Exercise 2
Pattern discovery in biological data

1. Pattern discovery in sequences
    a. Describe the general workflow of Benson's k-nucleotide tandem repeat discovery algorithm (refer to reading below if needed)

    b. Is this an efficient algorithm? Explain why or why not

2. Pattern discovery in high-throughput screening data: Hierarchical clustering
    a. Go through the file format help site at http://smd.stanford.edu/help/formats.shtml.
    b. Name and describe the purpose of the four experimental file formats used in hierarchical clustering (exclude xml)

    c. Download *TreeView* from http://sourceforge.net/projects/jtreeview/files/jtreeview/1.1.6r2/ and Install it

    d. Download the example *pcl* data file from http://people.sc.fsu.edu/~jburkardt/datasets/pcl/i29111.pcl. This data represents ~1000 genes and their expression in various conditions. Open it using TreeView.
    Note: Ensure the file was downloaded with an '*.pcl*' extension and that no extra file format was added

e. Can you make sense of the data? i.e. can you see which genes tend to show similar expression in certain conditions?

f. Now, download the clustering algorithm *WCluster* from http://function.princeton.edu/WCluster/ and understand the different parameters it accepts

g. Place the file *i29111.pcl* in the same directory as *WCluster* and run *WCluster* as follows:

```
java -Xmx2000m -jar WCluster.jar i29111.pcl no_weight E
i29111.cdt i29111.gtr i29111.atr
```

Note: *WCluster* uses only average-linking for the linkage criteria. This parameter cannot be changed so we will leave it as is for now

Using *TreeView*, Open the *i29111.cdt* file generated

h. Explain the main differences you see, i.e. what the different clusters represent

i. Re-run *WCluster* again using the command in **g**. Open the generated *cdt* file once again in TreeView. Is the clustering the same as the one generated in **g**? Explain why or why not

j. Experiment running *WCluster* with different distance metrics (**P** and **PA**). Explain why one would need to use different distant metrics when performing such clustering's

By the end of this exercise, you should:

- Have an idea of how tandem repeats or patterns are discovered in biological sequences
- Apprehend technical details of Hierarchical clustering i.e. what it does, the different parameters it uses, its file formats etc.

If you face any difficulties or have any questions, don't hesitate to email me: omar.wagih@utoronto.ca

Readings:

Benson G, Waterman MS. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Res.* 1994;**22**:4828–4836

The End

Exercise 2 (with solutions)
Pattern discovery in bioinformatics

1. Pattern discovery in sequences
    a. Describe the general workflow of Benson's k-nucleotide tandem repeat discovery algorithm (refer to reading below if needed)

    Scan the entire sequence looking for suspicious patterns: Every time a suspicious pattern is detected do the following:
    - ***Compute a similarity score for the pattern versus the sequence in the region where the pattern was found.***
    - ***If the similarity score exceeds a threshold:***
        - ***Compute an alignment of the pattern and the sequence.***
        - ***Determine a consensus pattern from the alignment and re-compute an alignment with the consensus pattern***
        - ***Report sequence identification information and the alignment.***

    b. Is this an efficient algorithm? Explain why or why not

    Yes, to some degree, it is. The algorithm searches for suspicious patterns by scanning for all k-mers. This involves scanning the whole sequence once (O(n), where n is the length of the sequence). Given that a suspicious pattern is found (not very often), a similarity score is computed and if it crosses a specific threshold it is aligned. Thus, not every part of the sequence is scored and not every suspicious sequence is aligned. However, it must be kept in mind that this algorithm is a heuristic and its specificity is not to be compared to other newer and improved algorithms.

2. Pattern discovery in high-throughput screening data: Hierarchical clustering
    a. Go through the file format help site at http://smd.stanford.edu/help/formats.shtml.
    b. Name and describe the purpose of the four experimental file formats used in hierarchical clustering (exclude xml)

    - ***pcl (pre-clustering file): Specifies a format for data that has not yet been clustered.***
    - ***cdt (clustered data table): contains the original data, as recorded in the pcl file but reordered to reflect the clustering***
    - ***gtr (gene tree file): contains the order in which genes (rows) were joined during clustering***
    - ***atr (array tree file): contains the order in which the array genes (columns) were joined during clustering***

c. Download *TreeView* from http://sourceforge.net/projects/jtreeview/files/jtreeview/1.1.6r2/ and Install it
d. Download the example *pcl* data file from http://people.sc.fsu.edu/~jburkardt/datasets/pcl/i29111.pcl. This data represents ~1000 genes and their expression in various conditions. Open it using TreeView.
Note: Ensure the file was downloaded with an '*.pcl*' extension and that no extra file format was added

e. Can you make sense of the data? i.e. can you see which genes tend to show similar expression in certain conditions?

Answer: No. It's not very informative

f. Now, download the clustering algorithm *WCluster* from http://function.princeton.edu/WCluster/ and understand the different parameters it accepts

g. Place the file *i29111.pcl* in the same directory as *WCluster* and run *WCluster* as follows:

```
java -Xmx2000m -jar WCluster.jar i29111.pcl no_weight E
i29111.cdt i29111.gtr i29111.atr
```

Note: *WCluster* uses only average-linking for the linkage criteria. This parameter cannot be changed so we will leave it as is for now

Using *TreeView*, Open the *i29111.cdt* file generated
h. Explain the main differences you see, i.e. what the different clusters represent

Different clusters represent groups of genes displaying similar expression profiles to one another.

i. Re-run *WCluster* again using the command in **g**. Open the generated *cdt* file once again in TreeView. Is the clustering the same as the one generated in **i**? Explain why or why not

No it is not. Every time hierarchical clustering is run, a different result is obtained. This is because, when building the clusters in an agglomerative

or bottom up approach, the cluster chosen to begin with is completely random for each run.

j. Experiment running *WCluster* with different distance metrics (**P** and **PA**). Explain why one would need to use different distant metrics when performing such clustering's

Different distance metrics calculate distances from clusters to clusters in separate ways. Euclidian (E), takes into account the direction and magnitude of the vectors (or profiles). Pearson (P), only accounts for direction. Pearson absolute (PA), would perform similarly to Euclidian but using Pearson correlations.

By the end of this exercise, you should:

- Have a general idea of how tandem repeats or patterns are discovered in biological sequences
- Fully understand why clustering is used in Bioinformatics and how to use it
- Apprehend technical details of Hierarchical clustering i.e. what it does, the different parameters it uses, file formats etc.

If you face any difficulties or have any questions, don't hesitate to email me: omar.wagih@utoronto.ca

Readings:

Benson G, Waterman MS. A method for fast database search for all k-nucleotide repeats. *Nucleic Acids Res.* 1994;**22**:4828–4836

The End