

# Robust Gene Network Topology Construction Based on the Evolutionary Algorithm and Artificial Neural Network

## Submitted By

Md. Tanvir Aunjum  
ID: 2014-1-60-030

Md.Nazmul Hasan  
ID: 2014-1-60-041

Md. Jahidur Rahman  
ID: 2014-1-60-055

## Supervised By

Md. Shamsujjoha  
Senior Lecturer  
Dept. of CSE, EWU

A thesis Submitted in Partial Fulfillment of the Requirements for the Degree of  
Bachelor of Science in Computer Science and Engineering



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
EAST WEST UNIVERSITY

August 2018

# Abstract

Design and implementation of automatic gene regulatory network are essential to construct and analyze the complex biological system. The recent study shows that Darwinian evolution can gradually develop higher topological robustness. In these consequences, this thesis presents an integrated scheme to simulate gene expressions dataset for identifying network topologies to find the robustness based on an evolutionary approach and artificial neural network. The final outcome is the most robust topology from a gene regulation dataset. The proposed method was verified using randomly sampled parameter spaces and threshold are generated by the network itself. Here, final result shed lights on the relationship among genes and corresponding transcription factors. Transcription factors are combined to specify the on-and-off states of genes. This binding form a regulatory network and constituting the wire diagram for a cell. The proposed network shows the whole combinatorial and co-association of transcription factors, co-relation and the robustness of human genes. Therefore, this research will play a crucial role in interpreting personal genome sequences and understanding basic principles of human health evolution in near future.

# Declaration

I hereby, declare that all the work presented in this project is the outcome of the investigation and research performed by us under the supervision of Md. Shamsujjoha, senior Lecturer, Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh. I also declare that neither it nor part of it has been submitted for the requirement of any degree or diploma or for any other purposes except for publications.

Signature of the Candidate

.....

Md. Tanvir Aunjum

ID: 2014-1-60-030

Signature of the Candidate

.....

Md.Nazmul Hasan

ID: 2014-1-60-041

Signature of the Candidate

.....

Md. Jahidur Rahman

ID: 2014-1-60-055

# Letter for Acceptance

This thesis entitled “**Robust Gene Network Topology Construction Based on the Evolutionary Algorithm and Artificial Neural Network**” submitted by Md. Tanvir Aunjum (ID: 2014-1-60-030), Md. Nazmul Hasan (ID: 2014-1-60-041), Md. Jahidur Rahman (ID: 2014-1-60-055), to the Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh is accepted by the department in partial fulfillment of requirements for the Award of the Degree of Bachelor of Science in Computer Science and Engineering on August, 2018.

Supervisor

---

**Md. Shamsujjoha**

Senior Lecturer  
Department of Computer Science and Engineering,  
East West University, Dhaka, Bangladesh

Chairperson

---

**Dr. Ahmed Wasif Reza**

Associate Professor and Chairperson,  
Department of Computer Science and Engineering,  
East West University, Dhaka, Bangladesh

# Acknowledgements

First, we are thankful and expressing our gratefulness to Almighty who offers me divine blessings, patience, mental and psychical strength to complete this thesis. The progression of this thesis could not possibly be carried out without the help of several people who, directly or indirectly, are responsible for the completion of this work. We deeply indebted to our thesis supervisor Mr. Md. Shamsujjoha. His scholarly guidance, especially for his tolerance with our persistent bothers and unfailing support. He gives us the freedom to pursue aspects of reversible fault tolerant computing which we found interesting and compelling. This helped our thesis to achieve its desired goals.

We wish to thank the great people of Department CSE at East West University. A special thank goes to all faculties for their well-disposed instructions and Encouragements.

Finally, we would like to thank our friends and family. Their continued tolerance with our moods and tendency to disappear for weeks at a time gave us a much needed break from the world computing.

# Table of Contents

<b>Abstract</b>	<b>I</b>
<b>Letter of Acceptance</b>	<b>II</b>
<b>Acknowledgements</b>	<b>III</b>
<b>Chapter 1: Introduction</b>	<b>1</b>
1.1. Motivation	1
1.2. Aims and Objectives	3
1.3. Overview	3
1.4. Methodologies of the research	3
1.5. Outline	4
1.6. Summary	4
<b>Chapter 2: Background Study</b>	<b>5</b>
2.1. Gene Regulatory Network	5
2.2. Transcription Factors	8
2.3. Robustness	11
2.4. Summary	13
<b>Chapter 3: Preprocessing of Dataset</b>	<b>14</b>
3.1. Collection of Data	14
3.2. Modification of Data	17
3.3. Summary	17

<b>Chapter 4: Working Procedure</b>	<b>18</b>
4.1. Flow Chart	18
4.2. Scatter Plot	19
4.3. Algorithm for Robustness	22
4.4. Dataset with Robustness	23
<b>Chapter 5: Result Analysis</b>	<b>25</b>
5.1. Artificial Neural Network	25
5.2. Prediction Model	29
<b>Chapter 6: Conclusion</b>	<b>31</b>
6.1. Future Work	31
<b>References</b>	<b>32</b>

# List of Tables

<b>Table 3.1.1:</b> sample transcription factors of genes	<b>15</b>
<b>Table 3.1.2:</b> sample target genes	<b>15</b>
<b>Table 3.1.3:</b> sample gene sequence data	<b>16</b>
<b>Table 3.1.4:</b> sample case dataset	<b>16</b>
<b>Table 3.1.5:</b> sample gene regulatory network in csv file	<b>17</b>
<b>Table 4.4.1:</b> Training Dataset for Neural Network	<b>23</b>



# List of Figures

<b>Fig 4.1.1:</b> working flow chart	<b>18</b>
<b>Fig 4.2.1:</b> Scatter Diagram Creation Using Orang3 Anaconda	<b>19</b>
<b>Fig 4.2.2:</b> scatter diagram sample1 vs sample2	<b>20</b>
<b>Fig 4.2.3:</b> Random Topology Selection	<b>21</b>
<b>Fig 4.3.1:</b> algorithm used for calculating robustness	<b>22</b>
<b>Fig 5.1.1:</b> artificial neural network	<b>25</b>
<b>Fig 5.1.2:</b> Example Set Statistics	<b>26</b>
<b>Fig 5.1.3:</b> Example Set Chart (Histogram)	<b>27</b>
<b>Fig 5.1.4:</b> Improved Neural Network	<b>28</b>
<b>Fig 5.2.1:</b> Example Set Apply Model	<b>29</b>
<b>Fig 5.2.2:</b> Prediction Model	<b>30</b>

# Chapter 1

## Introduction

A gene regulatory network is a collection of molecular regulators which are internally connected with each other. These molecular regulators are DNA, RNA, PROTEIN or complexes of them. The main players of gene regulatory network are transcription factors. A gene regulatory network is created by binding these transcription factors of genes. Gene regulatory network shows the co-association and co-relation among the transcription factors. From the gene regulatory network find out the robustness of the network and topology of the human genomic sequence. In this research we find out the robust network that causes damage to the human body. It will show us which genes are mostly cause diseases and enhance the medical science to prevent them.

### 1.1 Motivation

Gene regulatory network helps to find the factors which cause diseases like cancer. So it plays a vital role in human health research. The most common problem is to understand the regulation of the genes that controls gene expression. In recent years many researchers have proposed different approach to construct gene regulatory network (GRN), e.g. see reviews by Bansal *et al.* (2007) and Markowitz and Spang (2007). These include, among others, approaches that rely on linear models (D'haeseleer *et al.* 1999), information theory (ARACNE) (Margolin *et al.*, 2006), static and dynamic Bayesian networks (BANJO; Yu *et al.*, 2004) and Boolean networks and their probabilistic extensions (Shmulevich *et al.*, 2002). While these methods have been found useful in a

number of applications, they primarily model the data, not the underlying biological process. On the other hand, GRNs could be modeled in great detail with chemical reaction network models. However, there are major difficulties in inference with this modeling approach, e.g. lack of measurements from single cells and computational problems in inferring the model parameters and structure from data (Wilkinson, 2006). The exact models are commonly approximated by ordinary differential equations (ODE), which can be obtained as the expectation of the chemical master equation under certain assumptions, and are often coupled with linear, mass action, sigmoidal, Hill or Michaelis–Menten kinetics. A number of different modeling approaches using ODEs have been proposed, including, among others, estimation of model parameters (Cao and Zhao, 2008), inference for unknown transcription factor (TF) levels (Gao *et al.*, 2008), coupling ODE models with protein complexes (Wang *et al.*, 2007) and model structure inference NIR, TSNI and Inferelator; Bansal *et al.*, 2006; Bonneau *et al.*, 2006; Gardner *et al.*, 2003). Other related methods that combine aspects from ODEs and Bayesian modeling have been proposed, e.g. in Imoto *et al.* (2002), Perrin *et al.* (2003), Nachman *et al.* (2004) and Zou and Conzen (2005).

All ODE-based methods are essentially parametric, such as those proposed in Gardner *et al.* (2003), Perrin *et al.* (2003), Nachman *et al.* (2004), Bansal *et al.* (2006) and Bonneau *et al.* (2006). The work of Gao *et al.* (2008), however, shows a departure from standard parametric approaches in that latent protein activities are modeled using Gaussian processes, although the regulation function has a parametric form. Previously proposed non-parametric approaches, on the other hand, are essentially not based on differential equation-type modeling, such as those in Imoto *et al.* (2002) Yu *et al.* (2004) and Zou and Conzen (2005). Finally, most of the ODE-based approaches make use of frequentist inference (Bansal *et al.*, 2006; Bonneau *et al.*, 2006; Gardner *et al.*, 2003), which as such might have, e.g. the aforementioned problems of making hard decisions (although resampling methods can alleviate that problem).

## **1.2 Aims and objectives**

The objectives of the study are summarized below:

- Creation of gene regulatory network using gene-gene interaction and gene-transcription factor interaction
- Find the co-relation among the transcription factors and their target genes
- Find the robust genes from the gene regulatory network which is created in this research

## **1.3 Overview**

In this study we will show the robustness of the human genes. This study also shows the whole genomic expression of human gene network topology.

## **1.4 Methodologies of the research**

While working on this research, the following important steps are followed:

- First understand about gene regulatory network and approaches to find the gene expression, robustness of a gene regulatory network and transcription factors.
- Visualized the gene regulatory network by using machine learning tools and find the co-association among the transcription factors of the genes.

- Lastly in this research we find out the robust network by using algorithms.

## **1.5 Outline**

In next chapter (chapter 2) briefly discusses about gene regulatory network, transcription factors and robustness.

Chapter 3 discusses about background study about gene regulatory network. The study includes how gene regulatory networks works and the implementation of GRN. This study also includes robustness of genes.

Chapter 4 discusses preprocessing of dataset and the collection of data. In this chapter discusses about the dataset and how it is preprocessed before working with this dataset.

Chapter 5 discusses robustness of the gene regulatory network and result analysis.

Chapter 6 finally discusses conclusion and future work.

## **1.6 Summary**

This chapter demonstrates motivations and objective of this thesis. Then the methodologies of the research that is being followed are discussed here. A brief elementary instructional text of remaining chapters of this thesis has also been described.

## Chapter 2

### Background studies

In this chapter gene regulatory network, transcription factors and robustness is briefly discusses. This chapter helps to understand about gene regulatory network and robustness of a gene regulatory network. Here gene regulation and the importance of gene regulation is briefly discusses. Lastly methods to find the robust genes are discussed here.

#### 2.1 Gene regulatory network

Gene expression networks are networks inferred from microarray time series data and transcription factor networks are networks obtained from a new genome-wide technique that allows an identification of all of the DNA binding sites for each transcription factor (TF). While our knowledge of the transcription factor networks is limited, these networks provide insights into a regulatory core network of TFs that regulate each other, and drive all network interconnectivity. In addition to these global properties, the local properties of these gene expression networks can be used in data mining and classification. High throughput technologies allow a genome-wide interrogation of biological systems. There is a limited literature on transcription factor networks thus far, but early results show intriguing network features for these as well. The global network properties are discussed and it is seen that these inferred networks are scale free and exhibit small world properties. The global properties of such networks show the scale-free distributions of node connectivity indicative of a hierarchical network and also exhibit small world graph properties. The transcription factor networks, on the other hand, are a direct result of experimental observation of a physical association between a TF and a DNA binding site.

While these two networks address the same underlying questions, the regulation of gene expression, they are, by their nature, very different, and represent different manifestations of the underlying regulatory mechanism. To this end, systems-wide investigations have focused on specific functional network structures such as metabolic, signaling and gene regulatory networks. In this chapter we review progress in inferring and interpreting gene expression networks and transcription factor networks.

To understand the mechanism of gene expression, a detailed molecular picture of gene regulatory networks is required. The gene expression network, being inferred from dynamic analysis of time series data of gene expression profiles, must be considered phenomenological, reflecting dynamical observations from the data and an inherently incomplete modeling of this data. These networks are derived from a genome-wide identification of the entire DNA binding sites for each transcription factor (TF). We discuss a network growth model based on gene duplication that provides excellent agreement with the global network parameters derived from the analysis of experimental expression data. While these networks are of limited size, they also appear to show the scale free behavior seen in the expression networks. We conclude with a discussion of how these two networks can be compared and used in concert to create more complete quantitative models of gene regulation. While addressing the same underlying questions, these networks reflect different properties of gene regulation and provide different insights. We discuss how gene regulatory networks are inferred from time series data using simple linear dynamical models.

However, these networks are silent as to the dynamics of the network; they are strictly structural and do not the indication of the extent of control exerted by the TFs. The TF networks, on the other hand, are a direct result of experimental observation of a physical association between a TF and a DNA binding site, which (except for experimental noise) is unique. Two gene regulatory networks inferred from different types of data are considered in this chapter. Two very different types of networks associated with gene expression are considered here. In addition to the global properties, the local properties of

these networks provide powerful data mining tools. The gene expression network is inferred from dynamic analysis of time series data of gene expression profiles. These are networks derived from microarray data through the measurement of time series of mRNA levels on a genome-wide scale. The resulting networks show features that may be universal to biological systems. In both cases, the resulting networks show features that may be universal to biological systems. First, we consider gene expression networks. A network growth model based on gene duplication is described that provides excellent agreement with the global network parameters derived from the experimental data. The two complementary approaches provide a dynamical, but incomplete, phenomenological model of the structure of the network and a precise structural model with unknown dynamical properties. The molecular circuitry not only reveals how the expression of individual genes is controlled but also how one effectors or agent influences the entire network. These are not true, gene regulatory networks in the strict sense because they are correlative, and not necessarily causal, networks. No line of inference is required to generate these networks and they lead to direct mechanistic interpretation. This network allows the direct interaction and control of each gene to be identified from the observation of TF binding at elements upstream from DNA coding regions.

The second network to be discussed is the transcription factor network. These networks describe how the mRNA level of one gene influences the level of another. Such a picture can be developed by probing the interactions between signaling pathways, transcription factors and the TF binding sites of the cis regulatory region of a gene. These interactions constitute the molecular circuitry that describes how external influences can trigger signal transduction pathways to activate transcription factors for a specific set of genes. Thus, systems biology invariably means network analysis. These relationships are most easily represented by network structures or graphs. The current era of systems biology is marked by ongoing efforts to assimilate and integrate this avalanche of information into models of biological functions. To do this, the detailed information about molecular species cannot be considered in isolation but rather must be related to all of the other components of the system. These technologies permit the measurement of the many parameters and variables associated with life processes and reveal, in many cases, the



inherent complexities of these processes. An emerging problem in bioinformatics is to identify the relationships between the various components of a system and infer how one component influences another.

## **2.2 Transcription factors**

In addition to controlling the genes and transcription of other transcription factors, these protein complexes can also control the genes responsible for their own transcription, leading to complex feedback control mechanisms. The basal transcription factors increase the rate of transcription for all genes; indeed, RNA polymerase cannot bind to the promoter without them. An average gene may have several dozen specific factors involved in its regulation, giving the potential for very precise control of its expression. Gene-specific factors are known as activators or repressors, depending on whether they increase or decrease the rate of transcription. Many gene-specific factors bind to the promoter outside of the TATA box, especially near the transcription initiation site, the beginning of the DNA sequence that is actually read by RNA polymerase. Transcription factors are a very diverse family of proteins and generally function in multi-subunit protein complexes. For example, homeotic genes control the pattern of body formation, and these genes encode transcription factors that direct cells to form various parts of the body.

Transcription factors function in the nucleus, where genes are found, and nuclear transport (i.e., import or export) of transcription factors can influence their activity. Differential control of gene transcription is facilitated by gene-specific transcription factors. Transcription factors can have important roles in cancer, if they influence the activity of genes involved in the cell cycle (or cell division cycle). Basal, or general, transcription factors are necessary for RNA polymerase to function at a site of transcription in eukaryotes. Another important general mechanism controlling the activity

of transcription factors is posttranslational modification such as phosphorylation. In addition, transcription factors can be the products of oncogenes (genes that are capable of causing cancer) or tumour suppressor genes (genes that keep cancer in check). If a mutation occurs in any of the homeotic transcription factors, an organism will not develop correctly. Transcription factors are vital for the normal development of an organism, as well as for routine cellular functions and response to disease.

Transcription factors control when, where, and how efficiently RNA polymerases function. The binding sites of transcription factors can be determined by "DNA footprinting." Gene-specific factors work in a variety of ways. During development of multicellular organisms, transcription factors are responsible for dictating the fate of individual cells. Transcription factors can activate or repress the transcription of a gene, which is generally a key determinant in whether the gene functions at a given time. Transcription factors are a common way in which cells respond to extracellular information, such as environmental stimuli and signals from other cells. By interacting directly with DNA, transcription factors can open up otherwise inaccessible regions. They are considered the most basic set of proteins needed to activate gene transcription, and they include a number of proteins, such as TFIIA (transcription factor II A) and TFIIB (transcription factor II B), among others. Transcription factor, molecule that controls the activity of a gene by determining whether the gene's DNA (deoxyribonucleic acid) is transcribed into RNA (ribonucleic acid).

For example, in fruit flies (*Drosophila*), mutation of a particular homeotic gene results in altered transcription, leading to the growth of legs on the head instead of antenna; this is known as the antennapedia mutation. Growth factors and homeotic proteins also act as gene-specific factors or form complexes that do. Substantial progress has been made in defining the roles played by each of the proteins that compose the basal transcription factor complex. Some influence RNA polymerase's rate of escape from the promoter, or its return to it for another round of transcription. A hormone is not a transcription factor itself but binds to a receptor to form a gene-specific factor. The number of known gene-specific factors is currently in the low thousands and inevitably will grow as the genome

becomes better known. This arrangement keeps the DNA well ordered but also decreases its accessibility for transcription. Some interact with the basal factors, altering the rate at which they bind to the promoter.

The factors that bind to them come from elsewhere in the genome and are called "trans" acting factors. Some factors physically alter the local structure of the DNA, making it more or less accessible. A homoerotic protein can activate one gene but repress another, producing effects that are complementary and necessary for the ordered development of an organism. In eukaryotic organisms, DNA is wound around protein complexes called histones and is further looped, coiled, and condensed to allow efficient packing in the cell nucleus. Since enhancer and silencer sites are on the same DNA sequence as the gene they control, they are called "cis" regulatory elements (from the Latin word for "side").

They may bind directly to special "promoter" regions of DNA, which lie upstream of the coding region in a gene, or directly to the RNA polymerase molecule. The DNA sequences that activators bind to are called enhancer sites; repressors bind to silencer sites. The enzyme RNA polymerase catalyzes the chemical reactions that synthesize RNA, using the gene's DNA as a template. Because of the looped structure of DNA, these sequences are physically close to the promoter, despite being far away along the double helix. Some bind to DNA sequences hundreds or thousands of nucleotides away from the promoter. Others bind to sequences within the coding region of the gene, or downstream from it at the termination region. Hormones are an important class of molecules that regulate gene expression. Red blood cells should make lots of hemoglobin but not the digestive enzyme pepsin, while stomach lining cells should do the opposite. Once bound together, the hormone-receptor complex binds to DNA. However, not all genes should be transcribed at an equal rate all the time.

## 2.3 Robustness

Together with synthetic network biology, such studies are starting to provide insights into the transcriptional mechanisms that cause robust versus stochastic gene expression and their relationships to phenotypic robustness and variability. The computational modeling and analysis of GRNs, together with the field of synthetic biology, have provided numerous insights into the importance of network architecture and topology in generating differential gene expression and phenotypic outputs. In the last decade or so, the field of systems biology has extensively studied the mechanisms of differential gene expression at the level of gene regulatory networks (GRNs).

Biological robustness and stochasticity can be controlled, at least in part, at the level of differential gene expression. We also observed the existence of multiple genotypes giving rise to the same phenotype in accordance with the theoretical view that natural selection operates on phenotypes thereby accommodating variation in the genotype by fixing those changes that are phenotype-neutral. In any given cell, thousands of genes are expressed and work in concert to ensure the cell's function, fitness, and survival. Gene regulatory networks (GRNs) involving interactions between large numbers of genes and their regulators have been mapped onto graphic diagrams that are used to visualize the regulatory relationships. Recent advances have enabled the analysis of differential gene expression at a systems level.

Here, we describe examples of robustness and stochasticity at the organismal or cellular level, as well as at the gene expression level. For instance, developmental gene expression is extremely similar in a given cell type from one individual to another. We have developed a framework to analyze the effect of objective functions, input types and starting populations on the evolution of GRNs with a specific emphasis on the robustness of evolved GRNs. Although much work has been done in elucidating the transcriptional regulatory network, the underlying mechanisms that have possibly influenced the

evolution of these GRNs are still debatable. This study gives a proof-of-concept of the fact that robustness is an emergent property of GRNs as well as of the degeneracy of the network topology/function relationship analogous to the sequence/structure problem in proteins. We observed that robustness evolves along with the networks as an emergent property even in the absence of specific selective pressure towards more robust systems. The expression of other genes is more variable: Their levels are noisy and are different from cell to cell and from individual to individual. Gene Regulatory Networks (GRNs) have become a major focus of interest in recent years. Here, we discuss GRNs and their topological properties in relation to transcriptional and phenotypic outputs in development and organismal physiology. The further characterization of GRNs has already uncovered global principles of gene regulation. The regulation and expression of some genes are highly robust; their expression is controlled by invariable expression programs. Each gene, in turn, must be expressed at the proper time and in the proper amounts to ensure the appropriate functional outcome.

In addition, robustness was independent of the selective pressure, input types or the initial starting populations. Technological advances in high-throughput molecular biology have enabled the characterization of large sets of genes and their regulators. We discuss the GRN principles and mechanisms that generate these different types of biological outputs. However, responses to stress can be more stochastic, thereby providing a population of cells or organisms with different outputs to adapt or survive under adverse conditions. Biological processes can be deterministic and robust, or more stochastic and variable. For example, in development and differentiation, little deviation is tolerated. This can be highly beneficial in physiological responses to outside cues and stresses.

## **2.4 Summary**

In this chapter we have discussed about gene regulatory network, transcription factor and robustness. This study helps to understand about gene regulatory network.

## **Chapter 3**

### **Preprocessing of dataset**

Preprocessing data means before using any dataset make that dataset suitable for the research. In this research we need numeric dataset of genome but our dataset was in polynomial format. So, we need to preprocess for using dataset.

#### **3.1 Data collection**

In order to create a gene regulatory network first we need corresponding transcription factor of gene. We collect our data from website and combined them in a csv file. We collect various types of genes from different website and find the corresponding transcription factor for all genes. Then preprocess all the data and use that data for creating gene regulatory network and find the robust genes as well. We collect data from <http://bioinfo.icgeb.res.in/APA>. In this dataset there are 47721 number of genes, 1024 number of transcription factors, 9993 number of target genes, 22407 number of gene regulatory network, 14360 number of gene sequences.

<b>no</b>	<b>TF</b>
1	10002
2	10009
3	10014
4	10113
5	10127
6	10153
7	10168
8	10194
9	100062
10	10173

Table 3.1.1: sample transcription factors of genes

From this transcription factors and target genes we find gene sequences. Sample target genes and gene sequences are given below:

<b>No</b>	<b>TG</b>
1	100
2	1000
3	10000
4	10001
5	10002
6	1E+08

Table 3.1.2: sample target genes



ENTREZID	SYMBOL
1	A1BG
10	NAT2
100	ADA
1000	CDH2
10000	AKT3
10001	MED6
10002	NR2E3

Table 3.1.3: sample gene sequence data

From the gene sequence and target genes we create the gene regulatory network from applying case data and training dataset. Gene regulatory network I csv file are given below:

ENTREZID	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
10	3.41684	4.255501	3.732269	3.732269	1.646163	2.469886
100	6.375387	6.536675	12.1454	12.1454	9.038946	7.04756
1000	11.17589	11.23057	9.384007	9.384007	5.120186	11.24181
10000	5.684258	6.002928	6.076602	6.076602	3.662206	3.204767
1E+08	5.029011	5.421223	5.863443	5.863443	3.861955	5.656782
10001	6.211986	5.64501	6.524816	6.524816	5.76288	5.814807

Table 3.1.4: sample case dataset

TF	TG	type
10002	1025	TF-TG
10002	10514	TF-TG
10002	10607	TF-TG
10002	2099	TF-TF
10002	22907	TF-TG
10002	25942	TF-TG
10002	3065	TF-TG

Table 3.1.5: sample gene regulatory network in csv file

## 3.2 Modifying dataset

For creating dataset we need numeric numbers but in our dataset genes are in polynomial data. So first we need to modify it. We replace the gene name into their genomic id and then create gene regulatory network.

## 3.3 Summary

In this chapter discusses about dataset and the preprocessing of dataset and reason behind the change. In this study it clear that for making gene regulatory network genes are must be in numeric format and transcription factor and target genes should be there.

# Chapter 4

## Working procedure

In this chapter we will discuss about the flow chart to describe how the work is done in this research. In this study some steps are followed to make the work easier.

### 4.1 Flowchart

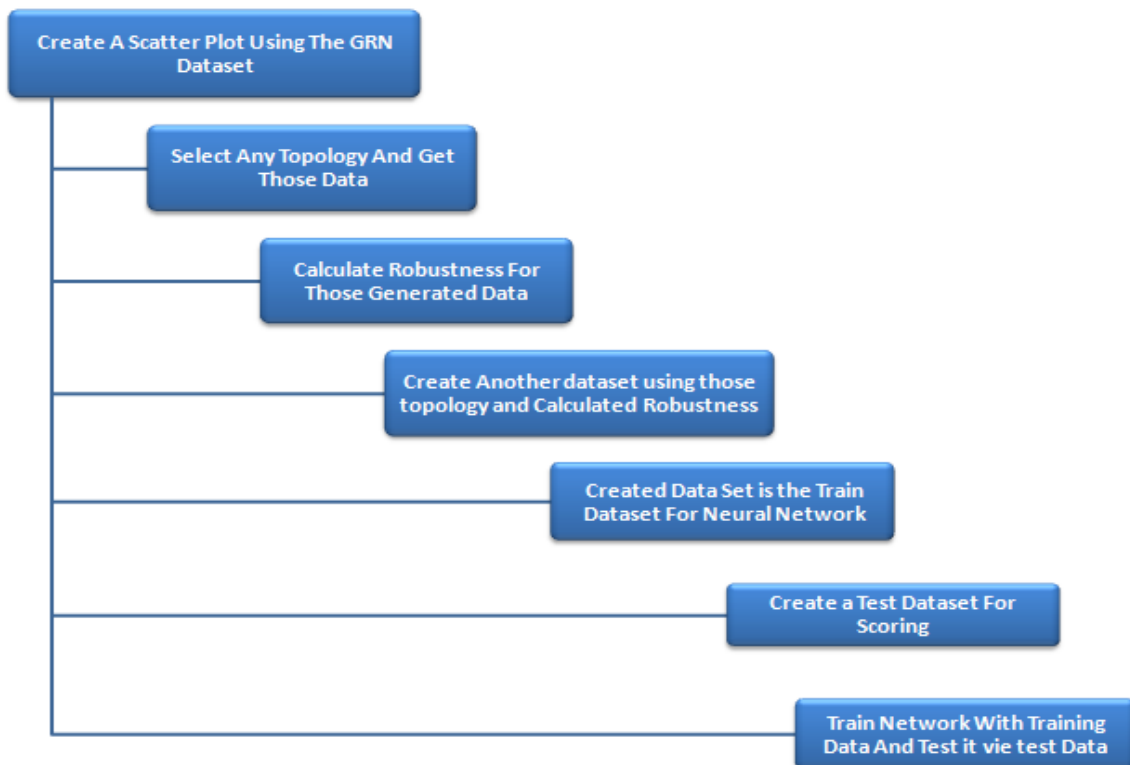


Fig 4.1.1: working flow chart

## 4.2 Scatter diagram

Working with neural network is challenging. To use neural network it requires numeric value for all attributes. We have discussed about data preparation in previous chapter. Using Orange3 tool available in Anaconda we can show the scatter plot and data selection flow chart. Three parts here one is file selection then primary data visualization secondly make a scatter plot then selecting data randomly.

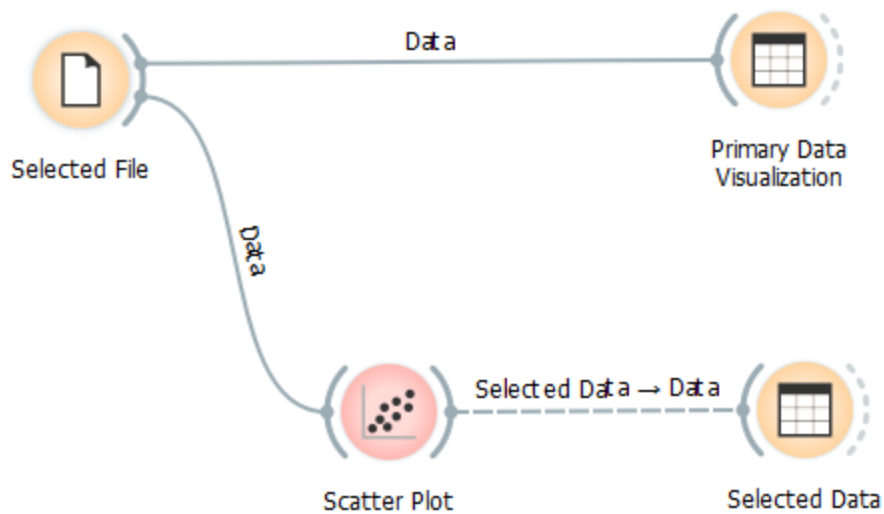


Fig 4.2.1: Scatter Diagram Creation Using Orange3 Anaconda

We used a Gene Regulatory Network Dataset and Orange3 tool available in Anaconda to create a scatter plot. From the Scatter plot we randomly selected a topology and marked those data.

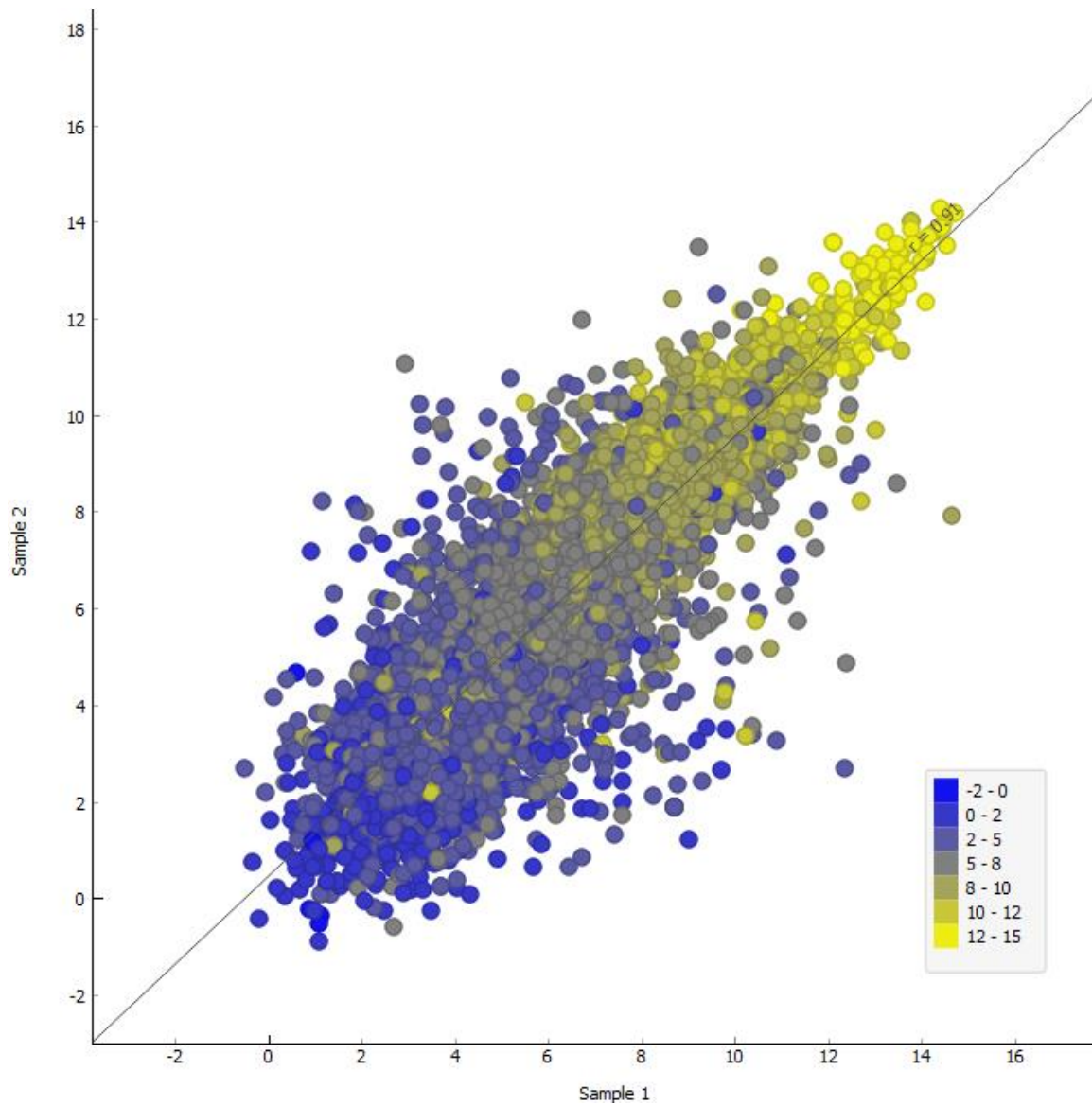


Fig 4.2.2: scatter diagram sample1 vs sample2

Scatter Plot Represents Data samples used in GRN dataset. Here we can see some human gene interactions.

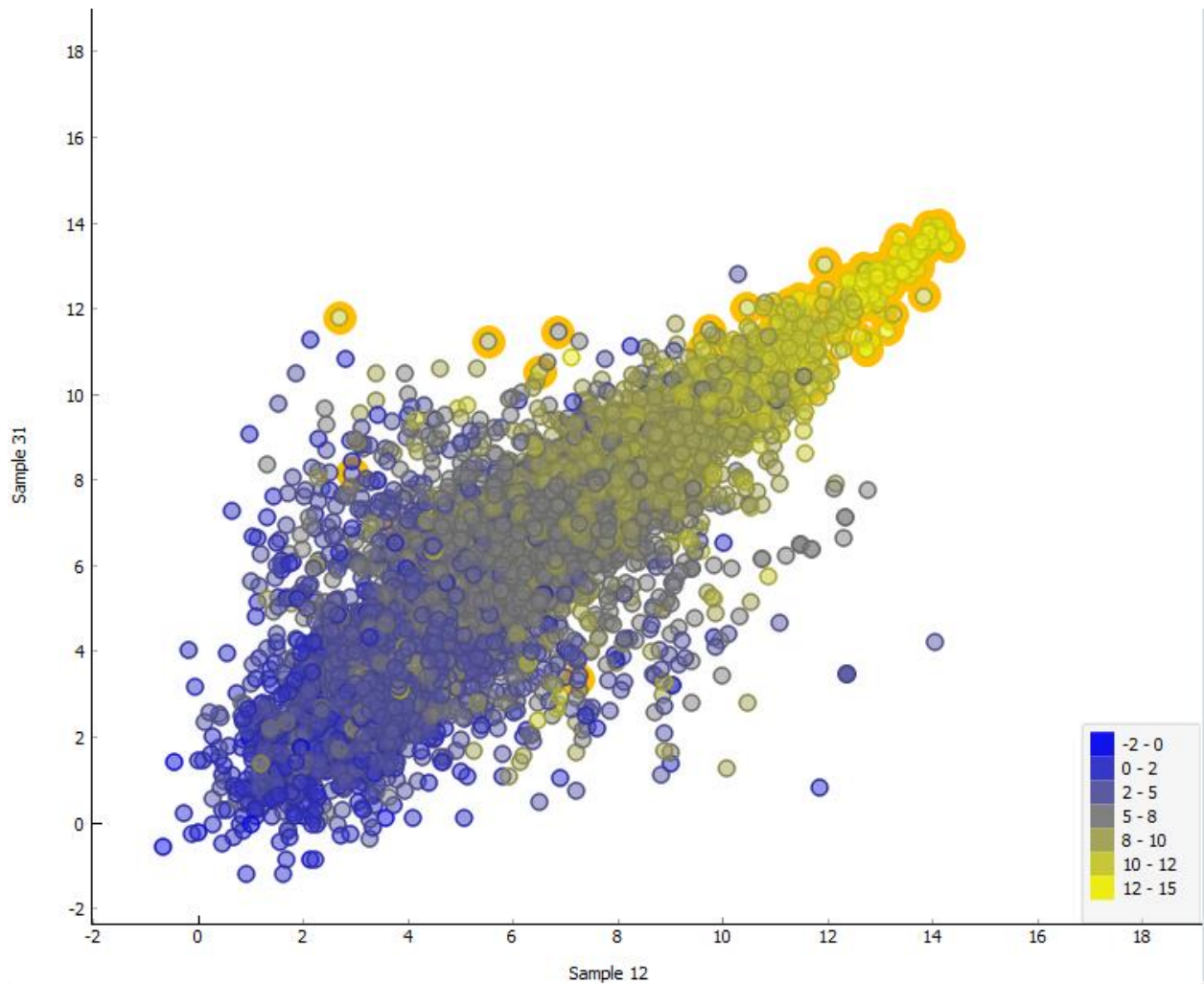


Fig 4.2.3: Random Topology Selection

We have randomly selected some data from the scatter plot. This will show the description of any topology in the scatter plot.

## 4.3 Algorithm for robustness

```
In [21]: import sys
import os

In [22]: path = r"C:\Users\Trimatrik\Desktop\Thesis\robustness-calcs-jupiternotebook\robustness-calcs-master\robustness-calcs\robustness-c
sys.path.append(path + '\metrics')

In [23]: def calc_all(random_topology, threshold, maximise=True):

    class robustnessMetric:
        def __init__(self, name, robustness):
            self.name = name
            self.robustness = robustness

    robustness = []
    robustness.append(robustnessMetric("RobustTopology", Net_Val(performance, maximise)))

    return robustness

In [24]: def Net_Val(random_topology, maximise=True):
    if maximise:
        robustness = [max(solution) for solution in random_topology]
    else:
        robustness = [min(solution) for solution in random_topology]

    return robustness

In [25]: if __name__ == "__main__":
    random_topology = [[0.941,0.938, 0.911, 0.908], [0.893, 0.888, 0.899, 0.881]]
    robustness = calc_all(random_topology, 0.717)
    for metric in robustness:
        print(metric.name, metric.robustness)

RobustTopology [0.941, 0.899]
```

Fig 4.3.1: algorithm used for calculating robustness

Evolutionary Algorithm Used to find the robustness of Created Network Topology. In this algorithm we have passed 2 parameter Network Topologies and Approximate Threshold value. Network topology is a 2D matrix. And Threshold value confirms the network creation is failed or passed. Threshold is to be between 0-1 we used 0.95.

In order to create higher quality solutions (offspring), crossover and mutation operators are applied on individuals (parents) selected from current generation based on some criteria.

Higher quality offspring replaces individuals from the current generation thus creating a new generation of population.

Subscript can take one of the values from  $\{1, -1, 0\}$  where 1 represents activation,  $-1$  represents repression and 0 represents no interaction.

Even with the Monte Carlo method the fitness estimation becomes very expensive as we need to simulate 10,000 network's behavior to quantify the robustness of each topology. In order to accelerate the robustness measurement procedure we utilized a fitness approximation technique in our algorithm.

If the fitness of a topology is reevaluated then we update the archive with the new robustness score for that topology if the score is higher than the stored value.

## 4.4 Dataset with robustness

In this study a new dataset is created with robustness found from the gene regulatory network. From this dataset creates training dataset and test dataset.

<b>TF</b>	<b>TG</b>	<b>Robustness</b>
6.995824	100101629	0.04
8.700682	100129361	84.51
7.84089	100130418	99.16
11.08241	100133941	93.97
6.907371	100271849	3.59
6.659496	100101629	78.73
8.862544	100129361	99.87

Table 4.4.1: Training Dataset for Neural Network



In previous we prepared a dataset of topologies and their corresponding robustness .We will use that data as our training data. Tool we have used has limit of 2000 data that's why we are taking dataset of 30 elements as our training model.

Test dataset limit to 5 elements and can be prepared by human gene-gene interaction or gene-transcription factor interaction.

# Chapter 5

## Result analysis

In this chapter we will discuss about the robustness of gene regulatory network, creating an artificial neural network, show prediction model and linear regression using machine learning approach.

### 5.1 Artificial neural network

An artificial neural network is created by using rapid miner tools. Here is the procedure how an artificial neural network is created by rapid miner.

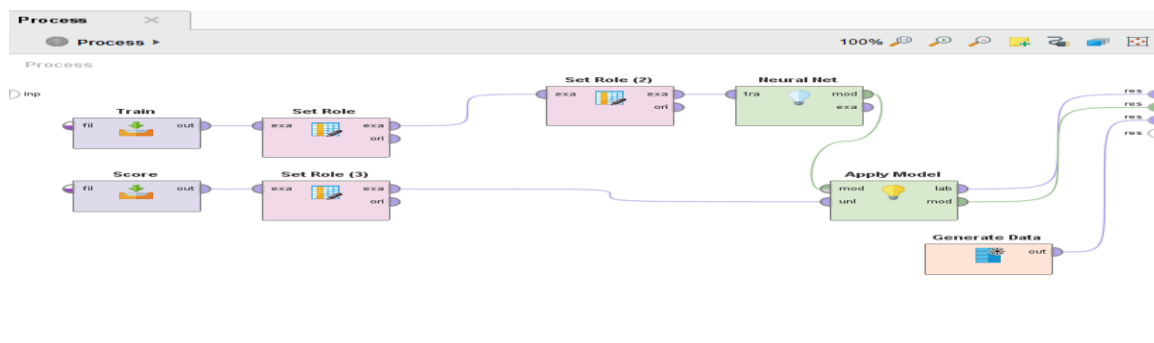


Fig 5.1.1: artificial neural network

Name	Type	Missing	Statistics		Filter (6 / 6 attributes): <input type="text" value="Search for Attributes"/>
Label label	Real	0	Min -434.610	Max 615.730	Average 5.321
att1	Real	0	Min -9.939	Max 9.798	Average 0.191
att2	Real	0	Min -9.661	Max 9.795	Average -0.016
att3	Real	0	Min -9.837	Max 9.612	Average -1.171
att4	Real	0	Min -9.968	Max 9.973	Average 0.343
att5	Real	0	Min -9.920	Max 9.843	Average 0.769

Fig 5.1.2: Example Set Statistics

This is the Example set of generated data we can use this data and visualize it in many ways. It is a complete relationship between dataset attributes. And no missing value found so our dataset is clean.

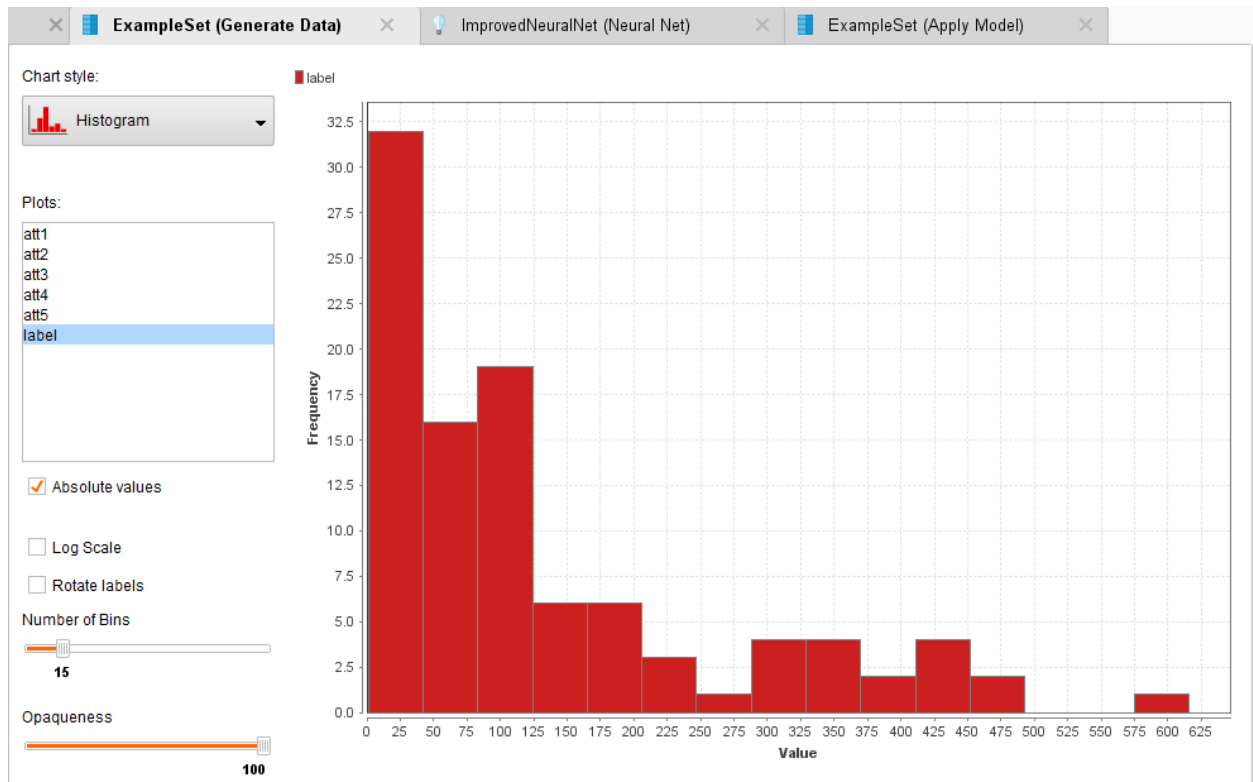


Fig 5.1.3: Example Set Chart (Histogram)

Label means the type (TF-TF or TG-TF) in GRN .We can show it absolutes values in a histogram to find their frequencies.

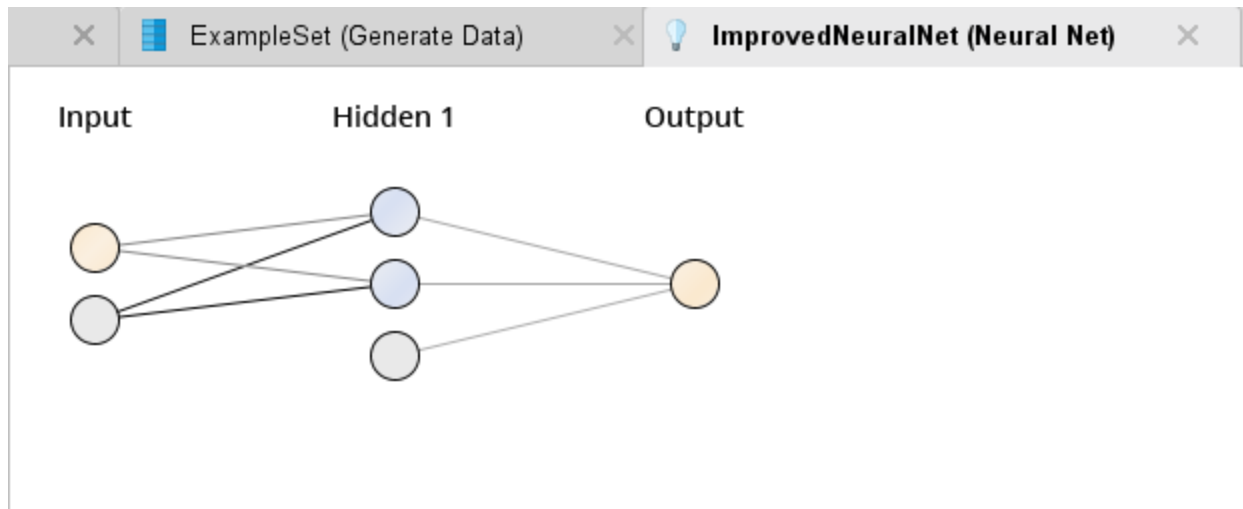


Fig 5.1.4: Improved Neural Net

This is the complete visual of neural network we have created. We can calculate the threshold value from here and see if the network creation fails of pass.

Hidden 1

=====

Node 1 (Sigmoid)

-----

B: 1.067

Bias: -2.160

Node 2 (Sigmoid)

-----

B: 1.046

Bias: -2.210

Output

=====

Regression (Linear)

-----

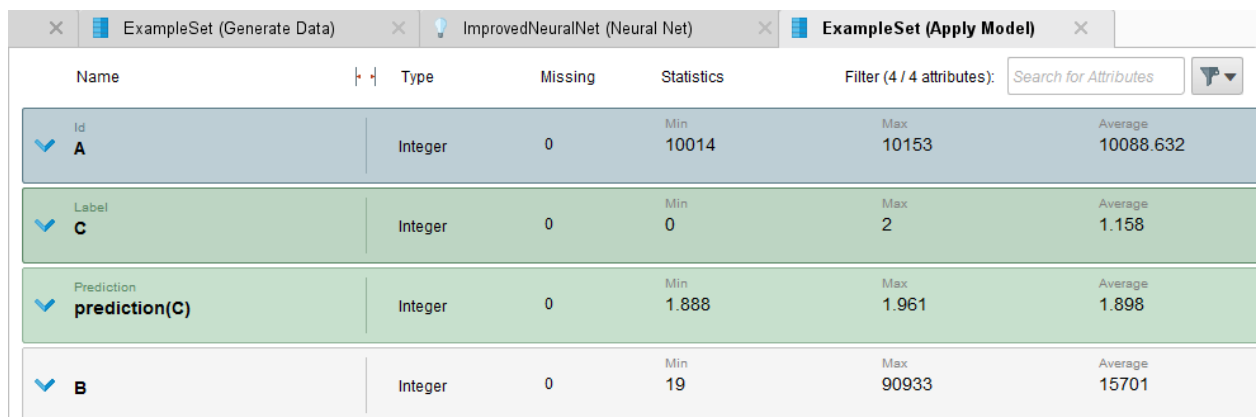
Node 1: 0.766

Node 2: 0.767

Threshold: 0.717

## 5.2 Prediction model

In this study a prediction model is created and shows the prediction values and can find the error rating from the prediction model.



Name	Type	Missing	Statistics	Filter (4 / 4 attributes):
Id A	Integer	0	Min 10014 Max 10153 Average 10088.632	<input type="text" value="Search for Attributes"/>
Label C	Integer	0	Min 0 Max 2 Average 1.158	
Prediction prediction(C)	Integer	0	Min 1.888 Max 1.961 Average 1.898	
B	Integer	0	Min 19 Max 90933 Average 15701	

Fig 5.2.1: Example Set Apply Model

After applying neural net this is our predicted model. In predicted result there are also no missing values. We have maximum value for A and minimum for C. So our prediction will be over C.

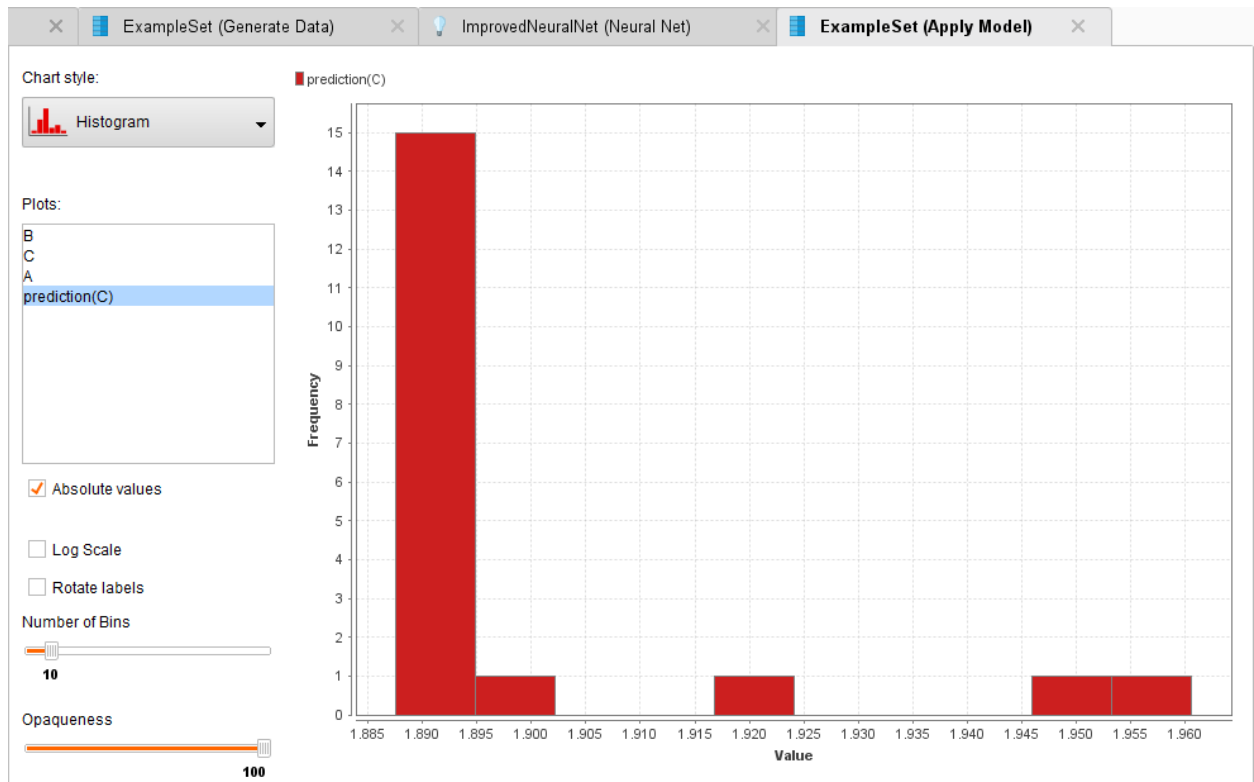


Fig 5.2.2: Prediction Model

Prediction model histogram with absolute value. We can find frequencies for each value here. We can calculate error rating in different cases.

## **Chapter 6**

### **Conclusion**

Finally, through this research we tried to combine machine learning with evolutionary algorithm and we did it successfully. There was a challenge to find out the perfect dataset for this research but we overcome that anyhow. In future researcher's can use those dataset for their research purposes. This dataset and research will be helpful in genetic analysis software in the days ahead. By using our train dataset they can train the neural network and test it through test data created in genetic research lab. In the field of cancer or any kind of disease detection this research works will be helpful.

### **6.1 Future works**

This research will enhance some future works like cancer gene detection and disease gene detection. Artificial neural network to indentify the topology of robust gene regulatory network can be implemented through code but that was not our approach. We tried to proof that this method is applicable. So we used some open source software's to fire up our research. If this method is using for software development then using raw code instead of using tool is mandatory.



## References

1. Wagner A (2007) *Robustness and Evolvability in Living Systems*. Princeton University Press, 1st ed edition.
2. Hilgers V, Bushati N, Cohen SM (2010) *Drosophila* micromRNAs 263a/b confer robustness during development by protecting nascent sense organs from apoptosis. *PLoS Biol* 8: e1000396. pmid:20563308
3. Zheng L, Sepúlveda LA, Lua RC, Lichtarge O, Golding I, et al. (2013) The maternal-to-zygotic transition targets actin to promote robustness during morphogenesis. *PLoS Genet* 9: e1003901. pmid:24244181
4. Graudenzi A, Serra R, Villani M, Colacci A, Kauffman SA (2011) Robustness analysis of a boolean model of gene regulatory network with memory. *Journal of Computational Biology* 18: 559–577. pmid:21417939
5. Holme P (2011) Metabolic robustness and network modularity: A model study. *PLoS ONE* 6: e16605. pmid:21311770
6. van Dijk AD, van Mourik S, van Ham RC (2012) Mutational robustness of gene regulatory networks. *PLoS ONE* 7: e30591. pmid:22295094
7. Hsiao TL, Vitkup D (2008) Role of duplicate genes in robustness against deleterious human mutations. *PLoS Genet* 4: e1000014. pmid:18369440
8. MacNeil LT, Walhout AJ (2011) Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res* 21: 645–657. pmid:21324878
9. Plata G, Vitkup D (2013) Genetic robustness and functional evolution of gene duplicates. *Nucleic Acids Research*. pmid:24288370
10. Kitano H, Oda K (2006) Robustness trade-offs and host-microbial symbiosis in the immune system. *Molecular Systems Biology* 2: msb4100039-E1–msb4100039-E10. pmid:16738567
11. Qiu Y, Zeltzer S, Zhang Y, Wang F, Chen GH, et al. (2012) Early induction of *ccl7* downstream of *tlr9* signaling promotes the development of robust immunity to cryptococcal infection. *J Immunol* 188: 3940–3948. pmid:22422883
12. Staniczenko PP, Lewis OT, Jones NS, Reed-Tsochas F (2010) Structural dynamics and robustness of food webs. *Ecology Letters* 13: pages 891–899. pmid:20482578
13. Evans DM, Poccock MJ, Memmott J (2013) The robustness of a network of ecological networks to habitat loss. *Ecology Letters* 16: 844–852. pmid:23692559
14. Tian T, Olson S, Whitacre JM, Harding A (2011) The origins of cancer robustness and evolvability. *Integr Biol* 3: 17–30. pmid:20944865

- 15.**Masuda M, Toh S, Wakasaki T, Suzui M, Joe AK (2013) Somatic evolution of head and neck cancer—biological robustness and latent vulnerability. *Mol Oncol* 7: 14–28. pmid:23168041
- 16.**Kitano H (2004) Biological robustness. *Nat Rev Genet* 5: 826–837. pmid:15520792
- 17.**Wagner A (2005) Circuit topology and the evolution of robustness in two-gene circadian oscillators. *PNAS* 102: 11775–11780. pmid:16087882
- 18.**Benítez M, Alvarez-Buylla ER (2010) Dynamic-module redundancy confers robustness to the gene regulatory network involved in hair patterning of arabidopsis epidermis. *Biosystems* 102: 11–15. pmid:20655358
- 19.**Kitano H (2007) Towards a theory of biological robustness. *Molecular Systems Biology* 3. pmid:17882156
- 20.**Dhaeseleer P, Lian S, Somojai R (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics (Oxford, England)* 16: 707–726. pmid:11099257
- 21.**Ciliberti S, Martin OC, Wagner A (2007) Robustness can evolve gradually in complex regulatory gene networks with varying topology. *PLoS Comput Biol* 3: e15. pmid:17274682
- 22.**Karlsson M, Weber W (2012) Therapeutic synthetic gene networks. *Current Opinion in Biotechnology* 23: 703–711. pmid:22305476
- 23.**Kwok R (2010) Five hard truths for synthetic biology. *Nature* 463: 288–290. pmid:20090726
- 24.**Lu TK, Khalil AS, Collins JJ (2009) Next-generation synthetic gene networks. *Nature biotechnology* 27: 1139–1150. pmid:20010597
- 25.**Wagner A (2008) Robustness and evolvability: a paradox resolved. *Proceedings Biological sciences / The Royal Society* 275: 91–100. pmid:17971325
- 26.**Hache H, Lehrach H, Herwig R (2009) Reverse engineering of gene regulatory networks: a comparative study. *EURASIP Journal on Bioinformatics and Systems Biology* 2009: 8. pmid:19551137
- 27.**Jostins L, Jaeger J (2010) Reverse engineering a gene network using an asynchronous parallel evolution strategy. *BMC systems biology* 4: 17. pmid:20196855
- 28.**Francois P, Hakim V (2004) Design of genetic networks with specified functions by evolution in silico. *Proceedings of the National Academy of Sciences of the United States of America* 101: 580–585. pmid:14704282
- 29.**Drennan B, Beer RD (2006) Evolution of repressilators using a biologically-motivated model of gene expression. In: *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*. Citeseer, pp. 22–27.
- 30.**Paladugu S, Chickarmane V, Deckard A, Frumkin J, McCormack M, et al. (2006) In silico evolution of functional modules in biochemical networks. *IEE Proceedings-Systems Biology* 153: 223–235. pmid:16986624

- 31.**Cao H, Romero-Campero FJ, Heeb S, Cámara M, Krasnogor N (2010) Evolving cell models for systems and synthetic biology. *Systems and synthetic biology* 4: 55–84. pmid:20186253
- 32.**Noman N, Palafox L, Iba H (2013) Evolving genetic networks for synthetic biology. *New Generation Computing* 31: 71–88.
- 33.**Rizk A, Batt G, Fages F, Soliman S (2009) A general computational method for robustness analysis with applications to synthetic gene networks. *Bioinformatics* 25: i169–i178. pmid:19477984
- 34.**Donzé A, Fanchon E, Gattepaille LM, Maler O, Tracqui P (2011) Robustness analysis and behavior discrimination in enzymatic reaction networks. *PLoS one* 6: e24246. pmid:21980344
- 35.**Siegal ML, Beraman A (2002) Waddington's canalization revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences* 99: 10528–10532. pmid:12082173
- 36.**Leclerc RD (2008) Survival of the sparsest: robust gene networks are parsimonious. *Molecular systems biology* 4. pmid:18682703
- 37.**Gonze D, Halloy J, Goldbeter A (2002) Robustness of circadian rhythms with respect to molecular noise. *Proceedings of the National Academy of Sciences* 99: 673–678. pmid:11792856
- 38.**Atkinson MR, Savageau MA, Myers JT, Ninfa AJ (2003) Development of genetic circuitry exhibiting toggle switch or oscillatory behavior in *Escherichia coli*. *Cell* 113: 597–607. pmid:12787501
- 39.**Conrad E, Mayo AE, Ninfa AJ, Forger DB (2008) Rate constants rather than biochemical mechanism determine behaviour of genetic clocks. *Journal of The Royal Society Interface* 5: S9–S15. pmid:18426770
- 40.**Novák B, Tyson JJ (2008) Design principles of biochemical oscillators. *Nature reviews Molecular cell biology* 9: 981–991. pmid:18971947
- 41.**Montagne K, Plasson R, Sakai Y, Fujii T, Rondelez Y (2011) Programming an in vitro dna oscillator using a molecular networking strategy. *Molecular systems biology* 7. pmid:21283142
- 42.**Bar-Or RL, Maya R, Segel LA, Alon U, Levine AJ, et al. (2000) Generation of oscillations by the p53-mdm2 feedback loop: a theoretical and experimental study. *Proceedings of the National Academy of Sciences* 97: 11250–11255. pmid:11016968
- 43.**Nelson D, Ihekwaba A, Elliott M, Johnson J, Gibney C, et al. (2004) Oscillations in nf-kb signaling control the dynamics of gene expression. *Science* 306: 704–708. pmid:15499023
- 44.**Baggs JE, Price TS, DiTacchio L, Panda S, FitzGerald GA, et al. (2009) Network features of the mammalian circadian clock. *PLoS biology* 7: e1000052. pmid:19278294
- 45.**Partch CL, Green CB, Takahashi JS (2014) Molecular architecture of the mammalian circadian clock. *Trends in cell biology* 24: 90–99. pmid:23916625
- 46.**Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403: 335–338. pmid:10659856
- 47.**Wilkins AK, Tidor B, White J, Barton PI (2009) Sensitivity analysis for oscillating dynamical systems. *SIAM Journal on Scientific Computing* 31: 2706–2732. pmid:23296349

- 48.**Gunawan R, Doyle III FJ (2006) Isochron-based phase response analysis of circadian rhythms. *Biophysical journal* 91: 2131–2141. pmid:16815910
- 49.**Ferrell Jr JE, Xiong W (2001) Bistability in cell signaling: How to make continuous processes discontinuous, and reversible processes irreversible. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 11: 227–236. pmid:12779456
- 50.**Gardner TS, Cantor CR, Collins JJ (2000) Construction of a genetic toggle switch in *escherichia coli*. *Nature* 403: 339–342. pmid:10659857
- 51.**Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* 19: ii122–ii129. pmid:14534181
- 52.**Zhu X, Huang Y, Doyle J (1996) Soft vs. hard bounds in probabilistic robustness analysis. In: *Decision and Control, 1996., Proceedings of the 35th IEEE Conference on. IEEE*, volume 3, pp. 3412–3417.
- 53.**Tenne Y, Goh CK (2010) *Computational Intelligence in Expensive Optimization Problems*, volume 2. Springer. <https://doi.org/10.1007/978-3-642-10701-6>