**Team Members:**

1. Shashank Koundinya Cheruku
2. Akhila Tummala
3. Chaitra Tirumalesh Kulkarni

**LSH (Locality Sensitive Hashing):**

For this assignment, we have used the dataset from the following website:

<http://osmot.cs.cornell.edu/kddcup/>

The dataset used is the bio\_train dataset from the above website. This dataset contains 145751 rows with 74 attributes in each row. Since the size of the dataset is very large, we have only retrieved 25000 rows with 74 attributes in each row. That brings upto around 3.8 million double values for the data array.

For the query points, we have taken the average of some of the rows in the dataset and generated 10 points with 74 dimensions. Since, we have taken the average values, all the values are nearer to each other. So, the number of visits are same for all the query points.

**Execution of the program in the server:**

In this email, we have attached only the code for the program. Since, the dataset is very large and cannot be attached to the email, we have uploaded the dataset on the Hrothgar server in the tmp folder. Please find the following directory in the tmp folder:

/tmp/shacheru\_lsh/

This directory contains three files:

1. The lsh.c program
2. The bio\_train\_bin.bin dataset binary file
3. The query\_bin.bin binary file of the query

**Compile Instructions:**

The C file can be compiled using the following command:

$gcc -lm lsh.c -o lsh

The output can be looked by executing the following command:

$./lsh

**Sample Output:**

Total Clusters = 252

Total visits = 24749

Minimum distance = 878.601242

Total visits = 24749

Minimum distance = 879.467754

Total visits = 24749

Minimum distance = 896.410423

Total visits = 24749

Minimum distance = 958.582773

Total visits = 24749

Minimum distance = 855.271610

Total visits = 24749

Minimum distance = 820.243821

Total visits = 24749

Minimum distance = 888.054864

Total visits = 24749

Minimum distance = 828.713731

Total visits = 24749

Minimum distance = 860.690623

Total visits = 24749

Minimum distance = 922.009246

**Results from kdtree:**

The same datasaet is used in the kdtree program and the following results are produced. The query dataset is also used in this program.

Points visited: 130

Minimum distance: 462.483744

Points visited: 129

Minimum distance: 535.017310

Points visited: 130

Minimum distance: 482.279395

Points visited: 130

Minimum distance: 489.883062

Points visited: 130

Minimum distance: 563.264904

Points visited: 98

Minimum distance: 419.915832

Points visited: 130

Minimum distance: 406.107845

Points visited: 98

Minimum distance: 441.125066

Points visited: 130

Minimum distance: 361.433432

Points visited: 160

Minimum distance: 526.319765

We weren’t able to compare the results from the Kmeans algorithm since, it is taking a very long time for the execution of the code with this large dataset.