DNA Matching Tool
User Manual

Colin Phillips

August 18, 2015

1 Calculate alignment score on cluster

The User Manual is divided into three successive components.

- Retrieving the sequence data from the cluster
- Calculating the alignment scores of distinct sequences, two at a time
- Plotting the histogram as well as normal distribution of the collected data from many samples

1.1 Retrieving sequence data

The python scropt **seqretsplit.py** is used to retrieve sequences from the database stored on the cluster. It is run using the command **python seqretsplit.py**. The default setting allows for 250 sequences to be retrieved with a single iteration of the program. The sequence textfile name as well as the directory the sequences are stored in must be specified in the source code. The directory must be created prior to running the program.

1.2 Calculating the alignment scores

The python script **cmbproj.py** is used to calculate the alignment scores of two distinct sequences at a time. The scores are then collated into a text file named **score.txt**. Other text files may exist in the same directory as **score.txt** but these files are used as working files in order to extract the alignment scores from. They're given unique filenames in order to differentiate the multiple processes that are run simultaneously. The script is run using the command **python cmbproj.py (filename of folder 1) (filename of folder 2)**, where folder 1 stores one set of sequences and folder 2 stores another set.

1.3 Plotting the histogram

The python script **histogram.py** is used to plot a histogram of value versus frequency of value. It accepts values from two different text files each containing a set of alignment values. Along with the histogram, a normal distribution curve of the data is plotted coincident with the histogram.The script is run using the command **python histogram.py** , .