repetitive elements, which can be on the order of a few hundred bases (SINEs – Alu elements) or a few thousand (LINES). In addition, regions of low complexity are present as well. Repetitive elements and methods to filter them out will be discussed during a later class period. In addition to repetitive elements, regions of a genome can be duplicated. The duplicated region can be found either as a direct repeat, meaning that it occurs in the same direction, or as an inverted repeat, meaning that the sequence of the duplicated region is found in the reverse complement direction. Dot plots can readily show regions of direct and inverted repeats.

Dot plots show all possible matches of residues between two sequences given a certain threshold level. Thus, the researcher can decide which alignments are the most significant.

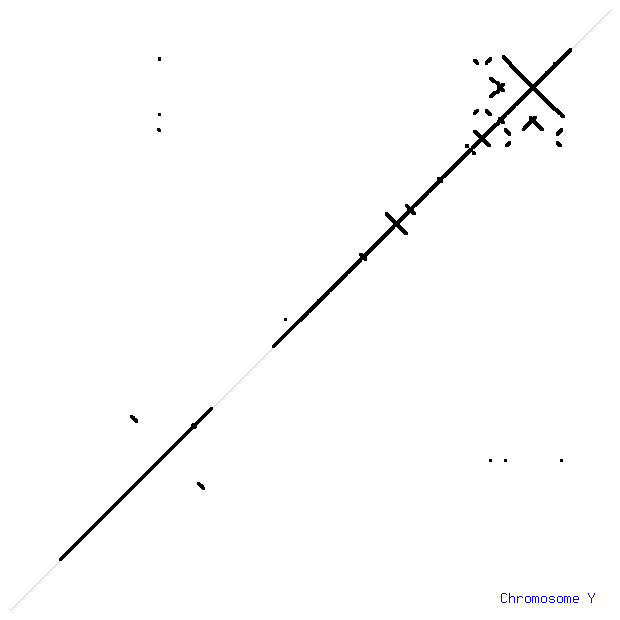


Example dot plots showing the presence of direct and inverted repeats.

Dot plots can also be used in order to compare two different assemblies of the same sequence. Below are three dot plots of various chromosomes. The first shows two separate assemblies of human chromosome 5 compared against each other. The second shows one assembly of chromosome 5 compared against itself, indicating the presence of repetitive regions. The final dot plot shows chromosome Y compared against itself, indicating the presence of inverted repeats.



Comparison of two assemblies of chromosome 5. The figure to the left indicates the alignment of two separate assemblies, while the figure to the right indicates the alignment of a single assembly against itself.



Self plot of chromosome Y. Indicated are several regions of both direct and inverted repeats.