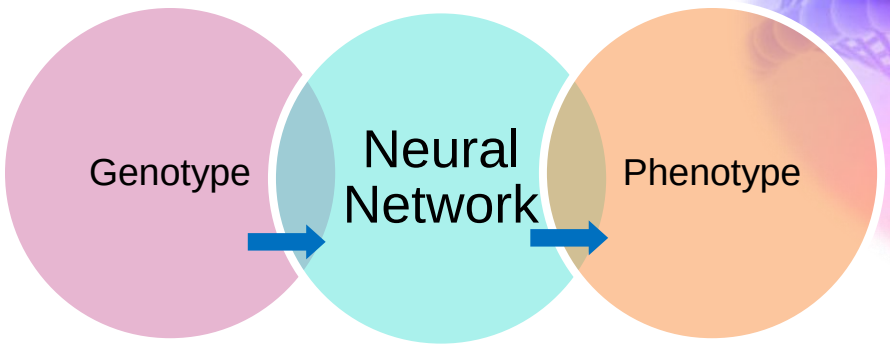
PhenoGene

[Graduation Project]

By

Rehab Sabry Ahmed 20080403
Salma Shehab Raouf 20090152
Computer Science Department
Prof. Hesham Hassan

TABLE OF CONTENTS



CHAPTER 1|INTRODUCTION

1.1 ABSTRACT

MOTIVATION

Prediction of complex phenotypes using genotypes is the central concept and recent interest in the advocated fields of research.

One of the implications was the public wonder if it's actually possible to predict a human phenotypic trait given only his genome DNA sequence/ genotype.

APPROACH

We're to use neural networks. The network is trained on a given database of DNA SNPs variants of a set of individuals having a specific trait described. Then the trained network is used to predict the individuals' trait.

We've specifically tested the eye color trait, analyzing a database grading a number of SNPs data markers associated to eye color variation.

PRODUCT

A computer-software that will be able to:

- Predict a human's phenotypic trait having been provided by a database of SNPs variations associated specifically to that trait.
- Study the effect of each gene variation related to that trait, giving a representation of its contribution or role in the appearance of such trait, and its effect on validation of such prediction.

The application is extendible for various human phenotypes.

1.2 INTRODUCTION

Understanding the human's genome and DNA sequencing and protein structure for example is a complex process in Biology, requiring the use of more complex machinery.

Analysis and interpretation of this data has been researchers' main concern, here comes the part of Computational Biology within bioinformatics where new software and tools are to be developed not only for the storage and querying data, but also for managing and applying algorithms and techniques for finding out new relations and further knowledge about it.

GWA studies will be our project's main concern, the project's willing to cover a critical issue, and we are to test several genetic variations (SNPs) for significance in predicting the human's eye color specifically for further aiding and conformation to them.

In Previous Genome Wide Association Studies, researchers have examined different genetic components of individuals and have been able to connect between one's genotype and phenotype, stating that the human eye color morphology has a strong genetic component.
This fact is of great importance, it has been their key to find out the relationship between genome encoded in DNA and hereditary information based on it.

Chapter 1, as you can see, is to give you an introduction about our project's idea and a decent summing up to the most interesting facts and scientific evidences.

Chapter 2 covers the background, motivation and usages of the idea and its contribution to DNA Forensics and research.

As for Chapter 3, it includes surveys of similar software and applications.

Chapter 4 describes the requirement analysis of the project, whereas Chapter 5 and 6 shows the design and implementation phases of the project.

Chapter 7 states the conclusions and Chapter 8 states the future work to be made upon this project, while Chapter 9 describes the user and programmer manuals.

At the end of this document, appendix A is provided for user manual and appendix B for a programmer guide sample. Also, a reference to the major studies will be provided.

CHAPTER 2|MOTIVATION & BACKGROUND

2.1 MOTIVATION

FORENSIC PURPOSES

DNA analysis is starting to be interesting, as well as important, for different and critical fields. One of which is forensic science, as mentioned before. Forensic science is to aid in criminalistics using trace evidence analysis and DNA forensic profiling.

GWA studies are of great contribution to forensics. To explain, if we are able to study human's genes and genes' variants to tell about their outer structure and traits, and further predict them, placing a suspect of a crime could be a lot more maintainable.

DNA markers related to eye color have been identified, making it feasible for predicting human's phenotype using genome variations, which are valuable to forensic studies.

RESEARCH-WISE

Furthermore, another one of the main motives is supporting researchers, mainly genetic researchers. This reason possesses huge interest, as the need for geneticists increase for more progress to identify new genome sequences, to study the effect of genes alleles that can lead to human morphology structure.
Our application provides potential help in this area, it allows researchers to test certain variants and study its effect on phenotypes, being provided by their database of SNPs.

2.2 BACKGROUND

BIOLOGICAL BACKGROUND

DNA

DNA is recorded using a sequence of nucleotides which are: guanine, adenine, thymine, and cytosine; and mostly referred using their 4 initial letters: G, A, T, and C.
So DNA, just like a binary compiled file that consists of 0's and 1's, is a compiled file consisting of these four letters.
Mainly, humans' DNA contains 3.2 billion letters.

And just like an executable binary file, the DNA is an executable biological file.
No two organisms can have the same DNA sequence, making it a genetic fingerprint-like identifier.

GENES

The DNA is being cut into smaller segments called "genes". Each gene is responsible for being translated into a specific protein. 30,000 genes have been identified so far.

Gene variants at certain loci are responsible for traits in human being such as eye or hair color.

SNP

Single-nucleotide Polymorphism is a DNA variation sequence occurring within a locus in a DNA fragment and differs from another one.

Almost all common SNPs have only two alleles. SNPs are used in GWA studies; with the assistance of bioinformatics, researchers have been trying to include more variations in a database of SNPs found in certain genes, which can be useful in genetic mapping, its influence on human traits and diseases.

SNP is an important concept in our application. An example of a SNP is “rs1800407” In gene OCA2 which is mainly responsible for the human eye color trait.

EYE COLOR GENETICS

Genetics of eye color is a very complex process; as GWA studies have demonstrated. Multiple genes could be involved; however, it’s been found out that the genes OCA2 and HERC2, existing at chromosome 15, are the two main genes that explain the variation in eye color.
We’ve used nine different SNP variations associated to eye color for prediction. Details and conclusions are to be provided later.

CHAPTER 3|SURVEY

STRANGER VISIONS

As quoted from their website:

> *"In Stranger Visions artist "Heather Dewey-Hagborg" creates portrait sculptures from analyses of genetic material collected in public places. Working with the traces strangers unwittingly leave behind, Dewey-Hagborg calls attention to the impulse toward genetic determinism and the potential for a culture of genetic surveillance."*

Heather was able to extract basic heredity traits from the genetic material, including hair and eye color, ethnicity and gender.

While a lot of guesswork had to be done to get the 3D portraits.

ANCESTARY-DNA

Speculate ones family trees' ethnicity/race using their DNA structure.

CHAPTER 4|PROBLEM STATEMENT

4.1 INTRODUCTION

PURPOSE

- To build software that associates an organism's genotype with its phenotype.
- To contribute an open-source software to the field of forensic sciences.
- To assist the genetic researchers, as well as interested individuals.

ISSUE STATEMENT

The DNA contains enormous information that we haven't fully understood yet, for instance, the relation between the genotype and phenotype of an individual. How, given the gene sequence of an individual, could we determine some phenotype trait of his?

OBJECTIVE

- The software should predict a desired physical trait given some chosen DNA SNP variations.
- The software should confirm a gene SNP is affecting a specific phenotype, given enough data.

PROBLEM SCOPE

- This software is phenotype independent, i.e. the user chooses what physical features to predict.
- This software is animal-type independent, i.e. it can be used on animals other than Homo sapiens.
- This software provides further analysis to the results, like error measurements and PDF reports.

USERS OF THE SYSTEM

- The program is intended for genetics researchers.
- Interested individuals with little or no knowledge of neural networks.

4.3 FUNCTIONAL REQUIREMENTS

1- TRAIN

Given multiple user's gene data and corresponding physical feature, train the neural network to learn and produce appropriate weights.

This is necessary for first-time use of the system, in which the user doesn't have a weights file to predict the physical feature of an unknown.

Output: produce weights file that represents the relation between trained input/output data.

2- TEST

Given one or more individual's gene data, and weights file, the system will attempt to predict the physical feature to each individual.

Output: predicts with an accuracy that corresponds to that specific at the training stage.

3- ANALYSIS

After the training stage the system will provide the user with error measurements and further analysis that should help the researcher in the decision make process.

Output: PDF report of the analysis.

CHAPTER 5|PROPOSED SOLUTION

5.1 DEVELOPMENT ENVIRONMENT

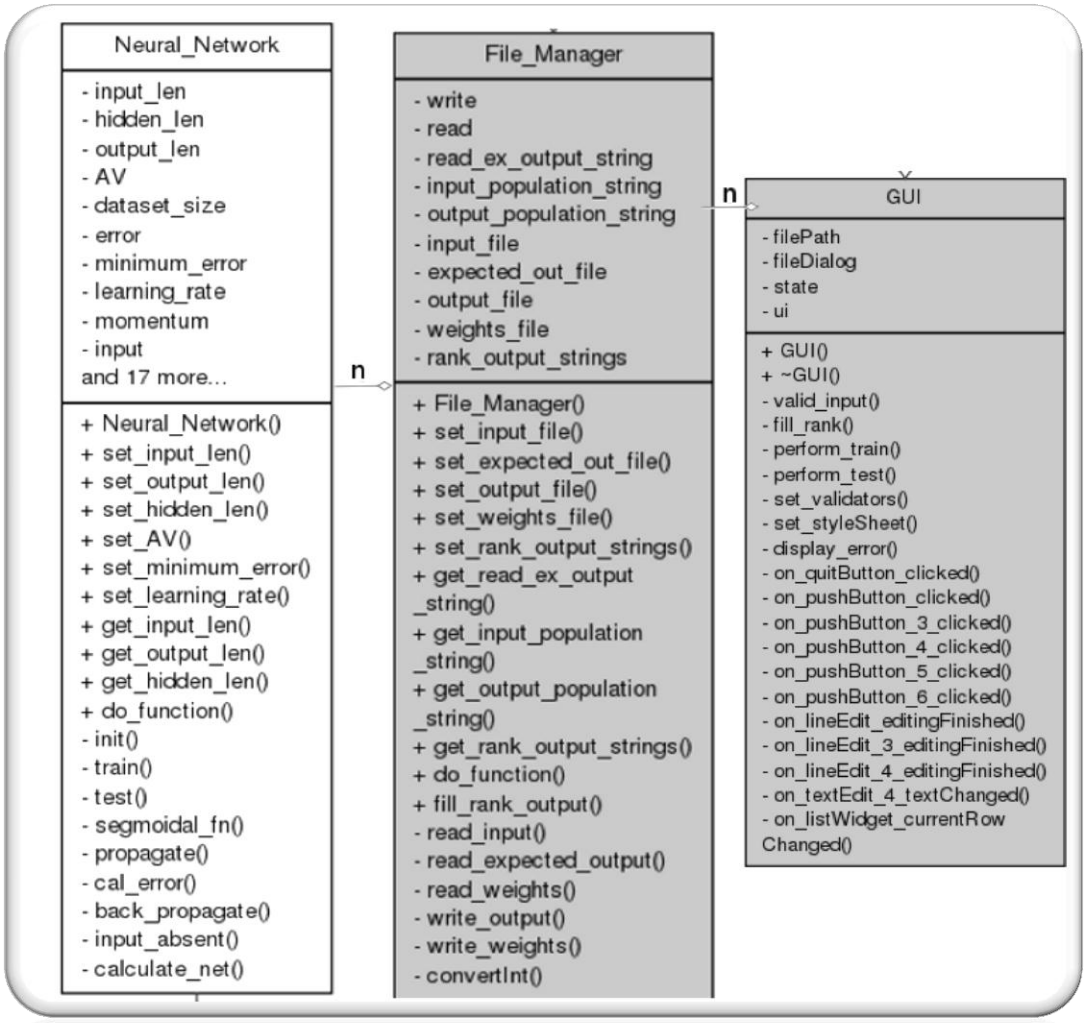
LANGUAGE AND TOOLS

- ✓ Git: used for source code management.
- ✓ Qt framework: used for development.
- ✓ Qt Creator: used for GUI development.
- ✓ C++: The main code.
- ✓ R: “ropensnp” to obtain the data from gene’s database.
- ✓ DoxyGen: used to create a website which serves as a programmer guide.

ALGORITHMS AND DATA STRUCTURES

- ✓ Feed forward Neural Network.
- ✓ Back propagation is used.
- ✓ Multilayer, input, hidden and output.
- ✓ Array, Vectors and Maps data structures are used.
- ✓ Softmax activation function is used.
- ✓ Cross entropy error function is used.

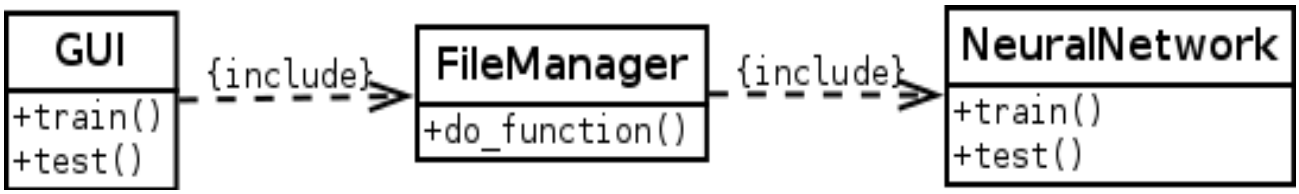
5.2 CLASS DIAGRAM



5.5 DESIGN PATTERN

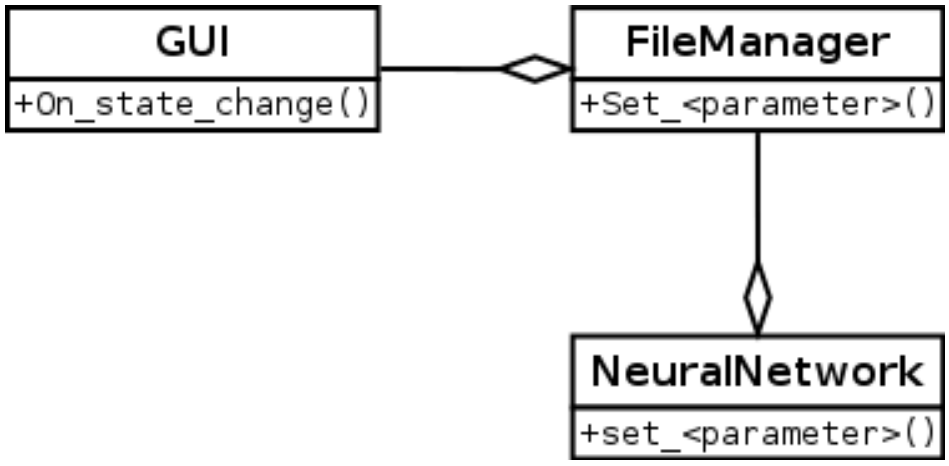
SINGELTON

Singleton is used implicitly in the implementation:



OBSERVER

Observer is also used implicitly, where any change in the GUI parameters results in internal changes in both the FileManager and the NeuralNetowk parameters.



CHAPTER 6|IMPLEMENTATION

6.1 DATA SOURCE

Ropensnp

Is a library that is developed under the R-project. It was used to obtain the genotypes and phenotypes of several individuals, in our case, around 300.

R script

```
library('ropensnp')
sink("output", append=TRUE, split=FALSE)
pheno <- phenotypes_byid(phenotypeid=1, return_ = 'users')
for (i in 1:400)
{
user <- pheno[i, 1]
us <- paste0("'0,",user,"'")
gen1<-gen2<-gen3<-gen4<-gen5<-gen6<-gen7<-gen8<-gen9<--1
try(gen1 <- genotypes(snp='rs12913832', userid=us ,df=TRUE),silent=TRUE)
try(gen2 <- genotypes(snp='rs1800407', userid=us ,df=TRUE),silent=TRUE)
try(gen3 <- genotypes(snp='rs1393350', userid=us ,df=TRUE),silent=TRUE)
try(gen4 <- genotypes(snp='rs12896399', userid=us ,df=TRUE),silent=TRUE)
try(gen5 <- genotypes(snp='rs16891982', userid=us ,df=TRUE),silent=TRUE)
try(gen6 <- genotypes(snp='rs12203592', userid=us ,df=TRUE),silent=TRUE)
try(gen7 <- genotypes(snp='rs7495174', userid=us ,df=TRUE),silent=TRUE)
try(gen8 <- genotypes(snp='rs6497268', userid=us ,df=TRUE),silent=TRUE)
try(gen9 <- genotypes(snp='rs11855019', userid=us ,df=TRUE),silent=TRUE)
cat(user,"\n")
result <- 1
result <- try(cat(gen1[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen1,"\n")
result <- 2
result <- try(cat(gen2[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen2,"\n")
result <- 3
result <- try(cat(gen3[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen3,"\n")
result <- 4
result <- try(cat(gen4[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen4,"\n")
result <- 5
result <- try(cat(gen5[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen5,"\n")
result <- 6
result <- try(cat(gen6[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen6,"\n")
result <- 7
result <- try(cat(gen7[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen7,"\n")
result <- 8
result <- try(cat(gen8[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen8,"\n")
result <- 9
result <- try(cat(gen9[[7]],"\n"),silent=TRUE)
if (!is.null(result))cat(gen9,"\n")
}
pheno[[2]]
sink()
```

6.2 SIGNIFICANT CODE FRAGMENTS

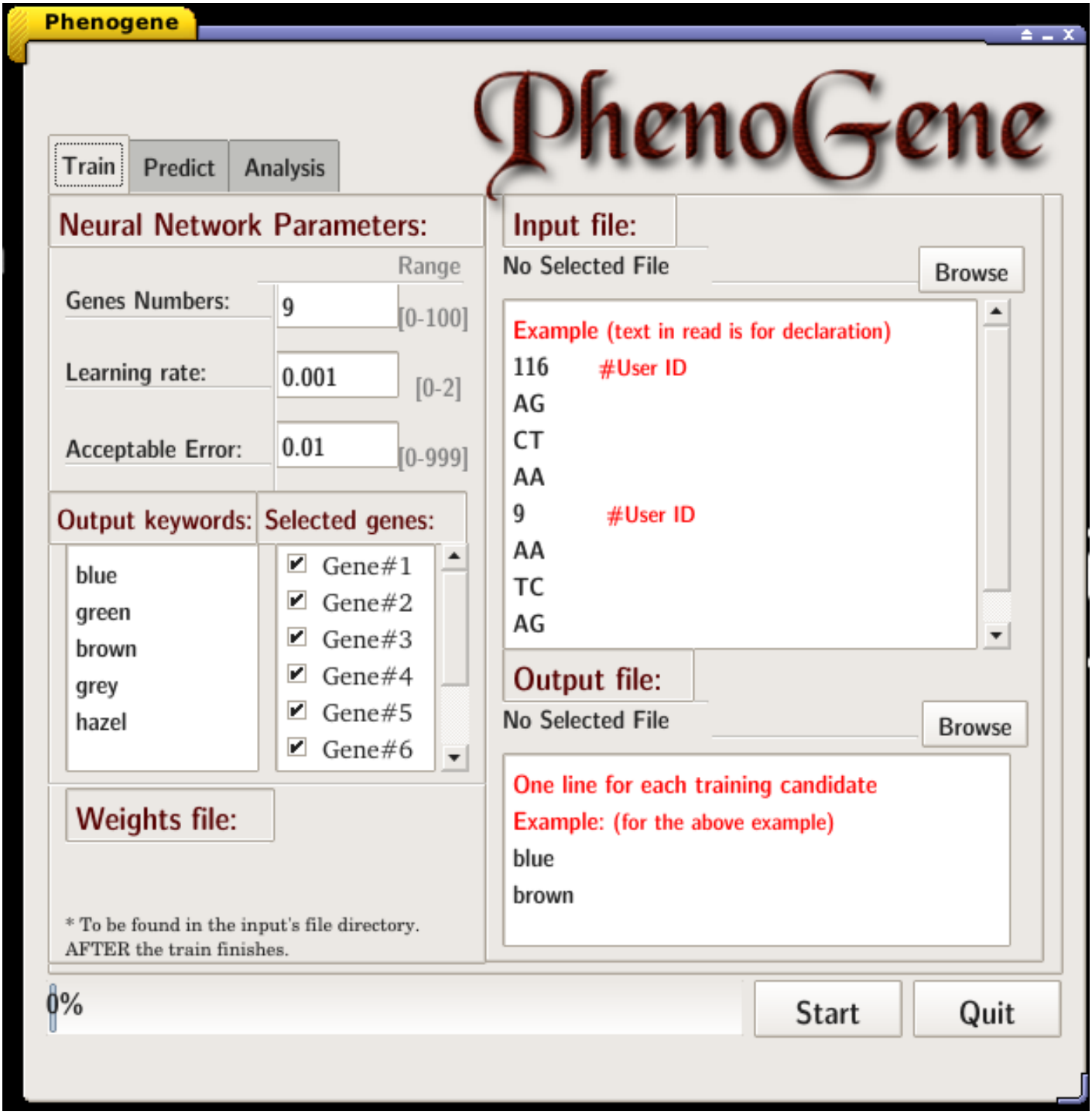
TEST

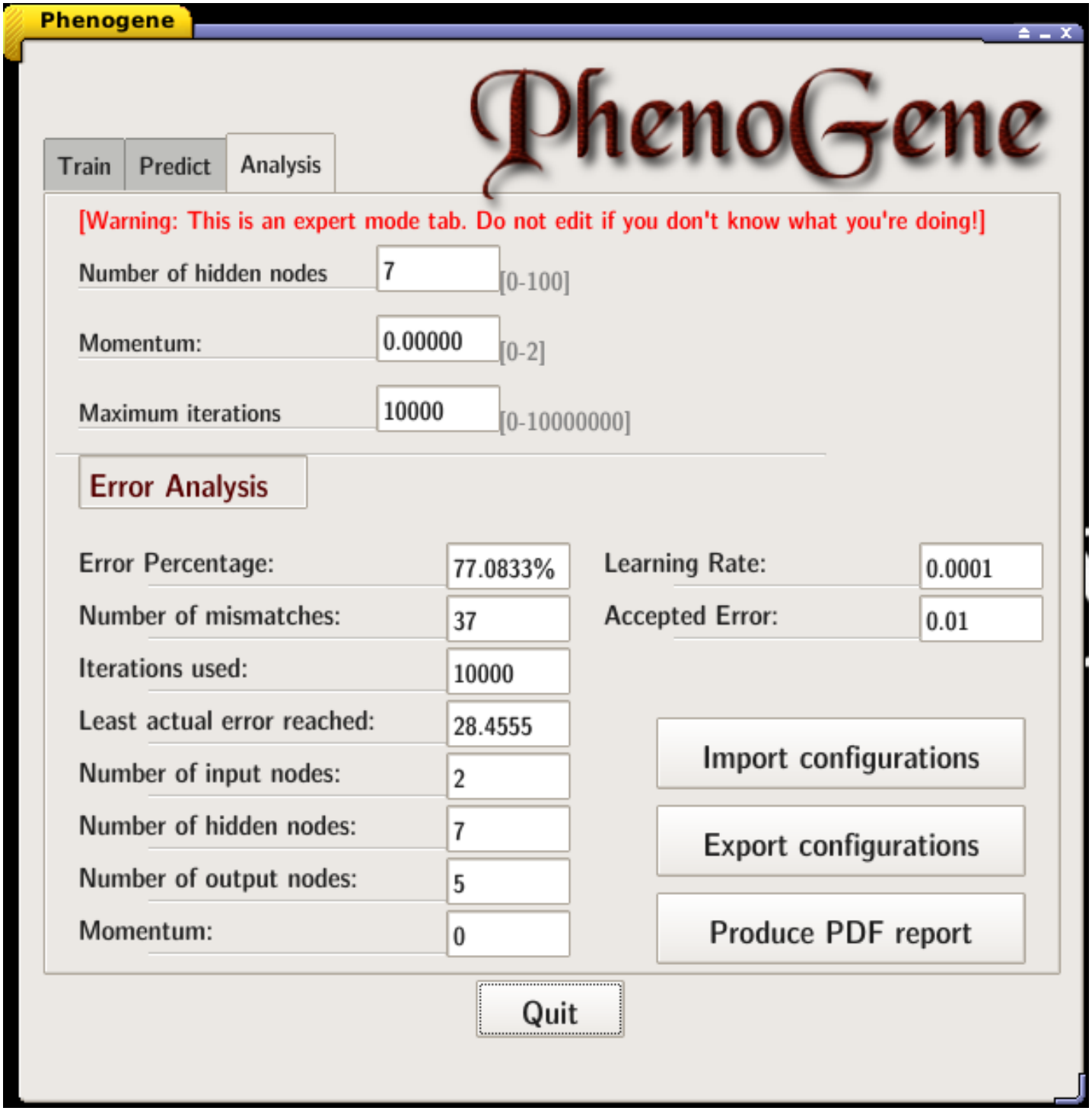
```
141 void Neural_Network::predict()
142 {
143     error = 0;
144     output_dataset.clear();
145     output_dataset.resize(dataset_size);
146     fork(0,dataset_size)
147     {
148         // Get current dataset of input
149         fori(0,input_len)
150                 input[i]=input_dataset[k][i];
151         // Calculate Output
152         propagate();
153         output_dataset[k].resize(output_len);
154         forj(0,output_len)
155                 output_dataset[k][j]=output[j];
156     }
157     return;
158 }
```

FEED FORWARD

```
168 void Neural_Network::propagate()
169 {
170     fill_n(hidden, hidden_len, 0);
171     fill_n(output, output_len, 0);
172
173     double max = -1000.0;
174     double maxO = -1000.0;
175
176     /*Calculate the net (weights x input)*/
177     calculate_net(max,maxO);
178     double exp_HT=0.0;
179     double exp_OT=0.0;
180
181     /*Softmax activation function*/
182     fori(0,hidden_len) exp_HT+=hidden[i]=exp(netH[i]);
183     fori(0,output_len) exp_OT+=output[i]=exp(netO[i]);
184
185     fori(0,hidden_len) hidden[i]=hidden[i]/exp_HT;
186     fori(0,output_len) output[i]=output[i]/exp_OT;
187     return;
188 }
```

6.3 GUI SCREEN SHOTS




Phenogene
PhenoGene
Train
Predict
Analysis
[Warning: This is an expert mode tab. Do not edit if you don't know what you're doing!]
Number of hidden nodes
7
[0-100]
Momentum:
0.00000
[0-2]
Maximum iterations
10000
[0-10000000]
Error Analysis
Error Percentage:
77.0833%
Number of mismatches:
37
Iterations used:
10000
Least actual error reached:
28.4555
Number of input nodes:
2
Number of hidden nodes:
7
Number of output nodes:
5
Momentum:
0
Learning Rate:
0.0001
Accepted Error:
0.01
Import configurations
Export configurations
Produce PDF report
Quit

CHAPTER 7|CONCLUSION

7.1 ERROR MEASURMENTS

ERROR TOLERANCE

As the bioinformatics field is fairly new, a 20-30% of correct results is accepted, that leaves a space of 70-80% of error tolerance.

The application reported 78% false predictions. That may be due to the small size of available dataset that is associated with phenotypic traits.

That, together with the probability that the SNPs tested weren't actually as significant as the research studies stated, resulted in the neural network not being able to learn and classify the input.

7.2 PARAMETERS DETERMINATION

LEARNING RATE

The learning rate is directly proportional to the dataset size.

We've noticed that as the dataset size decreases, a high learning rate would result in a higher error percentage in the training process. While a smaller learning rate would provide better results.

MOMENTUM

The momentum is only useful when the training is slow and the learnt information is too high for the neural to classify fast.

We've noticed that setting the momentum to any value higher than zero when training small datasets would result in a failure to learn.

CHAPTER 8|FUTURE WORK

PHENOTYPES

The software might be tested in other phenotypic traits, even in other organisms, to affirm or refute the results of research findings. As well as for prediction of unknown phenotypic traits upon a successful train.

NEURAL NETWORK DESIGN

Further improvements to the design of the neural network may be conducted. Using of different error function, activation function or a mix of different activation functions on the different layers.

BETTER DATA

The software may be tested as more phenotype/genotype data becomes available.

OPEN SOURCE

The software will be released as an open-source software in which other interested programmers may contribute to it.

CHAPTER 9|MANUALS

8.1 PROGRAMMER MANUAL

- ✓ The manual was generated using “doxygen”, which produced an HTML website to help the programmer navigate the code easily.
- ✓ The code commenting format follows one understood by doxygen. A programmer might view these comments in the website as well.
- ✓ The website is attached with the provided CD’s.

8.2 USER MANUAL

A user manual is attached to appendix A. The manual should guide the user through the steps of preparing their data, using the application, and make use of the analysis tools the application provides.

CHAPTER10|REFERENCES

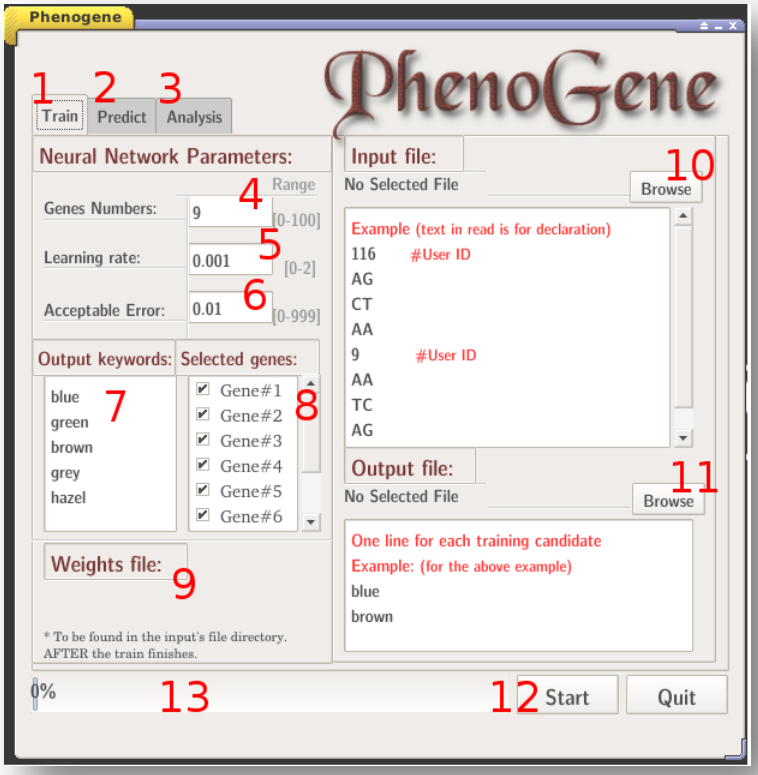
1. **Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand:**
   G.E. Nasr, E.A. Badr and C. Joun
   School of Engineering and Architecture
   Lebanese American University, Byblos, Lebanon
   e-mail: genasr@lau.edu.lb
2. **The Error Entropy Minimization Algorithm for Neural Network Classification:**
   Jorge M. Santos, Luís A. Alexandre and Joaquim Marques de Sá INEB - Instituto de Engenharia Biomédica
   R. Roberto Frias, Porto, Portugal
   e-mail: {jms@isep.ipp.pt, lfbaa@di.ubi.pt, jmsa@fe.up.pt}
3. **Backpropagation Learning:**
   15-486/782: Artificial Neural Networks
   David S. Touretzky
4. **A Three–Single-Nucleotide Polymorphism Haplotype in Intron 1 of OCA2 Explains Most Human Eye-Color Variation:**
   David L. Duffy,* Grant W. Montgomery,* Wei Chen, Zhen Zhen Zhao, Lien Le, Michael R. James, Nicholas K. Hayward, Nicholas G. Martin, and Richard A. Sturm
5. **Eye color and the prediction of complex phenotypes from genotypes:**
   Current Biology Vol 19 No 5 R192
   Fan Liu1, Kate van Duijn1, Johannes R. Vingerling2,3, Albert Hofman2, André G. Uitterlinden2,4, A. Cecile J.W. Janssens2 and Manfred Kayser1,*
6. **Free Executive Summary - The Evaluation of Forensic DNA Evidence:**
   ISBN: 978-0-309-05395-2, 272 pages, 6 x 9, hardback (1996)

Additional references.

7. **Softmax activation function:**
   http://cseweb.ucsd.edu/~elkan/254/apr3.html

APPENDIX A | USER MANUAL SAMPLE

TRAIN TAB



1. Training tab
2. Predicting tab
3. Analysis tab
4. Number of input genes, i.e. how many gene SNPs do you have for each user?
5. Learning rate: This is usually a number between 0.3-0.7, if the dataset is too small choose even a smaller number. This number is directly proportional to your dataset size.
6. The least error you'd be tolerant about. Usually if this is the first time you're training the network you won't know this value, choose a small number or even zero then.
   The network will then continue to train until maximum iterations number is reached.
7. Phenotypic traits you're predicting. These keywords should match with your expected_output_file keywords.
8. Select which genes to train. This is mainly used for studying the effect of excluding a gene or more on the prediction and error rates.
9. The weights file will be produced upon successful training. You can find it in the same directory as the input and expected files you browsed earlier.

Mainly, the weights file describes the relation between the input and the output, and it will be used later on for actual prediction on new data.

10. Browse your input file. Choose one that follows the correct format specified below:

    Your data file is a text file that has many individual records. For each individual the record starts with an integer number, his ID, follows the 2-letters gene SNPs of that individual each in a line. Then the second ID for the second individual, follows his gene SNPs and so on so forth.
    Example:
    1
    AA
    AT
    2
    CC
    CG
11. Browse your expected file. Choose one that follow the correct format specified below:

    Your data file is a text file that has many individual records. For each individual there's exactly one line. The line may have one keyword or more than one keyword separated by a comma.
    Example:
    blue
    green

12. Press this button when you have filled all the data.
13. Progress bar, this is updated each time the network makes a progress.

APPENDIX B | PROGRAMMER GUIDE

This website is to be found on the provided CD's as well.