INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER (UOW)

BEng/BEng.(Hons) in Software Engineering

Final year Project 2014/2015

Software Requirement Specification

For

Project Title: **Identify Inherited Diseases based on DNA (IIDDNA)**

By

Iddamalgodage Don Lahiru Manohara - 2010070

Supervised By: Mr. Achala Chathuranga Aponso

…….. ……..

Signature of Supervisor Signature of Student

Table of Contents

1. METHODS OF REQUIREMENT ELICITATION 1
1.1. Literature Review 1
1.2. Online Survey 1
1.3. Email Correspondences with Domain Experts 1
1.4. Interviews with domain experts 1
2. SURVEY FINDING – EXPERT SURVEY 2
3. STAKEHOLDERS 8
4. USE CASES 9
5. FUNCTIONAL REQUIREMENTS 10

USE CASESTable 1: Stakeholders and Their Roles 9
Table 2: Functional Requirement 10

Figure 1: Experience of Domain Experts 2
Figure 2: Method Usefulness 3
Figure 3: Biological Area Related to Inherited Diseases 3
Figure 4: Relation of The SNPs and Most Of the Inherited Diseases 4
Figure 5: Disease SNPs identification Method 5
Figure 6: Methods for Prioritize Candidate Disease Genes 6
Figure 7: Classification Methods for Disease Gene Prioritization 6
Figure 8: Development tools 7
Figure 9: Onion Model Showing Stakeholders of The IIDDNA 8
Figure 10: Usecase Diagram 9

1. METHODS OF REQUIREMENT ELICITATION

Following requirement were gathered from different source, and based on requirement engineering techniques in order to cover wide range of the domain and impartiality.

Literature reviews, online surveys, written surveys, e mail correspondence with domain experts and interviews were considered to gather requirements for this research project.

1.1. Literature Review

Literature review was conduct to find out the sate-of-art techniques in SNPs (Single Nucleotide Polymorphism) based approaches for identifies inherited diseases, Gene prioritization for identify inherited diseases, Protein-Protein interaction network for identify inherited diseases.

1.2. Online Survey

Questionnaires were prepared for target audience such as bioinformatician, domain experts, doctors and medical students. The survey was assist to understand aspect of the data analysis techniques of biological methods and clinical system which were not concerned by the author.

1.3. Email Correspondences with Domain Experts

Email correspondences is a grate techniques stayed in touch with the research community and the medical experts in order to understand their opinions on particular problem domain.

1.4. Interviews with domain experts

The author interviewed domain experts to gain more technical knowledge in order to identify the depth of the problem and possible solution. The domain expert group was Dr Prashanth Suravajhala (PHD studies from LJR lab at Rokilde and Aalborg University, Denmark), Lahiru Prabodha (PHP student).

The author also interviewed few medical experts to gain knowledge about diseases which are inherited and identify the relation between genes and diseases.

2. SURVEY FINDING – EXPERT SURVEY

Following are the result of the expert survey.

Experience of domain experts

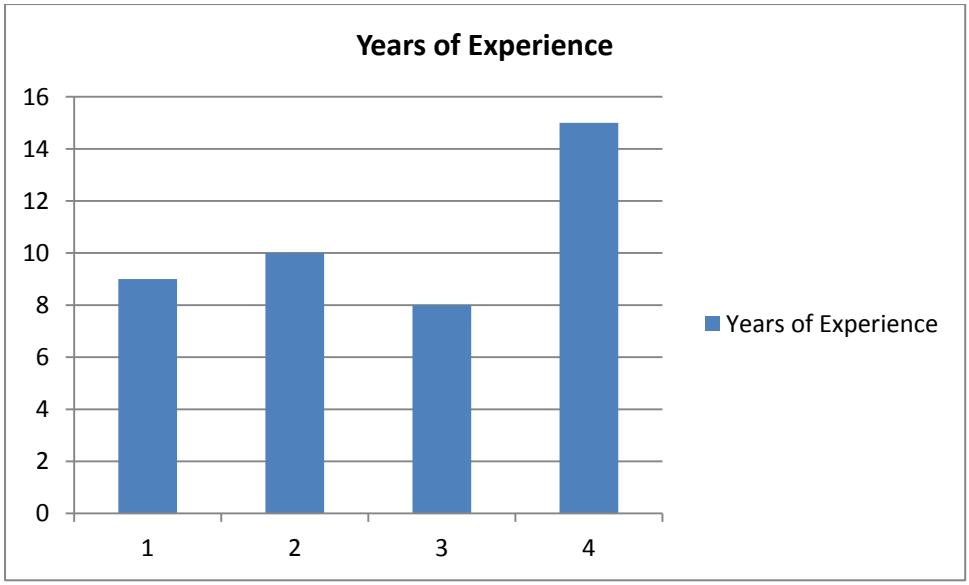

Figure 1: Experience of Domain Experts

Designation

- gene-gene interactions analysis
- Research Scientist
- Senior Lecturer

Country/City please specifies your current location.

- Russia/Kursk
- Denmark
- England/Bristol
- London

What are the medical domain which you used machine learning techniques?

- I did not use it
- ANN
- Cancer

Is it useful method to identify inherited diseases based on DNA related data?

Figure 2: Method Usefulness

Please specify the reason?

- it is designed for other purposes
- The best thing would be take the known unknown regions and classify the variants as per NN
- We have been successful at predicting mutations that are associated with heritable traits using machine learning.
- There are too many variables and far too few data points and so the problem is severely under powered.

Please specify which areas are mostly related to inherited diseases?

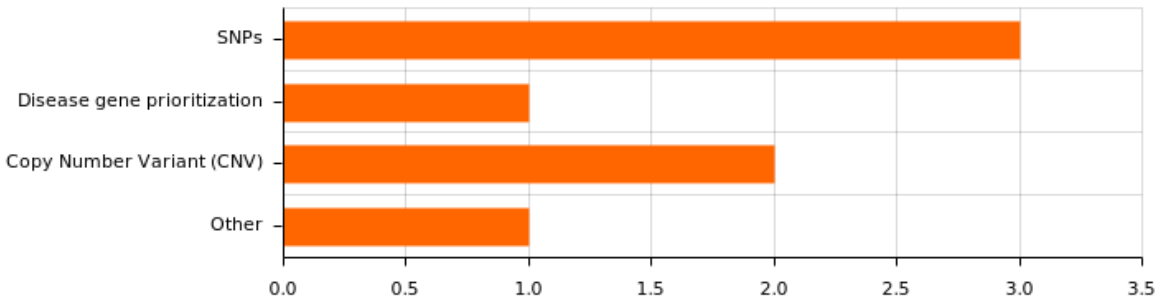

Figure 3: Biological Area Related to Inherited Diseases

Please specify the reason?

- The idea candidates would be CNVs though as per my obvious reason
- We have only done research in one of these areas, so I do not have enough expertise in the remaining areas to make a determination.
- It depends on the disease but most inherited diseases start from an SNP but in cancer for example CNV is very important as a somatic mutation which leads to a tumour cell.

What are the methods you used to analysis DNA data to identify inherited diseases?

- PCR based methods
- KNN
- Support vector machine classifiers
- cluster analysis of gene expression

Please specify the reason why you were using particular method?

- cheep, reproducible
- We are incorporating several different kinds of data into our classifiers. While there are numerous methods for accomplishing this (e.g., ensemble learning, multiple-kernel learning, multivariate regression), support vector machines currently yield state-of-the-art performance in many domains.
- There is a huge amount of publicly available data.

Is nsSNPs causing most of the inherited diseases?

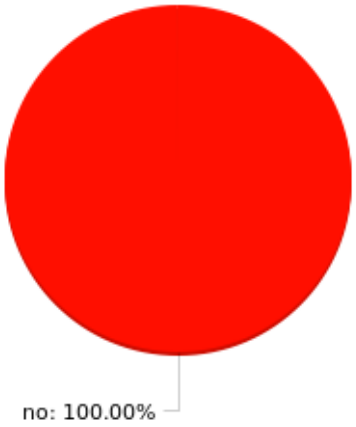

Figure 4: Relation of The SNPs and Most Of the Inherited Diseases

Please specify the reason?

- they are neutral ones
- Not necessarily. One needs to study the ns substitution rates that are linked to SNPs and diseases in tandem.
- "Not sure" should be one of the answers: there hasn't been nearly enough research to make that determination yet! For example, we have only the most elemental understanding of how post-transcriptional modifications such as alternative splicing, ubiquitination or RNAi may influence disease.
- Most SNPs that cause disease are in regulatory regions that are neither synonymous not non-synonymous. nsSNPS are the easy to identify disease causing mutations in things like thalessemia, sickle cell anaemia and cystic fibrosis but these are not examples of common genetic diseases such as cancers, Alzheimers, Parkinson's etc.

What are commonly using computational methods to identify inherited diseases associated SNPs?

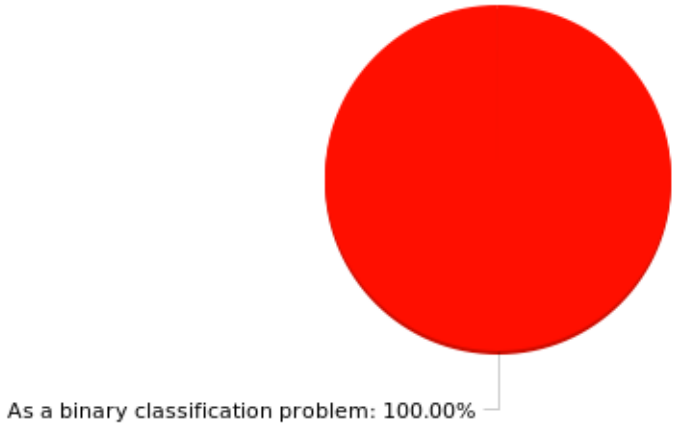

Figure 5: Disease SNPs identification Method

Please specify what inherited diseases are identify from SNPs analysis?

- many disorders
- A huge variety of different diseases, including cancers.
- Breast cancer

What is the method you used to prioritize candidate disease genes related to inherited diseases?

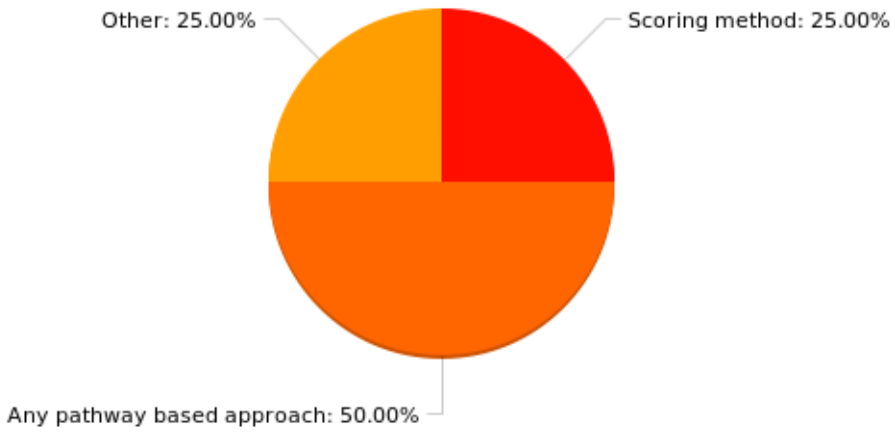

Figure 6: Methods for Prioritize Candidate Disease Genes

Please specify the reason?

- it is closer to disease pathogenesis
- We used six point classification scoring schema and used ML approaches in training a network of proteins across 22 classifiers.
- All of these may be used to score genes, mutations, indels, etc.
- The other methods are artificial garbage that does not capture the biology correctly.

Which classification method you used to candidate disease gene prioritization?

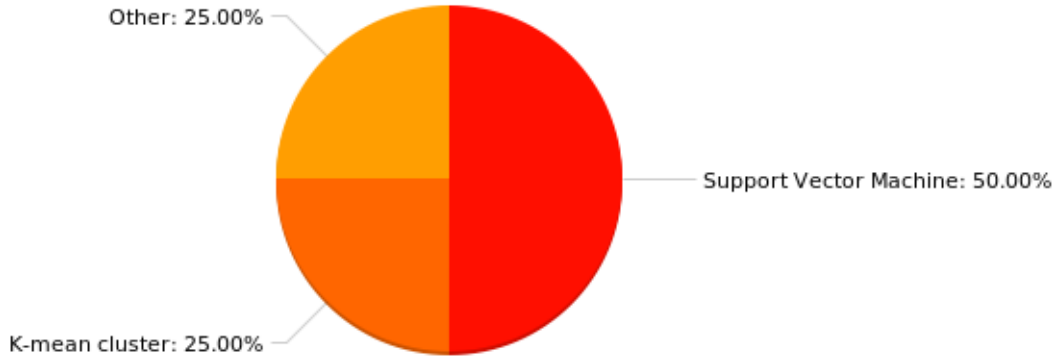

Figure 7: Classification Methods for Disease Gene Prioritization

Please specify what are the inherited diseases can be identify form the candidate gene prioritization method?

- Down syndrome
- A huge variety of different diseases, including cancers.
- I have used these in no cases as lung cancer is not inherited.

What are the tools you have used to diseases identification applications?

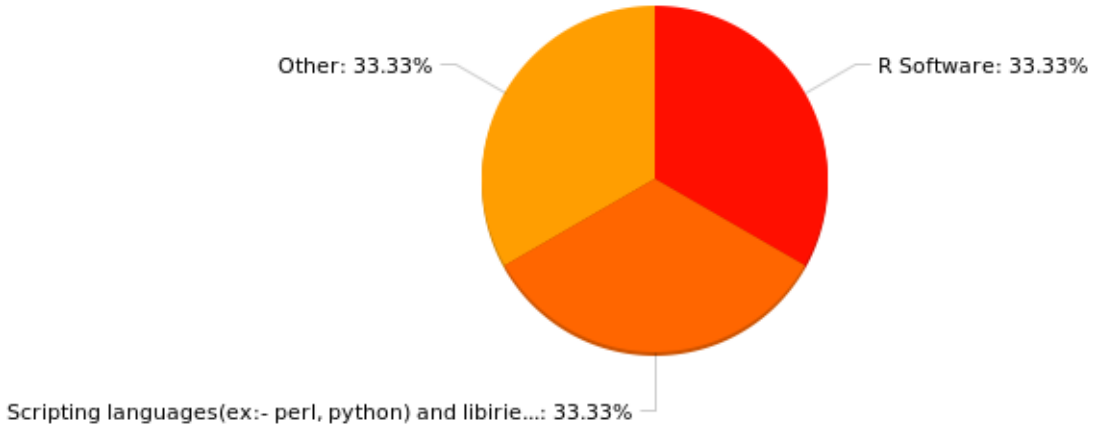

Figure 8: Development tools

Is there any specific reason for used particular tool?

- No
- Perl and PHP could be a choice as well although I am not sure how.
- These are the most commonly used tools for machine learning. I use Python for most things; my colleagues use Matlab or R, bash scripts for pulling data from online databases, and other tools (e.g., gnuplot) for data and graphical analysis.
- It has a huge number of bioinformatics libraries in the Bioconductor packages.

What tool is mostly recommended for this project?

- Weka
- R-Bioconductor

Please specify the reason?

- It has a flair for all specifics for this problem.
- There is no single tool that is best for performing machine-learning research.

Please add some comments on this project?

- I think the idea is perfectly good. Please try considering the hypothetical protein candidates as well. Any help, please do let us know.

3. STAKEHOLDERS

The following stakeholders have been identified for inherited diseases identification system. The role of each stakeholder is specified below in Table

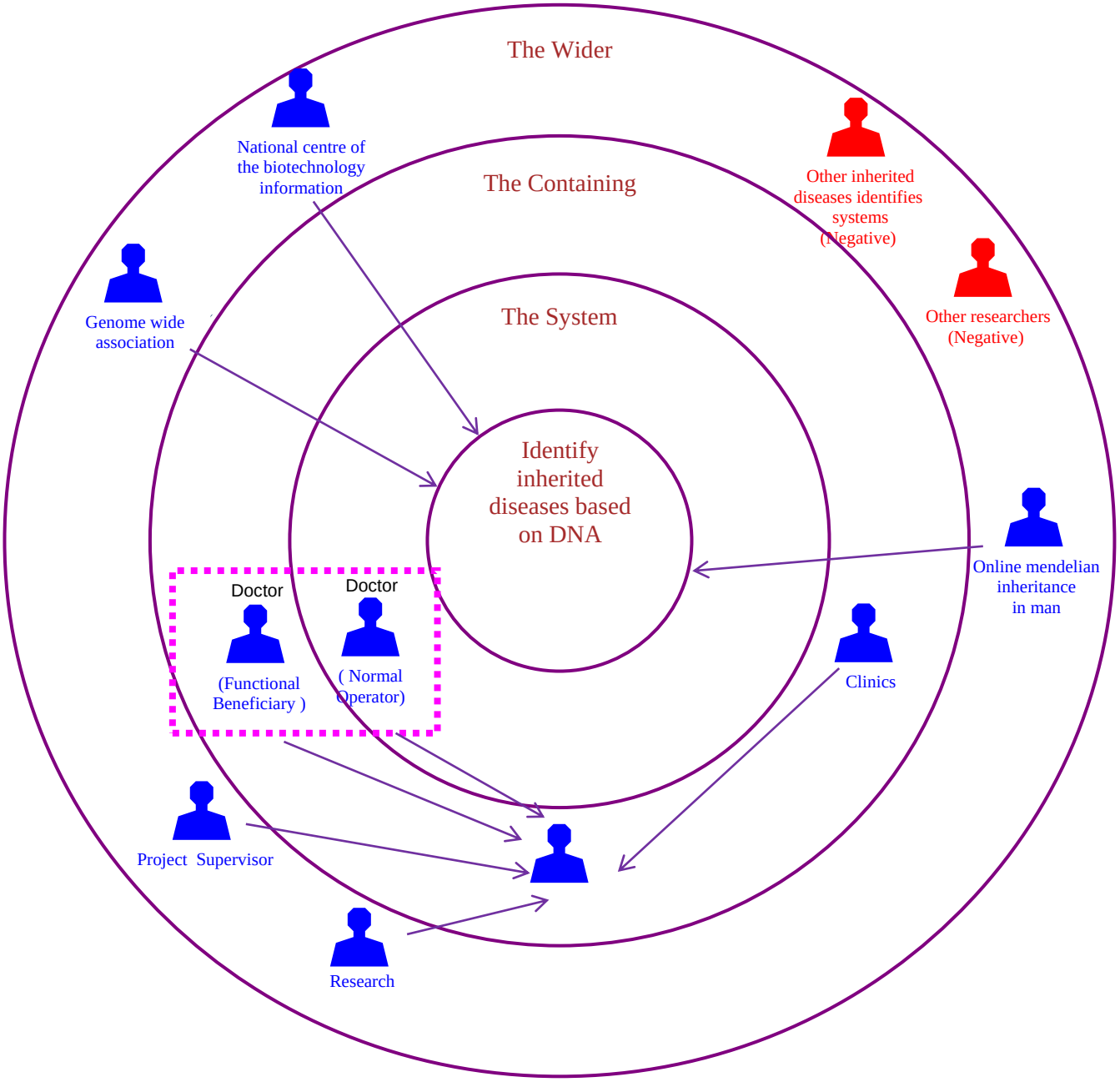

Figure 9: Onion Model Showing Stakeholders of The IIDDNA

Stakeholder	Role	View point
Doctor	Functional Beneficiary, Operator	Get useful information from the diseases identifies system
Researcher	Intellectual Beneficiary, development	Researcher and develop diseases identifies system.
Public	Regulator	Public will be able to get advantages from the inherited diseases identifies system.
Clinics	Financial Beneficiary	This system should be recommended to the doctors to find inherited diseases for patient.

USE CASESTable 1: Stakeholders and Their Roles

4. USE CASES

Use Case Diagram explains the functionality of a system and users of the system. The diagram includes the following elements

- Actors - Which represent users of the system
- Use Cases – Which represent functionality provide by the inherited diseases identification system to user

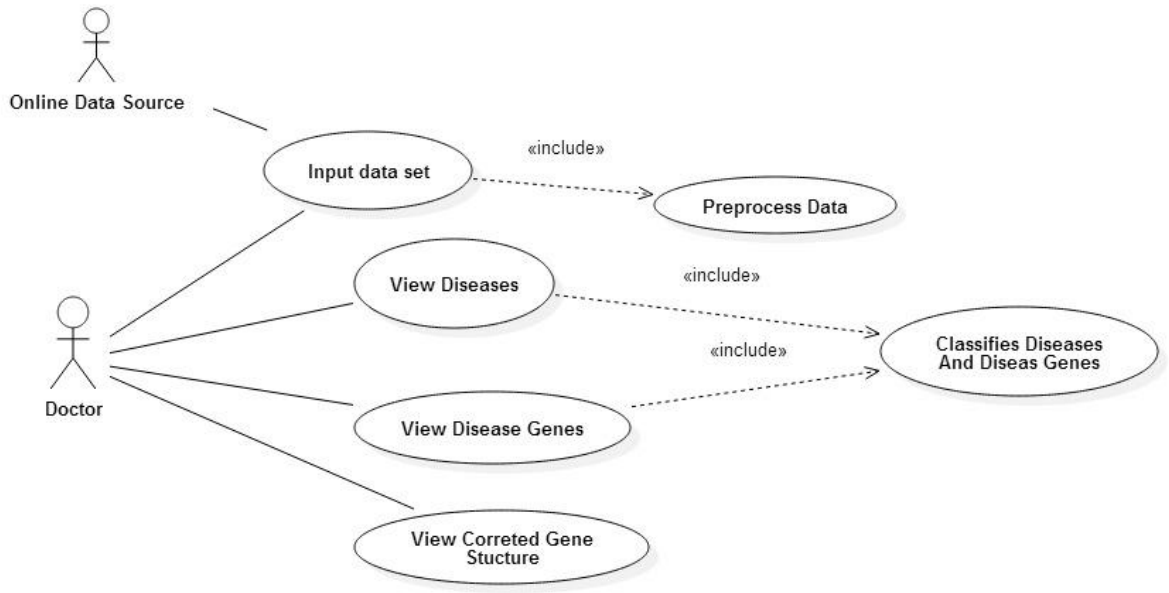

Figure 10: Usecase Diagram

5. FUNCTIONAL REQUIREMENTS

The main functional requirements are specified bellow in table

(C=Critical, D=Desirable, L=Luxury)

Id	Functional requirement	Priority	Use case(s)
FR1	The system requires genetic data set as input and data will preprocess	C	Input Data Set, Preprocess Data
FR2	Diseases should be identify by the system	C	View Diseases, Classifies Diseases And Diseases Genes
FR3	Disease genes should be identify by the system	C	View Diseases, Classifies Diseases And Diseases Genes
FR4	User should be able to view the corrected gene structure of the particular disease	D	View Corrected Gene Structure
FR5	User should be able to ask question from the system of the particular disease.	D	View Diseases
FR6	Application should generalize to identify most of the inherited diseases	L	View Diseases, Classifies Diseases And Diseases Genes

Table 2: Functional Requirement