INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

PROJECT REPORT

Similarity Analysis : For T-2 Diabetes (T2DM) using Moment of Inertia Tensor

---

Author:

Shisheer S KAUSHIK

A report is submitted in fulfillment of the requirements

for the coursework of Master of Technology

in the

Department of Mathematics

November 10, 2024

INDIAN INSTITUTE OF TECHNOLOGY JODHPUR

Abstract

Gaurav Bhatnagar

Department of Mathematics

Master of Technology

Similarity Analysis : For T-2 Diabetes (T2DM) using Moment of Inertia Tensor

by Shisheer S KAUSHIK

Significant advancements have been made in understanding the molecular associations related to Type II Diabetes Mellitus (T2DM). Numerous investigations have documented findings regarding the participation of different genes in the advancement of the disease. Nonetheless, the continuous advancement of bioinformatics tools has led to a significant increase in the prediction of various genes associated with the progression of T2DM, resulting in a high level of complexity when it comes to targeting these genes for further investigation. The candidate genes were prioritized through an analysis of sequence similarity with established T2DM genes. To achieve this objective, we treat the protein sequence as a rigid body possessing mass. Subsequently, we present the moment of inertia derived from the physiochemical characteristics of amino acids. The tensor converts the sequences into vectors pertaining to the moment of inertia. The Euclidean distance serves as a metric for assessing similarities.

...

Contents

Abstract iii

1 Introduction 3

1.1 Type 2 Diabetes(Mellitus) . . . 3

1.2 Data Collection . . . 4

2 Materials and Methods 7

2.1 Construction of protein Sequences as a 3D Model . . . 7

2.1.1 Distribution in X-Y Plane . . . 8

2.1.2 Distribution in Z Axis . . . 9

2.2 Moment of Inertia of a 3D Model . . . 11

2.2.1 Methods . . . 11

2.2.2 Results . . . 14

2.2.3 Data Collection (E-Fetch) . . . 14

2.2.4 Sorting and Prioritization . . . 15

3 Procedure 17

3.1 Flow Chart . . . 17

3.1.1 Data Collection (E-Fetch) . . . 18

3.1.2 Sorting and Prioritization . . . 19

3.2 Results . . . 20

3.2.1 Tensor Analysis of Known and Candidate Genes . . . 20

List of Figures

2.1 Caption 8
2.2 Caption 9
2.3 Caption 10
2.4 The information of the sequences used in our first test 14
2.5 Phylogenetic Tree of 9 species 15

3.1 Database Flowchart E-fetch 18
3.2 Sorting and Prioritization 19
3.3 DATA SHEET 21

Chapter 1

Introduction

1.1 Type 2 Diabetes(Mellitus)

Type 2 diabetes mellitus is a heterogeneous disorder with hyperglycemia as a common denominator. The most important pathophysiological features are impaired insulin secretion and decreased insulin sensitivity (insulin resistance), the latter related to the liver and extrahepatic tissues, mainly skeletal muscle and adipose tissue. (1) In type 2 diabetes, as in all other forms of diabetes, there is often development of late complications in several organs because of microangiopathy and/or other deleterious processes in, e.g. retina, kidneys and nerves. In addition, macroangiopathy leads to severalfold increased risk of cardiovascular disease. (2)

With advancements in molecular techniques, systems biology, bioinformatics, next-generation sequencing, microarray, and so forth, there has been a substantial improvement in understanding the underlying molecular mechanisms. In recent years, there has been an enormous accumulation of data regarding the prediction of disease candidate genes through biological network analysis. However, the experimental validation of each predicted gene is highly costly, and the resources cannot be wasted on this vast number of candidate proteins. Unfortunately, the accumulated data on candidate genes for cervical cancer is becoming redundant as few bioinformatics tools are

available for prioritizing the candidate genes for T2DM. Therefore, it is imperative to develop a bioinformatic method to prune and prioritize the genes for further evaluation. Various methods have been developed to analyze the sequence similarity between protein sequences to understand the functional similarity of the proteins. However, alignment-free methods have more benefits than alignment-based methods. Recently, Piotr Wąż and Bielin´ska-Wąż developed a technique using the concept of moment of inertia tensor for similarity analysis of DNA sequences. Later, Hou and co-workers introduced a method applying the same idea of tensor to measure the sequence similarity between the proteins.

Moment of inertia tensor is an alignment-free based method that is fast, efficient, and reliable in comparing the sequence similarity. We prioritized the candidate cancer genes through sequence similarity analysis between the known T2DM genes (KDGs) and the candidate T2DM genes (CDGs) using the moment of inertia tensor. This computational approach helps reduce the wastage of resources and decreases the efforts in designing the drugs for the candidate proteins.

1.2 Data Collection

We collected the list of genes that cause or involved in cancer from the database of Network of Cancer Genes (NCG6.0), which is a manually curated repository on systems-level properties of cancer. The genes list consists of 711 known cancer genes and 1661 candidate cancer genes based on the approach they were identified in various cancer studies. In addition, we collected the genes list datasets related to cervical cancer progression from the Cervical Cancer gene DataBase (CCDB). It is a manually curated catalog consisting of 537 genes involved in the different stages of cervical carcinogenesis. By mapping the list of cervical cancer genes obtained from CCDB with the list of cancer genes present in NCG6.0, we found a total of 128 genes that are

common in both databases. Among these 128 genes, 76 genes are identified as known cancer genes, and 52 genes are identified as candidate cancer genes. Further, these 128 genes are investigated for their association with cervical cancer in DisGeNET (a database of gene-disease associations) and found that 82 genes have experimentally validated evidence associated with cervical cancer. So, we considered these 82 genes as KCCs and the remaining 46 genes as CCCs. We retrieved the protein sequences of 128 cervical cancer genes from the Uniprot database through the biomaRt library on the R platform. Further, we curated the sequences using the BioStrings library on the R platform to their canonical sequences in FASTA format.

Chapter 2

Materials and Methods

2.1 Construction of protein Sequences as a 3D Model

A protein sequence consists of twenty different amino acids while a DNA sequence consists of only four bases. The delay in the emergence of graphical representation of protein sequences is partially because of the diversity in the amino acids.

In this section, we outline a 3-D graphical representation based on the physicochemical properties of amino acids. First, according to the classification of amino acids by different properties, we locate the amino acid on a unit circumference. Then we build a 3-D model to describe the sequences respectively. With the application of the tensor for moments of inertia, we calculate the similarities among different protein sequences. We test our scheme using different data, and it is demonstrated that our results agree with evolutionary relations satisfactorily.

It is generally accepted in bioinformatics that amino acid sequences determine the spatial structure of proteins.In our approach, we extract two main physicochemical properties of amino acids as the descriptor to reflect the innate relation among different proteins: the hydrophobicity and molecular mass.

For the twenty kinds of amino acids, we divide them into two groups by their different hydrophobicity:

Hydrophobic amino acids H F, L, I, Y, W, M, V, A, P, C;

Hydrophilic amino acids P S, N, K, D, R, T, H, Q, E, G.

Then, for a further classification, the amino acids are divided into four types :

Strong hydrophilic amino acids SP S, N, K, D, R

Weak hydrophilic amino acids WP T, H, Q, E, G

Strong hydrophobic amino acids SH F, L, I, Y, W

Weak hydrophobic amino acids M, V, A, P, C

2.1.1 Distribution in X-Y Plane

According to the hydrophobicity, different types of amino acid are located into different quadrants. The amino acids are ordered along the circumference, which is of unit radius, alphabetically according to the abbreviations of their names. The 20 points on the circumference of the circle have the coordinates given by:

$$x_i = cos(2\pi i/20), y_i = sin(2\pi i/20), i = 0, 1, 2...19.$$

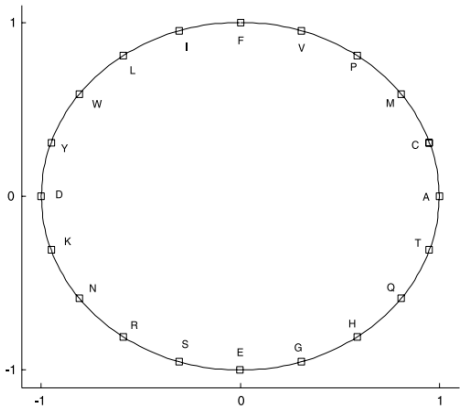


FIGURE 2.1: Caption

The hydrophobic amino acids are placed in the first and second quadrant

while hydrophilic ones are placed in the third and fourth quadrant. The distribution of amino acids in two dimensional Cartesian coordinates is illustrated in Fig.

2.1.2 Distribution in Z Axis

The z-axis coordinate of amino acid is determined by their relative residue weight. According to the weight, 20 amino acids are ranked as:

G < A < S < P < V < T < C < I = L < N < D < Q < K < E < M < H < F < R < Y < W.

Amino acids have equal residue weight, and we arrange the symbol in alphabetic order. The z-axis coordinates of ten amino acids with smaller molecular mass are labelled by -1. Other amino acids are labelled by 1 on the z-axis. The z-axis values of 20 amino acids are listed in the Table.

Aminoacid	Symbol	Residue. wt	Z
Alanine	A	71.08	−1
Cysteine	C	103.14	−1
Methionine	M	131.19	1
Proline	P	97.12	−1
Valine	V	99.13	−1
Phenylalanine	F	147.17	1
Isoleucine	T	113.16	−1
Leucine	L	113.16	−1
Tryptophan	W	186.21	1
Tyrosine	Y	163.18	1
Asparticacid	D	115.09	1
Lysine	K	128.17	1
Asparagine	N	114.10	−1
Arginine	R	156.19	1
Serine	S	87.08	−1
Glutamic acid	E	129.12	1
Glycine	G	57.05	−1
Histidine	H	137.14	1
Glutamine	Q	128.13	1
Threonine	T	101.11	−1

FIGURE 2.2: Caption

For a specific protein sequence, from the start to the end, we locate the amino acid in the corresponding position, connecting every point in turns.

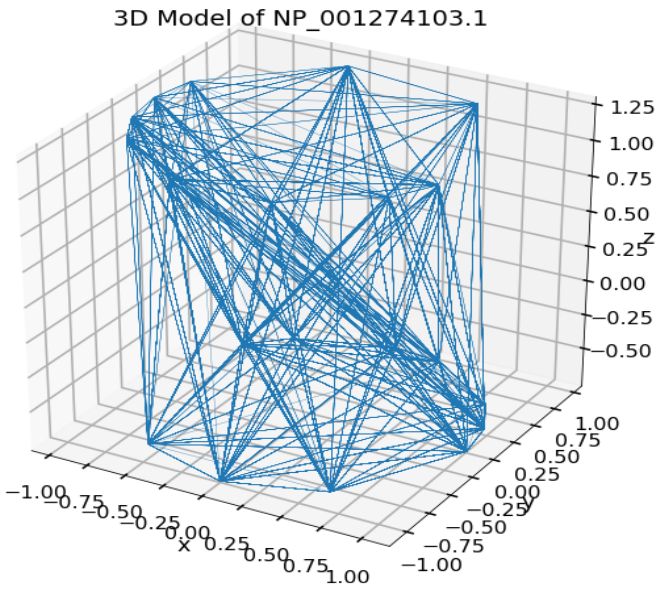


FIGURE 2.3: Caption

2.2 Moment of Inertia of a 3D Model

$$\vec{L} = \vec{A} \times \overrightarrow{\vec{b}} = m\vec{F} \times (\vec{w} \times \vec{r})$$

$$L = \sum_i m_i \left(\vec{\omega} \left(\vec{r}_i \circ \vec{n}_i\right) - \vec{r}_i \left(\vec{r}_i \circ \vec{\omega}\right)\right)$$

$$L = \sum_i m_i \left[\vec{\omega} \left(x_1^2 + y_i^2 + z_i^2\right) - \vec{r}_i \left(x_i\omega_z + y_i\omega_y + z_i\omega_z\right)\right]$$

$$L_x = \sum_i m_i \left[w_x \left(y_i^2 + z_i^2\right) - z_i u_i w_y - x_i z_i w_z\right]$$

$$L_y = \sum_i m_i \left[w_y \left(z_i^2 + z_i^2\right) - y_i x_i - w_2 - y_i z_i w_z\right]$$

$$L_z = \sum_i m_i \left[w_z \left(x_i^2 + y_i^2\right) - z_i y_i w_x - z_i y_i w_y\right]$$

$$\left(L_i = \sum_{j=1}^{3} I_{ij}\omega_j\right)$$

$$\begin{pmatrix} L_x \\ \\ L_y \\ L_y \end{pmatrix} = \begin{pmatrix} \sum_i m_i \left(y_i^2 + z_i^2\right) & -\sum_i m_i x_i y_i & -\sum_i m_i x_i z_i \\ -\sum_i m_i \cdot y_i x_i & \sum_i m_i \left(x_i^2 + z_i^2\right) & -\sum_i m_i y_i z_i \\ -\sum_i m_i z_i x_i & -\sum_i m_i z_i y_i & \sum_i m_i \left(x_i^2 + y_i^2\right) \end{pmatrix} \begin{pmatrix} w_x \\ w_y \\ w_i \end{pmatrix}$$

2.2.1 Methods

The moments of inertia of a 3D graph are applied in our method. It is first introduced in bioinformatics by Wąż and Bielińska-Wąż in (Reference). They model a DNA sequence as a set of "material points" in the 3D space. Then, they characterize the sequence by moments of inertia. In the present work, we apply the method to the analysis of protein sequences. A material point represents each amino acid in the protein sequence. Similarly, to simplify

the calculation, we assigned the mass m=1. The points are distributed as described before in the 3-D Cartesian coordinates. The coordinates of the centre of mass of the 3-D graph in the Cartesian coordinate system are defined as

$$\mu_x = \frac{\sum_i m_i x_i}{\sum_i m_i}, \mu_y = \frac{\sum_i m_i y_i}{\sum_i m_i}, \mu_z = \frac{\sum_i m_i z_i}{\sum_i m_i}$$

where x_i, y_i, z_i are the coordinates of material point m_i

The tensor of the moments of Inertia is defined by the matrix:

$$\hat{I} = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix}$$

$$I_{xx} = \sum_i m_i \left(\left(y_i^\mu\right)^2 + \left(z_i^\mu\right)^2 \right)$$

$$I_{yy} = \sum_i m_i \left(\left(x_i^\mu\right)^2 + \left(z_i^\mu\right)^2 \right)$$

$$I_{zz} = \sum_i m_i \left(\left(x_i^\mu\right)^2 + \left(y_i^\mu\right)^2 \right)$$

$$I_{xy} = I_{yx} = \sum_i m_i x_i^\mu y_i^\mu$$

$$I_{yz} = I_{zy} = \sum_i m_i y_i^\mu z_i^\mu$$

$$I_{xz} = I_{zx} = \sum_i m_i x_i^\mu z_i^\mu$$

where $x_i^\mu, y_i^\mu, z_i^\mu$ are the coordinates of mi in the Cartesian coordinate system for which the origin has been selected at the center of mass. We calculate the eigenvalues of matrix $\hat{I}$, which is labeled by λ_1, λ_2 *and* λ_3 . Let us define

the vector $\overrightarrow{v}(S) = (\lambda_1, \lambda_2, \lambda_3)$ to represent the protein sequence S, we obtain the similarity of two sequence S1 , S2 from the Euclidean distance

$$D\left(S^1, S^2\right) = \left\| \vec{v}^{(S^1)} - \vec{v}^{(S^2)} \right\|_2$$

2.2.2 Results

The ND5 protein sequences from 9 species are widely used in different articles and are considered a standard to evaluate the model. All the sequences are picked from the NCBI database. We notice that the pairs(blue whale, fin whale),(common chimpanzee, gorilla), (human, common chimpanzee), (pigmy chimpanzee, common chimpanzee) and (pigmy, chimpanzee, gorilla) have a shorter distance according to our models.

The primates, such as the human, common chimpanzee, pygmy chimpanzee and gorilla, are on the same tree branch. Human and common chimpanzee has the shortest distance. Besides, the blue and fin whales, rats and mice are similar in our calculation. These results agree with the classical evolution theory.

2.2.3 Data Collection (E-Fetch)

Abbreviation	Accession No.	Length
Human	AP_000649	603
Gorilla	NP_008 222	603
Common chimpanzee	NP_008 196	603
Pigmy chimpanzee	NP_008 209	603
Blue whale	NP_007066	606
Fin whale	NP_006899	606
Rat	AP_004902	610
Mouse	NP_904338	607
Opossum	NP_007 105	602

FIGURE 2.4: The information of the sequences used in our first test

2.2.4 Sorting and Prioritization

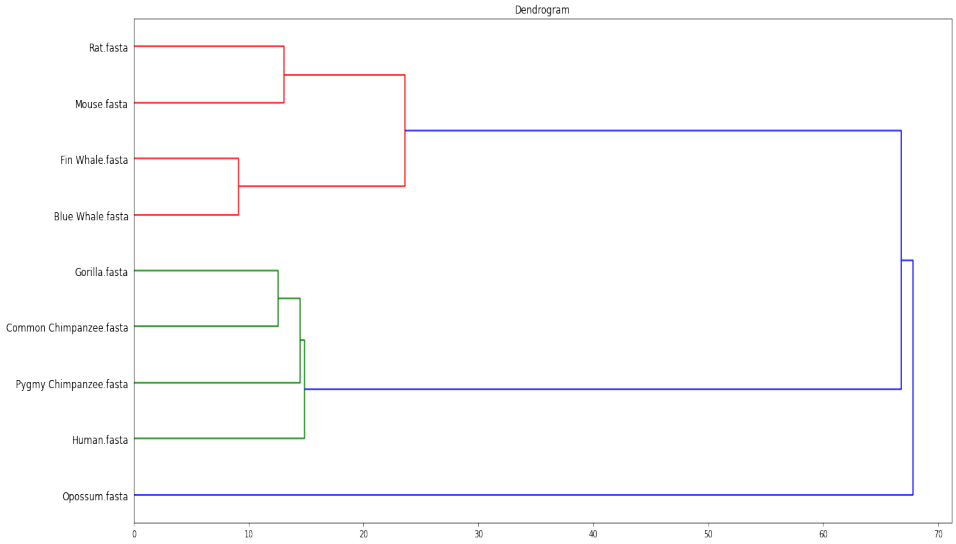
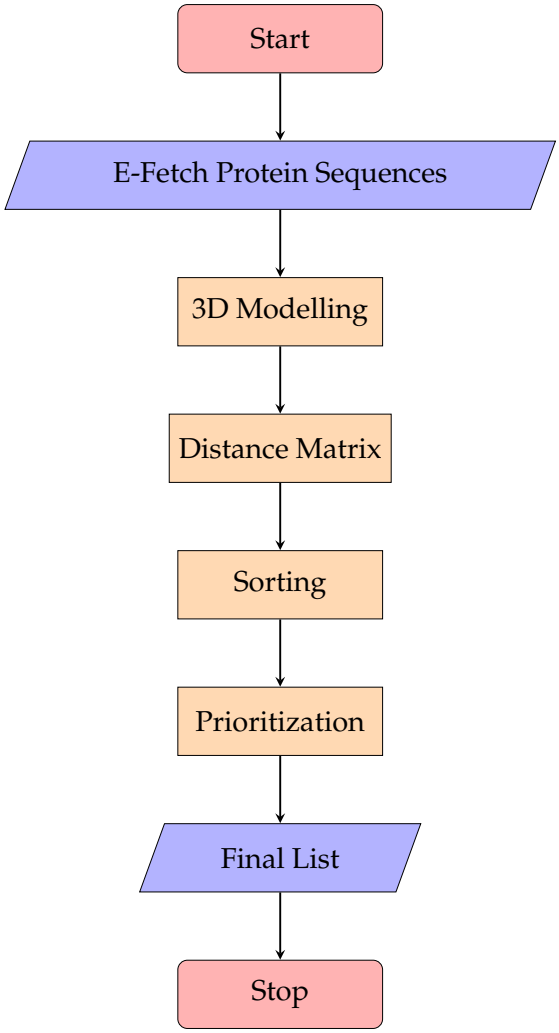


FIGURE 2.5: Phylogenetic Tree of 9 species

Chapter 3

Procedure

3.1 Flow Chart



3.1.1 Data Collection (E-Fetch)

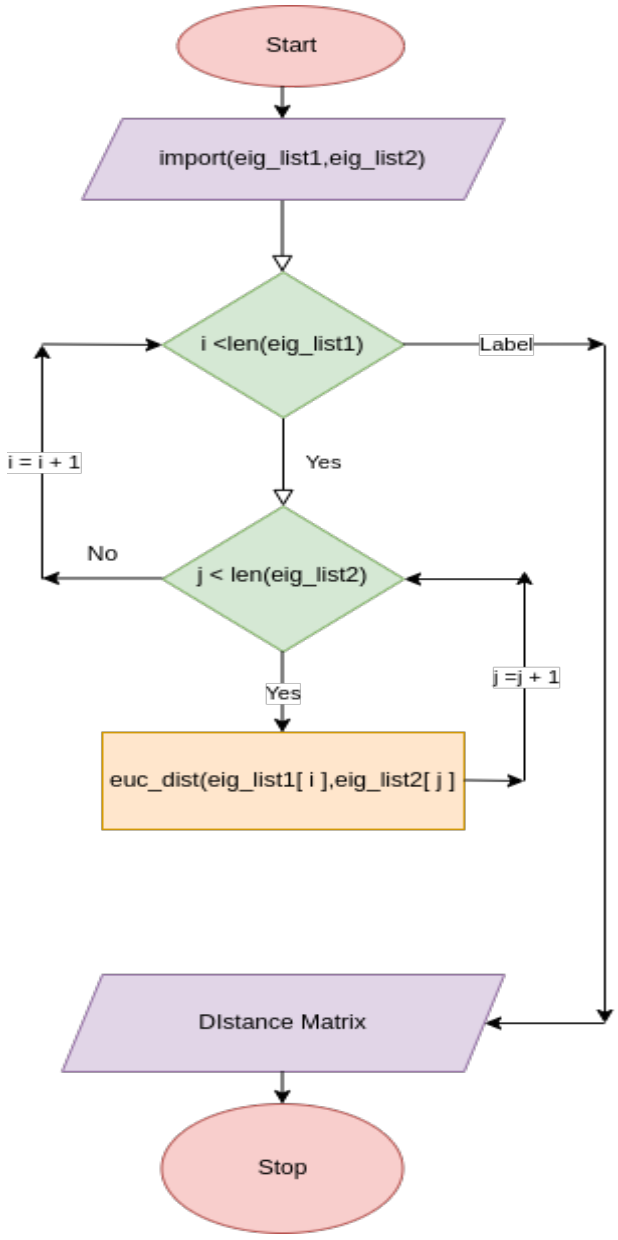


FIGURE 3.1: Database Flowchart E-fetch

3.1.2 Sorting and Prioritization

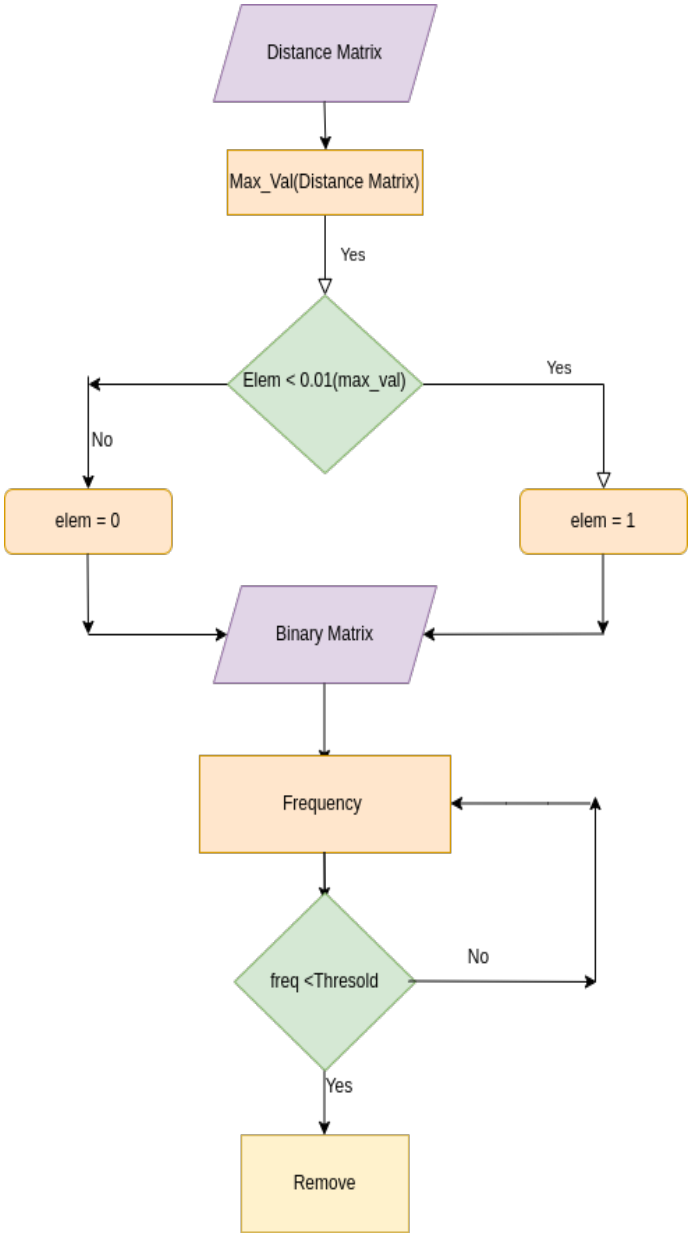


FIGURE 3.2: Sorting and Prioritization

3.2 Results

3.2.1 Tensor Analysis of Known and Candidate Genes

Using the tensor for Moment of Inertia, we analyzed the similarities between known genes(Experimentally Verified) and the Candidate Protein. Each protein sequence's moment of inertia matrix is considered to calculate the Euclidean distance between any two protein sequences. We constructed the distance matrix by calculating the Euclidean distance between the known and candidate proteins. From the distance matrix, we constructed a dendrogram. The resulting distance matrix ranges between 16.227 and 80710.615. The maximum distance (slightest similarity) is observed between QP66V0 and P01275.

Using the distance matrix, we prioritized the Candidate proteins that showed 0.1 percent or less distance (99.9 percent similarity or more) from Known proteins concerning the maximum distant (least similar) proteins. Further, we considered the proteins that showed similarities with more Known proteins. In our study, we picked the proteins that show at least fourteen or more associations with Known proteins. At the end we sorted 94 genes out of 2856 Candidate Genes which had shown 99.9 percent accuracy with the Known Genes.

FIGURE 3.3: DATA SHEET

serial no.	candidate protein	Frequency
1	P00568	14
2	P48047	16
3	Q07812	14
4	O43521	15
5	P18075	15
6	Q96NL8	16
7	Q08708	15
8	P60033	14
9	P46527	14
10	O60543	15
11	P16410	16
12	Q07507	15
13	P14416	14
14	Q14213	15
15	Q8WWZ3	15
16	P05305	16
17	P98173	14
18	Q92520	14
19	Q96BQ1	14
20	Q92914	14
21	Q14314	15
22	Q9H6D8	15
23	Q8NAU1	15
24	P28676	15
25	P39905	14
26	P55789	16
27	Q9BX51	15
28	B5MD39	16
29	P07203	14
30	P62993	15
31	P08263	14
32	P09488	15
33	Q9NRV9	14
34	P09429	15
35	P01112	14
36	P04792	15
37	Q9UJY1	14
38	P28335	14
39	Q9Y6W8	14
40	P01562	14
41	P05019	14
42	P29459	14
43	Q8TAD2	14
44	Q14116	14
45	Q9NZH6	15
46	P05231	16

serial no.	candidate protein	Frequency
47	O14713	14
48	P26718	15
49	P80188	15
50	Q9H9Z2	17
51	O95237	14
52	O75608	15
53	P13727	14
54	Q9H2W2	14
55	O60682	15
56	Q05195	14
57	P62166	14
58	Q9Y4Z2	15
59	P25208	14
60	O75469	15
61	P49763	14
62	Q8TCI5	14
63	P41236	14
64	Q06830	15
65	P32119	15
66	Q9ULZ3	14
67	P62491	14
68	P61106	14
69	Q9NP72	15
70	P51159	14
71	P61020	14
72	P02753	16
73	Q6ZTI6	17
74	P61586	14
75	O00212	17
76	Q99578	15
77	Q96AT9	14
78	P10301	15
79	Q9NP50	14
80	O00161	14
81	P60880	15
82	O15524	17
83	O14543	15
84	Q9BQB4	14
85	P30626	15
86	P35625	17
87	P58753	17
88	Q96LR5	15
89	P09936	14
90	P62760	14
91	O15498	15
92	Q9H8U3	14

serial no.	candidate protein	Frequency
93	Q6FIF0	15
94	Q15915	14