THE DOT MATRIX OR DIAGRAM METHOD FOR COMPARING SEQENCES

Author : Deepak Khatri

History

This method is described by A.J. Gibbs and G.A. McIntyre in 1970 for comparing two amino acid and nucleotide sequences in which a graph was drawn with one sequence writ- ten across the page and the other down the left-hand side. Whenever the same letter appeared in both sequences, a dot was placed at the intersection of the corresponding sequence positions on the graph The resulting graph was then scanned for a series of dots that formed a diagonal, which revealed similarity, or a string of the same characters, between the sequences. Long sequences can also be compared in this manner on a single page by using smaller dots.

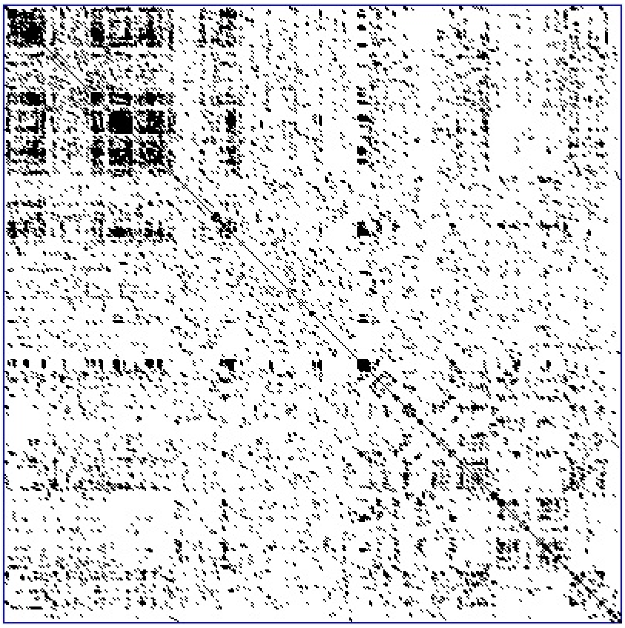


Fig 1. A DNA dot plot (https://en.wikipedia.org/wiki/Dot_plot_(bioinformatics)) of a human zinc finger transcription factor (GenBank ID NM_002383), showing regional self-similarity. The main diagonal represents the sequence's alignment with itself; lines off the main diagonal represent similar or repetitive patterns within the sequence. This is a typical example of a recurrence plot.

Why this method?

The dot matrix method quite readily reveals the presence of insertions or deletions between sequences because they shift the diagonal horizontally or vertically by the amount of change. Comparing a single sequence to itself can reveal the presence of a repeat of the same sequence in the same (direct repeat) or reverse (inverted repeat or palindrome) orientation. This method of self-comparison can reveal several features, such as similarity between chromosomes, tandem genes, repeated domains in a protein sequence, regions of low sequence complexity where the same characters are often repeated, or self-comple-mentary sequences in RNA that can potentially base-pair to give a double-stranded struc-ture. Because diagonals may not always be apparent on the graph due to weak similarity, Gibbs and McIntyre counted all possible diagonals and these counts were compared to those of random sequences to identify the most significant alignments.

Note : *Maizel and Lenk (1981) later developed various filtering and color display schemes that greatly increased the usefulness of the dot matrix method.*

Algorithm :

Step 1 : Calculate the matrix space

```
// for every i within length of sequence 1
// loop j form 0 to legth of sequence 2 and
// check the equality of both seqence to get
// mXn matrix where m,n = length of seq1, seq2

for(i=0; i<seq1.length(); i++)
    for(j=0; j<seq1.length; j++)
        data[i][j] = (seq1[i] != seq2[j])
```

Step 2: Plot the data

```
imshow(data)
```

In [5]:

```
# importing necessary modules
from matplotlib import pyplot as plt
```

In [6]:

```
# sequence data
seq1 = "agctaggacggta"
seq2 = "gactaggcagcag"
```

In [7]:

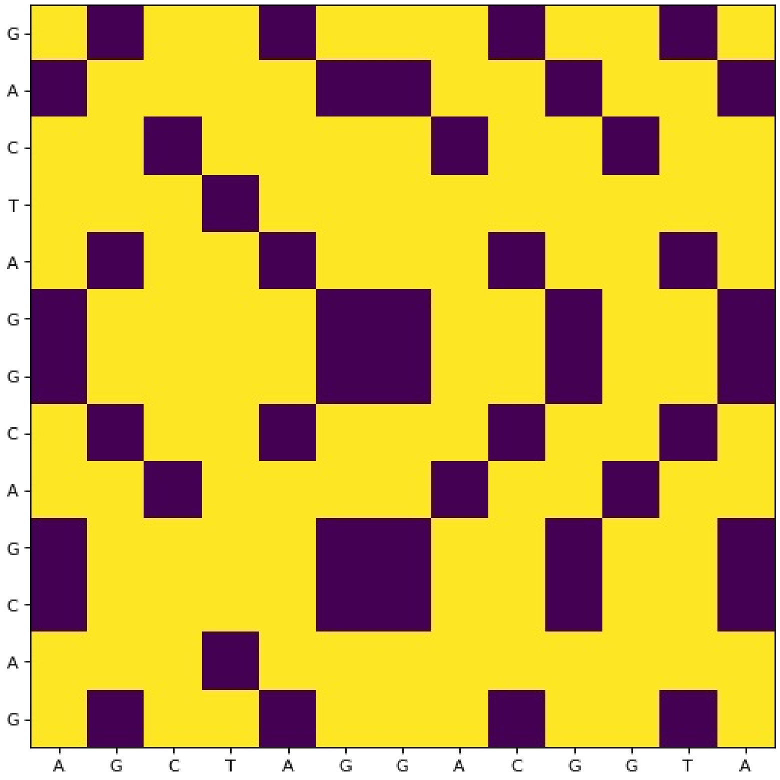
```
# calculating the matrix
data = [[int(seq1[i:i + 1] != seq2[j:j + 1]) for j in range(len(seq1))] for i in range(len(seq2))]
data
```

Out[7]:

```
[[1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1],
 [0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0],
 [1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1],
 [1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 [1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1],
 [0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0],
 [0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0],
 [1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1],
 [1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1],
 [0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0],
 [0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0],
 [1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1],
 [1, 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1]]
```

In [8]:

```
# plotting the data
plt.figure(figsize=(8,8), dpi=80)
plt.xticks([xt for xt in range(len(seq1))],seq1.upper())
plt.yticks([yt for yt in range(len(seq2))],seq2.upper())
plt.imshow(data)
plt.show()
```


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