UNIVERSITY OF HYDERABAD

MASTERS THESIS

Prioritizing the candidate genes related to T2DM using Moment of Inertia Tensor

Author:
Haris ANSARI

Supervisor:
Dr. P MANIMARAN

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Complex Network Analysis
School of Physics

June 10, 2022

Declaration of Authorship

I, Haris ANSARI, declare that this thesis titled, "Prioritizing the candidate genes related to T2DM using Moment of Inertia Tensor" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Thanks to my solid academic training, today I can write hundreds of words on virtually any topic without possessing a shred of information, which is how I got a good job in journalism."

Dave Barry

UNIVERSITY OF HYDERABAD

Abstract

P Manimaran

School of Physics

Master of Science

Prioritizing the candidate genes related to T2DM using Moment of Inertia Tensor

by Haris ANSARI

There is a tremendous improvement in realizing the underlying molecular associations in Type -II Diabetes(Mellitus)(T2DM). Several studies reported evidence for the involvement of various genes in the disease progression. However, with the ever-evolving bioinformatics tools, there has been an upsurge in predicting numerous genes responsible for T2DM progression and making it highly complex to target the genes for further evaluation. We prioritized the candidate genes based on the sequence similarity analysis with known T2DM genes. For this purpose, We consider the protein sequence as a rigid body with mass. Then we introduce the moment of inertia based on the physiochemical properties of amino acids. The sequences are transformed into vectors by the tensor for the moment of inertia. The Euclidean distance is employed as a measurement of the similarities. . . .

Acknowledgements

I wish to express my sincerest gratitude to Prof.P Manimaran for the continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve the targets by explaining all the tasks to be done and made sure that I understood all the concepts which kept me motivated for the project. He was always available to help and clear any doubts regarding difficulties in this project. Without his constant support and motivation, this project would not have been successful. I would like to thank Gerin Prince, and Sana Fatima, my friends and PhD Scholars, for their help in various aspects of the project. Throughout my work, they have been constant support; they were always willing to help, without which the work would not have gone smoothly.. . .

Contents

List of Figures

For/Dedicated to/To my…

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 several-fold increased risk of cardiovascular disease. (2)

GENETICS VS. ENVIRONMENT

It is generally agreed that the pathogenesis of the disease has got strong genetic and environmental components (3) (Fig. 1). So far, the genetic background has been characterized in only a minor part of the patients; among those, the MODY (maturity- onset diabetes of the young) types constitute 5±10autosomally dominant inheritance is mediated by mutations in the genes coding for glucokinase (MODY-2), hepatocyte nuclear factors (HNF)-4a, -1a and -1b (MODY-1, MODY-3 and MODY-5, respectively) and islet proliferation factor (IPF; MODY-4). The HNFs and IPF are transcription factors with decisive importance for the development and function of the pancreatic b-cells, while glucokinase plays a role for b-cell as well as liver metabolism. Some other subgroups (5±10patients with latent autoimmune diabetes of the adult, or LADA (5) , and diabetes secondary to rare genetic syndromes (6) .

The majority, 70±85polygenic inheritance which acts in concert with environmental factors in the development of the disease via a stage of impaired glucose tolerance. The environmental, or acquired, factors often relate to lifestyle, and comprise overweight, low physical activity, and tobacco use (7) . In fact, many of these patients present themselves with several features of the metabolic or insulin resistance syndrome (‘syndrome X&#39;), (8) . In this syndrome, there is clustering of several clinical and biochemical alterations, as abdominal obesity, hypertension, glucose intolerance, insulin resistance, dyslipidaemia (low HDL cholesterol as well as high LDL cholesterol and triglyceride levels). (9)

Epidemiology of Diabetes in India

Diabetes is one of the largest global health emergencies of this century, ranking among the 10 leading causes of mortalitytogether with cardio vascular disease (CVD), respiratory disease, and cancer.[1,2] According to the World Health Organization (WHO), non communicable diseases (NCDs)accounted for 742019, of which, diabetes resulted in 1.6 million deaths, thus becoming the ninth leading cause of death globally.[2] By the year 2035, nearly 592 million people are predicted to die of diabetes.[3] Type 2 diabetes, which constitutes 90earlier considered to be a disease of the affluent “Western” countries, has now spread globally, and has become a major cause of disability and death affecting even younger age group.[1] Diabetes has reached epidemic proportions in many developing economies, such as China and India.[1] According to WHO, the prevalence of diabetes is growing most rapidly in low-.and middle-income countries.[4] The rapid socioeconomic change in conjunction with urbanization and industrialization are the major factors for the global increase in the diabetes epidemic, with other associated risk factors such as population growth, unhealthy eating habits, and a sedentary lifestyle also playing an important role.[5] Diabetes is a progressive disorder that leads to serious complications, which are associated with increased costs to the family, community, and healthcare system. Uncontrolled diabetes leads to increased risk of vascular disease and much of the burden of type 2 diabetes is caused by macrovascular (cardiovascular (CV), cerebrovascular, and peripheral artery disease) and microvascular (diabetic retinopathy, nephropathy, and neuropathy) complications.[5,6]

Prevalence of diabetes and trends over time In India, the burden of diabetes has been increasing steadily since 1990 and leaps and at a faster pace from the year 2000.Fig. 3 shows the increasing trend in diabetes prevalence in India during the past decade in India as per IDF.[1,12-16] The prevalence of diabetes in India has risen from 7.1% in 2009 to 8.9% in 2019. Currently, 25.2 million adults are estimated to have IGT, which is estimated to increase to 35.7 million in the year 2045. India ranks second after China in the global diabetes epidemic with 77 million people with diabetes. Of these, 12.1 million are aged &gt;65 years, which is estimated to increase to 27.5 million in the year 2045. It is also estimated that nearly 57million. The mean healthcare expenditure on diabetes per person is 92 US dollars, and total deaths attributable directly to diabetes account for 1 million.

INSULIN SENSITIVITY AND GLUCOSE TOLERANCE

Insulin sensitivity and impaired glucose tolerance are other factors that have been investigated as possible predictors of the development of type 2 diabetes. Results of a 5-year prospective study of insulin resistance in Pima Indians showed a clear relationship between impaired glucose tolerance and the subsequent development of type 2 diabetes (12). Furthermore, analysis of 2-h insulin profiles showed that there was a nearly linear relationship between insulin concentration and the development of type 2 diabetes; a similar relationship was seen with fasting insulin concentrations (Fig. 2). According to this report, low insulin response and increased insulin resistance are both predictors of type 2 diabetes, and each variable acts as an independent risk factor (12). When an individual with diabetes presents to a physician, two questions are typically on their minds: firstly, 'why did I develop diabetes and how could I have possibly avoided it?'; and secondly, 'how can you cure me?' Although we are still at a loss to explain the complete pathophysiology and pathogenesis of Type 2 diabetes at the molecular level, epidemiological studies provide us with adequate guidelines for identifying the individuals at highest risk for its development. Biochemical criteria remain the best predictors for the future development of Type 2 diabetes. The presence of diabetes during pregnancy yields a 33 impaired glucose tolerance (IGT) develops into Type 2 diabetes at a rate of approximately 7 patients at risk who may then undergo biochemical assessment for disease. Family history of diabetes, being a member of an ethnic minority, obesity, age .65 years and simultaneous presence of hypertension and/or hypertriglyceridaemia make the possibility of IGT greater. In the USA, 11 has IGT; this figure increases to 18 highest risk of developing diabetes allows the introduction of interventions that may delay or prevent the progression of diabetes. Interventions designed to reduce insulin resistance and protect the b-cell from failure could theoretically prevent or delay the progression to Type 2 diabetes. Although prevention strategies are the best hope for stemming the epidemic of Type 2 diabetes, they are of limited value in aiding those patients already afflicted with the disease (Table 2). The United Kingdom Prospective Diabetes Study (UKPDS) has clearly demonstrated that Type 2 diabetes is a progressive disease with consistent deterioration in glycaemic control over time, which demands increased pharmacological intervention.6 In time, at least 33 hyperglycaemia.

Table 2. Cures for diabetes

- Closed-loop insulin infusion system
- Pancreas transplants
- Islet-cell transplants

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

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:

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

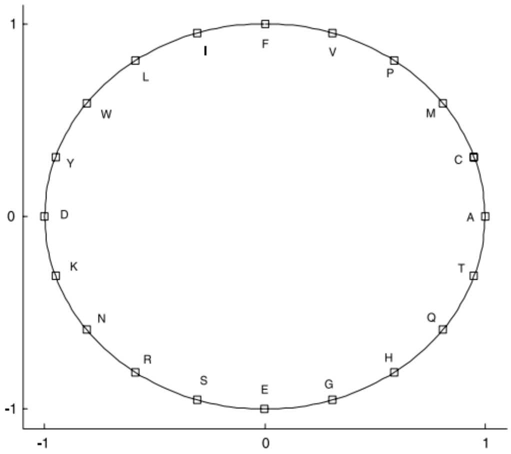


FIGURE 2.1: Caption

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.

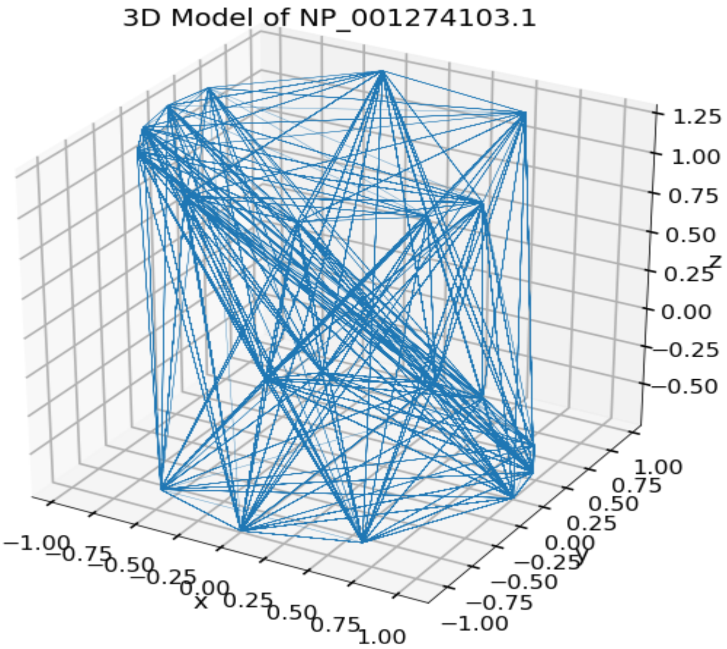


FIGURE 2.3: Caption

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 Î, 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.

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

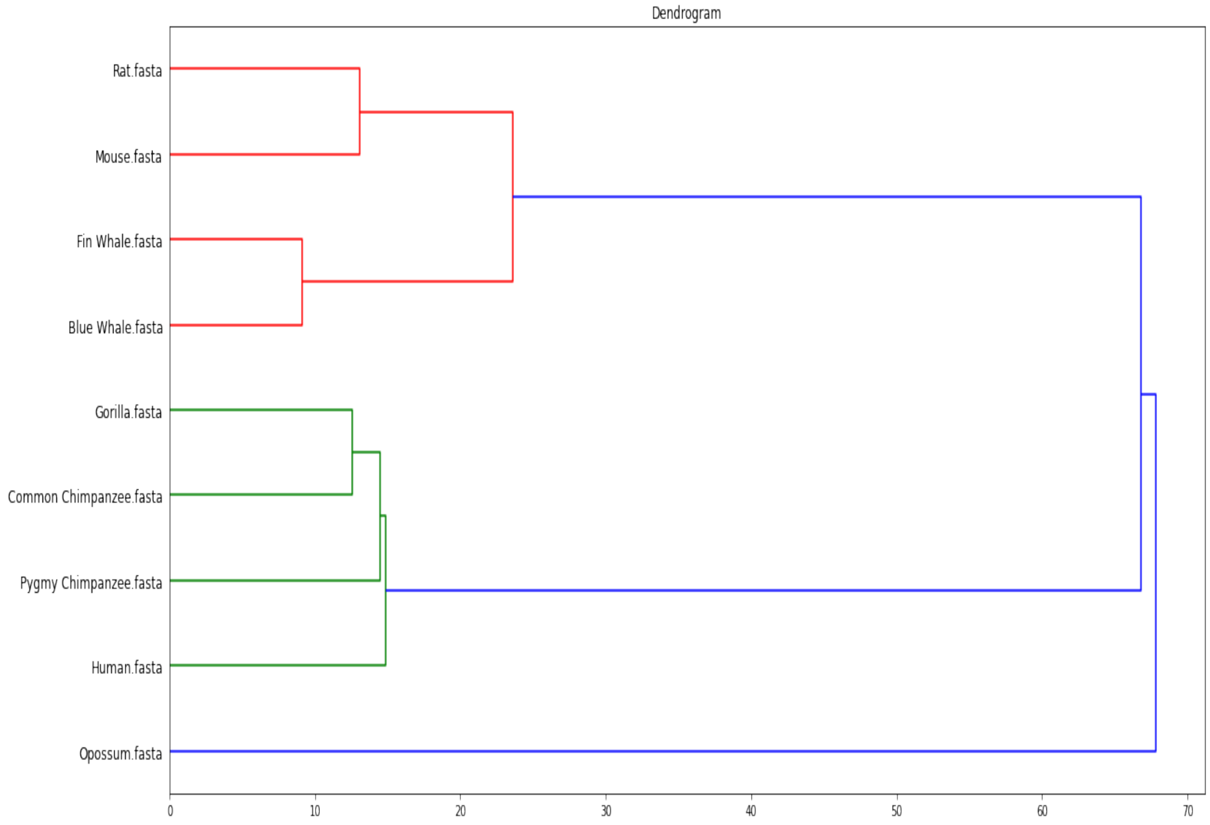
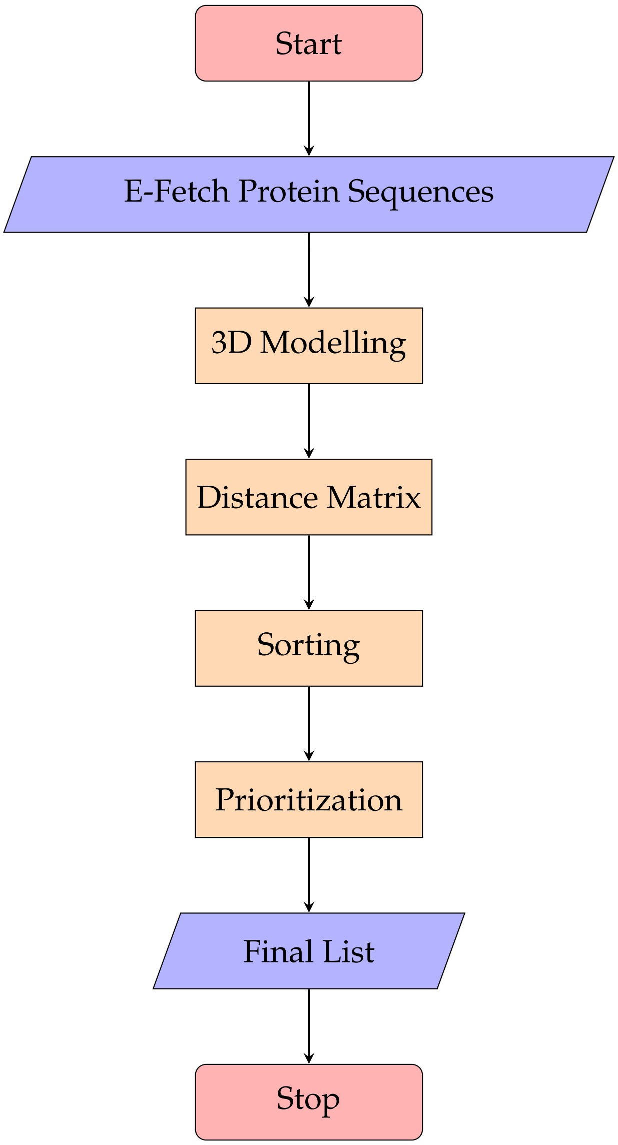


FIGURE 2.5: Caption

Chapter 3

Procedure

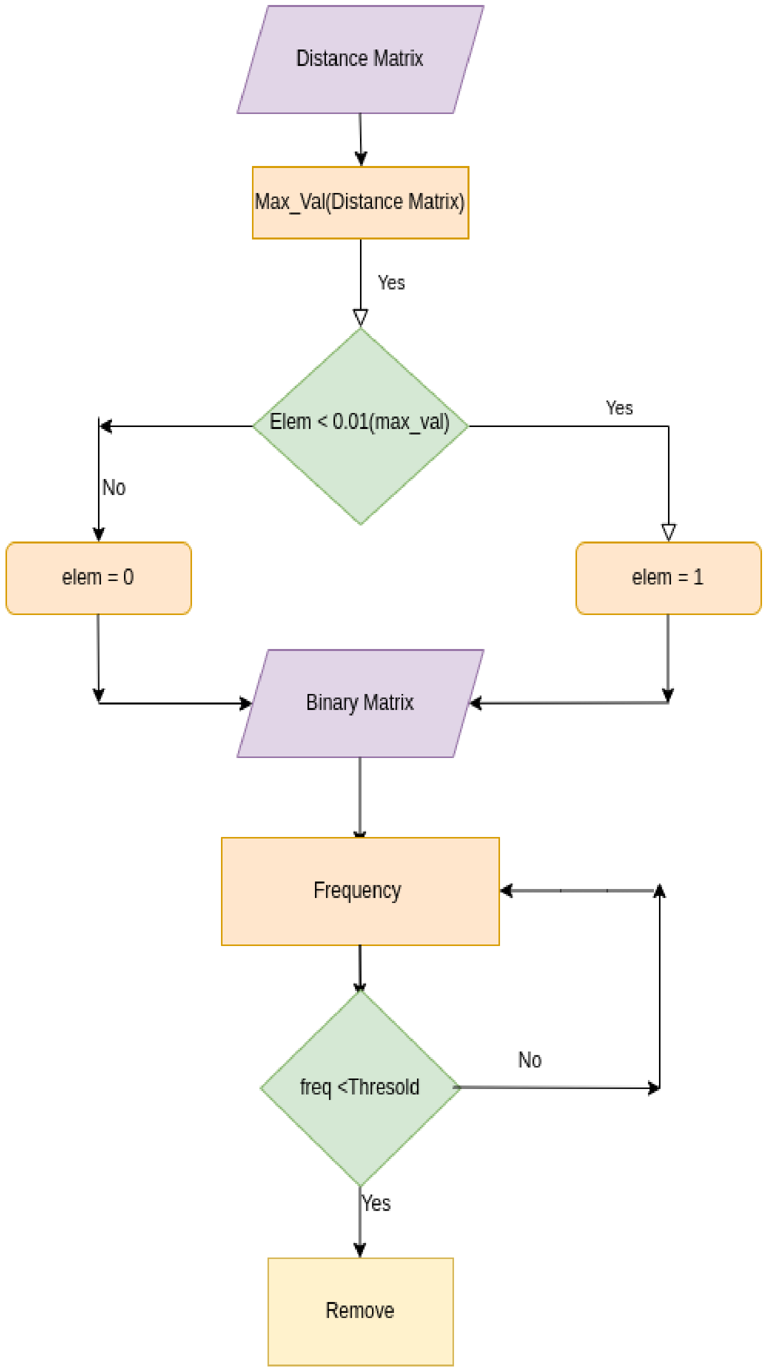
3.1 Flow Chart



3.1.1 Data Collection(E-Fetch)

3.1.2 Database(E-Fetch)

3.1.3 Sorting and Prioritization



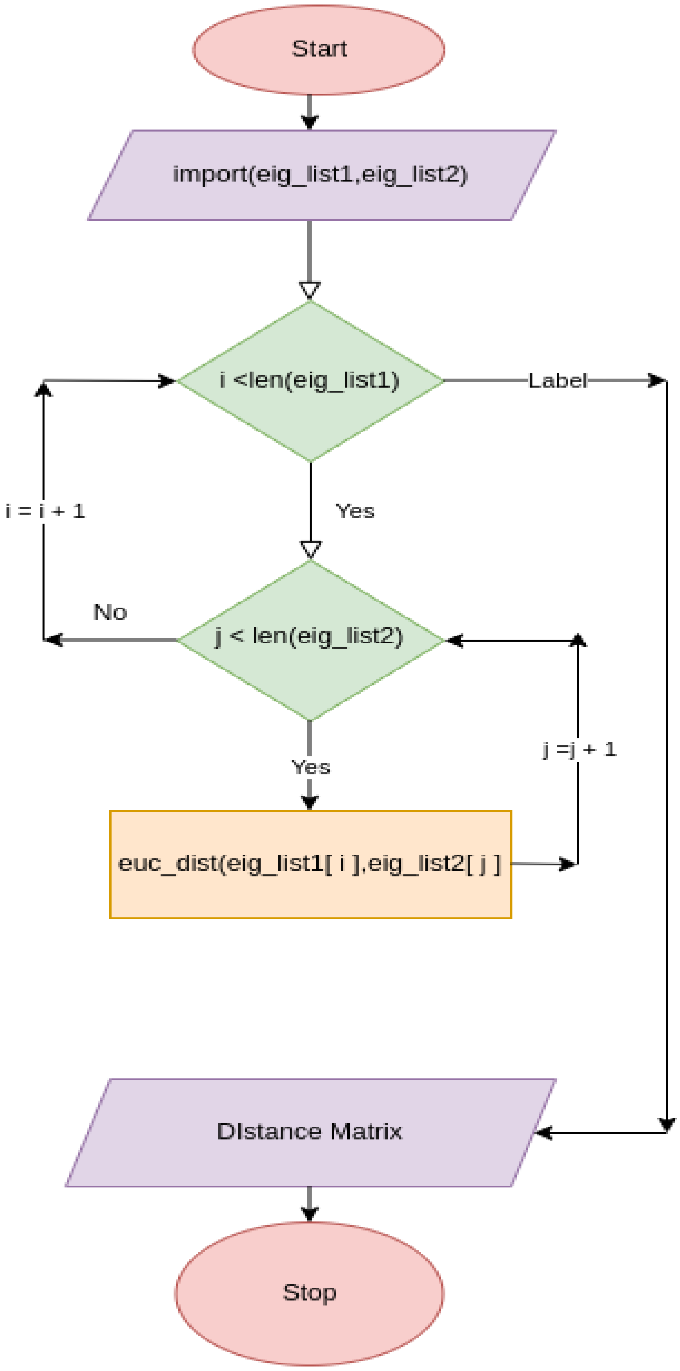


FIGURE 3.1: Database Flowchart E-fetch

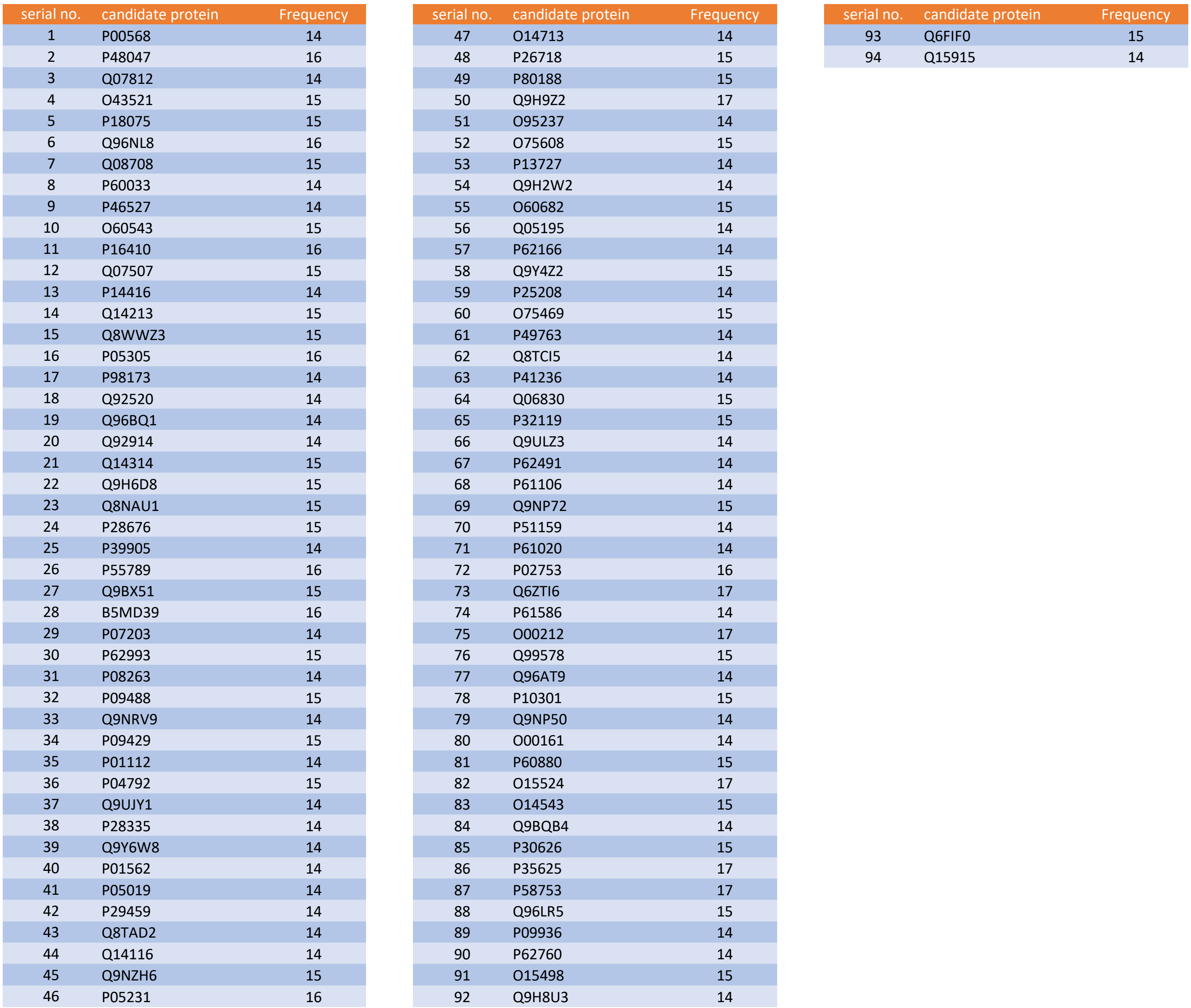
3.2 Results

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 3.2: 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
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
93	Q6FIF0	15
94	Q15915	14

Bibliography

DeFronzo, Ralph A, Riccardd C Bonadonna, and Eleuterio Ferrannini (1992). "Pathogenesis of NIDDM: a balanced overview". In: *Diabetes care* 15.3, pp. 318–368.