# **DNA Sequence Similarity Measure using LSH**

This application is designed to extract pairs of similar DNA Sequences from the given genomic [**dataset**](https://www.kaggle.com/thomasnelson/human-data) and user provided query. Query containing similar sequences are marked as a similar pair. *Locality sensitive hashing* is used to find similar pairs and Jaccard distance is used in the process. For each distance measure, a set of predicted similar Genomic Sequences are returned.

**Features**

* Returns pairs of ***similar*** sequences, which are answers to a question in our dataset.
* Data preprocessing of Human Genome dataset. **Jaccard distance** is used to find similar documents.
* The algorithm shows the **number of correct documents returned** for each distance measure.
* The signature matrix needs to be **generated** only **once** for each distance measure.
* Fully documented code.
* Necessary documents,files and code included pertaining to the assignment

## **How to run**

1. Extract the files.
2. Ensure Python 3.7 is installed, and in your system PATH.
3. Install necessary dependencies using pip install *<dependancy\_names>*

*<dependancy\_names>:*

* Numpy
* Matplotlib
* Pytest-timeit
* Random

1. Generated shingle-document matrix will be stored in **doc\_shingle.txt**..
2. To run the LSH algorithm:LSH.py. (for windows)

python3 LSH.py (for linux)

1. Enter shingle size
2. Enter query sequence
3. Enter Quit to exit

CS 469:IR Assignment-II

Abhishek Bapna 2018A7PS0184H

Hardik Parnami 2018A7PS0062H

Ashna Swaika 2018A7PS0027H