INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER (UOW)

BEng/BEng.(Hons) in Software Engineering

Final year Project 2014/2015

Terms of Reference

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. Project Background

"What are the chances that we will one day discover that DNA has absolutely nothing to do with inheritance? They are effectively zero."(Harris, 2014). DNA is established at connection, and does not change throughout human's life time. Human receive one-half of DNA from their mother and the other half from their father. Genetic mutation occur when DNA changes, altering genetic instructions. This may result in inherited diseases. Analyzing DNA, and identify inherited diseases will show future traits of human life. Parents can be decided their unborn children's has future risk of having inherited diseases.

When someone went to the doctor for some medication for disease, most probably doctor will ask from a patient regarding particular disease affected before one of their family members. This question will explain inheritance cause some of the diseases. The Table 1 shows some diseases and percentages of the inheritance involvement. According to this table inheritance percentage is high, indicates the genetic contribution is a major factor of developing particular disease. Example: - Pyloric stenosis is high probability than for peptic ulcer or congenital heart disease. "These observations have important practical implications for both research and for patient management. If the heritability is high then a search for susceptibility genes is justified."(Nicolavich, 2010).

Diseases	Heritability (%)
Schizophrenia	85
Asthma	80
Cleft lip / cleft palate	76
Pyloric stenosis	75
Ankylosing spondylitis	70
Club foot	68
Coronary artery diseases	65
Hypertension (essential)	62
Insulin-dependent diabetes	60
Peptic ulcer	37
Congenital heart disease	35
Insulin-nondependent diabetes	30

Table 1 Estimates of heritability of various diseases (Nicolavich, 2010)

Extract of meaningful information from a large experimental data set is a key element of bioinformatics. DNA pattern analysis and extract meaningful information is a technique to identify inherited diseases. Different algorithms are in data mining techniques involve finding diseases, and visualize sample data set. There are many researches are going into DNA analysis and genetic disease. GWAS (genome-wide association study) is a one of major institute support scientist to identify gene involved in human disease. "Identifying patterns of copy number variants in case control studies of human genetic disorders" research conduct to bring missing DNA data in cell, apart from them it can identify genetic disorders (Alqallaf, 2009).The different of DAN sequence are called single

nucleotide polymorphisms(SNPs) . Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. Research like "A SNP and KEGG Based Approach to Mine Risk Pathways Associated with Bipolar Disorder" proved the SNPs are help to identify disorders (Zhang, 2008).

The Support Vector Machine (SVM) based approach uses to find out gene contributes to inherited disease. SVM is used to classify extracted features from the genes, based on gene ontology (Xie, 2012). Various data manning algorithm has been applied including linear regression, neural network and support vector regression, and create "Hierarchical learning Approach to Calibrate Allele frequencies for SNP Based Genotyping of DNA pools". This framework explains for basic genomic to a rage of new application and studies (Hellicar, 2014). Domain-domain interaction network to detect diseases-associated nsSNPs research explains how to find genetic variation and inherited disease based on statistical based scoring method (Jiang, 2012).

Even though outcomes of research were very successful, DNA analysis and find inherited disease based researches are rare. Most of the researchers are explain the approaches to determine genetic disease. But there is no clear path explain of inherit DNA analysis and identify diseases. "The inference of genes that are associated with human inherited diseases (diseases gene) has been a task of grate challenging in biological and medical studies" (Peng, 2012, p. 1). Making these research approaches are apply to identify inherited disease based on DNA analysis.

The purpose of "IIDDNA" project is identify inhered disease based on DNA analysis data set which applying different data mining techniques and bioinformatics algorithms to predict probability of having inherited diseases.

2. Aims

The aim of the project is to research, design, develop and evaluate a prototype outcome of the research, identify inherited diseases based on DNA analysis data. Also demonstrate the algorithms and methods which found from the research.

Further elaborate on the aim, the system will feed by standard data set and correctly classify which inherited diseases are exists. When creating classification model, machine learning methods for bioinformatics such as supervised learning and unsupervised learning algorithms will help to correctly classify the inherited diseases.

3. Scope

First part of the project is search, and finds approaches which researchers were taken in their researches. Based on these approaches try to find out solution to identify inherited disease using visualizes, and running different algorithms on data set in different statistical analysis tools.

When successful approach found from the testing and running different algorithms on data sets, develop the prototype application with limited futures. User input sample test data with given format, the application runs the algorithm and display probability of having particular disease.

Due to the time constrains the prototype supports essential functionalities but yet sufficient to demonstrate outcome of the research. The feature like visualize data set is not supported in prototype.

For some disease heritability is directly involve but some disease are not (Nicolavich, 2010). These types of diseases not count only inheritance is major factor. So this system has only find the inherited disease such as Schizophrenia, “Schizophrenia is a chronic, severe, and disabling brain disorder that has affected people throughout history”(NIH, 2014), Asthma, “Asthma (AZ-ma) is a chronic (long-term) lung disease that inflames and narrows the airways”(NIH, 2014), Cleft lip / cleft palate, “Cleft lip and cleft palate are birth defects that occur when a baby's lip or mouth do not form properly”(NLM.NIH, 2014), pyloric stenosis, “Pyloric stenosis is a narrowing of the pylorus, the opening from the stomach into the small intestine”(NLM.NIH, 2014), ankylosing spondylitis, “Ankylosing spondylitis is a type of arthritis of the spine”(NLM.NIH, 2014).

4. Objectives

1. Writing Terms of Reference report to identify the problem domain of the project and define scope, objectives and deliverables.

❖ **Output Artifact: Terms of Reference Document**

2. Conducting a literature survey on the following topics to gain depth knowledge and analyze past works done in those areas and estimate the future work needed to be done.
   - Different approaches for analysis DNA patterns.
   - Existing application and prototype for DNA analysis, and identify diseases.
   - Data mining techniques related to DNA analysis and classify diseases.
   - Predictions algorithms for probability of having classified disease.

❖ **Output Artifact: Literature Review Document**

3. Identify and analyze requirements by conduction online surveys and interviews for a prototype evaluation. This phase helps to improve prototype features.

❖ **Output Artifact: Software Requirement Specification (SRS)**

4. Design the application incorporating the requirements prioritized through analysis and using appropriate design methodology to guide the implementation.
5. Select appropriate development technologies and tools which are very ideal for the development of the application. Identifying the appropriate technology is significant to develop an efficient prototype.

❖ **Output Artifact: Design Specification**

6. Develop a prototype for demonstrating proposed features.
   Developing the following components is necessary to fulfill the goals of the project.
   - Development prototype to demonstrate a solution found from the research.

❖ **Output Artifact: Application to demonstrate a solution found from the research.**

7. Test the prototype by unit testing and also by a black box testing approach to ensure all the requirements gathered in the requirement gathering phase are incorporated and functional.
8. Evaluate the prototype by allowing potential users to use the prototype and obtain feedback to ensure the outcome is good to be usable and user-friendly.
   8.1. Enhance prototype according to user feedbacks.
9. Submit all the project deliverables with in the dead line.

5. Features of Prototype

- Plugin data source or data set

  The application is able to plug in different data sources with pre-defined data format to analysis, and produce the appropriate result.

- Implemented algorithms found from the research

  The best approach which found from the research, implement the application with appropriate libraries and custom logic to develop the business logic of the application.

- User Interface for display result in understanding manner

  Eliminate complexity in the process of analyzing data, change threshold in algorithm and easily understand the test result etc. Design simple user interface to understand the process of the analysis and plugin data sources. User will be able to understand process of the application without much hassle.

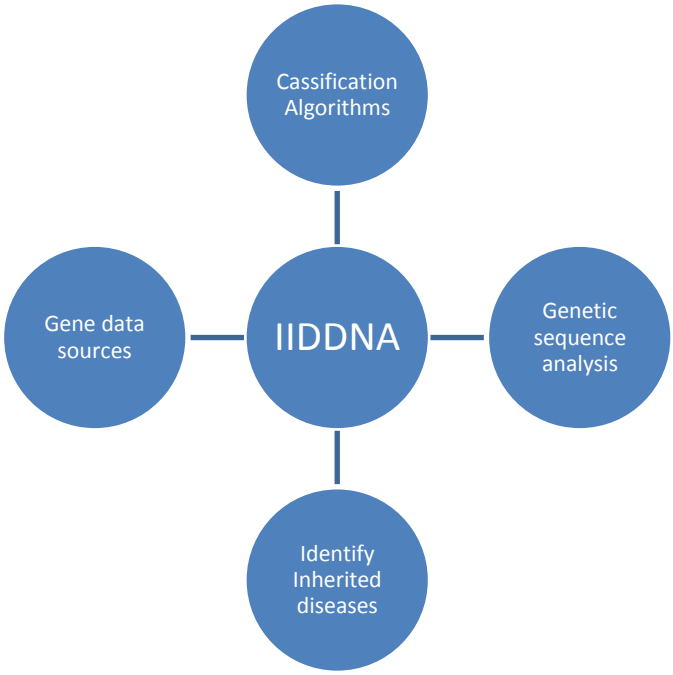

Figure 1 High level view of the IIDDNA

The following illustration shows usage of different features of the Identify inherited diseases based on DNA system.

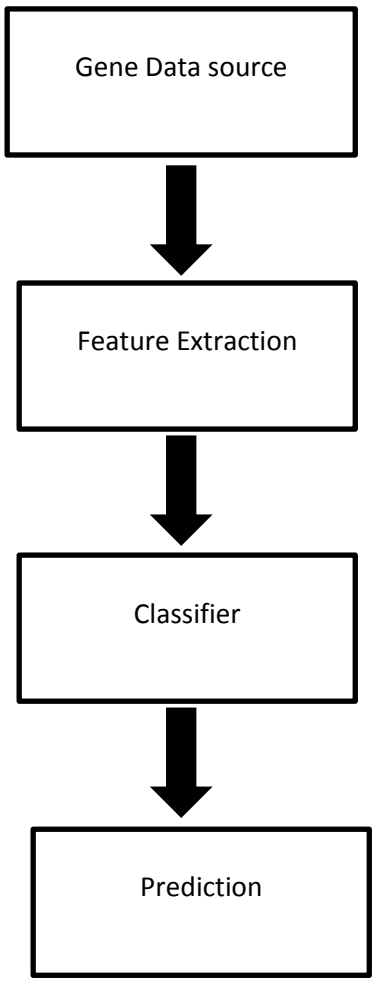

Figure 2 Application flow of the IIDDNA

6. Project Deliverables

- Terms of Reference

  Term of Reference report includes initial project idea. It is clearly state project background, aim, objectives, project deliverables, features of prototype, and activity schedule.

- Literature Review Document

  Literature Review report discuss, and evaluates past work which other researchers work carried out during their researches.

- Requirement Specification Document

  Requirement Specification document is a description of what a system should do. It contains functional and nonfunctional requirements need to develop the prototype.

- Design Specification

  Design Specification is an illustration of high level view of the project architecture

- Prototype

  Prototype is an incomplete version of proposed solution which has only basic features are implemented. This is used to get feedback from users.

- Prototype Report

  Prototype Report explains detailed information of functionality of the prototype.

- Test Report

  Test Report is a report which prototype application is working properly on given inputs. Also includes verification and validation of the basic functionalities.

- Interim Report

  Interim Report analyses how the project is proceeding. The progress of the research and what are the problems encountered, and how project will proceed forward.

- Final Report

  Final Report consists of all the tasks and documents carried out during the project milestones.

7. Resource Requirement

Software Requirements

- MATLAB or Octave
  These are high level languages for prototyping purpose and data visualization. The MATLAB has tool called "BIONIFOMATIC", it gives more features to analysis gene sequencing and prototype bioinformatics related researches.
- R software
  Statistical data analysis software for calculate statistic of the data sources.
  Ex: - Calculate mean distribution of the given data set.

Hardware Requirement

- Dual Core 2.0 GHz CPU
- 2GB DD2 Ram
- 5GB Hard Disk Free Space
- Intel HD Graphics Card

Note – Above requirements are subject to change

8. References

1. Alqallaf, A. K., Tewfik, A. H., Krakowiak, P., Tassone, F., Davis, R., Hansen, R., Hertz-Picciotto, I., Pessah, I., Gregg, J. and Selleck, S. B. (2009) Identifying patterns of copy number variants in case-control studies of human genetic disorders, *GENSIPS 2009. IEEE International Workshop on,* in *Genomic Signal Processing and Statistics, 2009.* pp. 1-4.
2. Ankylosing Spondylitis. Available from: http://www.nlm.nih.gov/medlineplus/ankylosingspondylitis.html. [Accessed 8th October 2014]
3. Cleft Lip and Palate. Available from: http://www.nlm.nih.gov/medlineplus/cleftlipandpalate.html [Accessed 8th October 2014]
4. Dawy, Z., Sarkis, M., Hagenauer, J. and Mueller, J. C. (2008) Fine-Scale Genetic Mapping Using Independent Component Analysis. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 5*(3), pp. 448-460
5. Gong, P., Qu, W. and Feng, D. D. (2005) An Ontology for the Integration of Multiple Genetic Disorder Data So, urces*IEEE-EMBS 2005. 27th Annual International Conference of the*, in *Engineering in Medicine and Biology Society, 2005.* pp. 2824-2827.
6. Hellicar, A. D., Smith, D., Rahman, A., Engelke, U. and Henshall, J. (2014) A hierarchical learning approach to calibrate allele frequencies for SNP based genotyping of DNA pools, *International Joint Conference on*, in *Neural Networks (IJCNN).* pp. 183-189.
7. He, P. and Jiang, R. (2012) Integrating multiple gene semantic similarity profiles to infer disease genes, *31st Chinese*, in *Control Conference (CCC), 2012.* pp. 7420-4725.
8. *How genetic conditions are inherited*, 2014. Available from: http://www.nhs.uk/Conditions/Genetics/Pages/Facts.aspx [Accessed 8th July 2014]
9. Liangcai, Z., Lina, C., Yan, Z., Liangde, X., Yukui, S., Qian, W. and Xia, L. (2008) A SNP and KEGG Based Approach to Mine Risk Pathways Associated with Bipolar Disorder, *ICNC '08. Fourth International Conference on*, in *Natural Computation, 2008.* pp. 34-38.
10. nicolavich, v. (2010) Multifactorial Diseases In Internal Medicine. in,kursk state medical university ,kursk ,russia: kursk state medical university.
11. Pyloric stenosis. Available from: http://www.nlm.nih.gov/medlineplus/ency/article/000970.htm. [Accessed 8th October 2014]

12. Rui, J. and Jiaxin, W. (2011) Integrating sequence conservation features and a domain-domain interaction network to detect disease-associated nsSNPs, *IEEE International Conference on*, in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011.* pp. 262-267.
13. Sam Harris, *The End of Faith: Religion, Terror, and the Future of Reason* [online] Available from: https://www.goodreads.com/quotes/261279-what-are-the-chances-that-we-will-one-day-discover [Accessed 9th September 2014]
14. *What are single nucleotide polymorphisms (SNPs)?.* Available from: http://ghr.nlm.nih.gov/handbook/genomicresearch/snp [Accessed 8th September 2014]
15. What Is Asthma?. Available from: http://www.nhlbi.nih.gov/health/health-topics/topics/asthma/ [Accessed 8th October 2014]
16. What Is Schizophrenia?. Available from: http://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml [Accessed 8th October 2014]
17. Xie, B., Agam, G., Sulakhe, D., Maltsev, N., Chitturi, B. and Gilliam, T. C. (2012) Prediction of candidate genes for neuropsychiatric disorders using feature-based enrichment. In Paper Presented to the Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, Orlando, Florida.

9. Activity Schedules

ID	Task Name	Duration	Start	Finish
1	**Term Of Reference**	**42 days**	**Fri 8/8/14**	**Mon 10/6/14**
2	Indentify Project Scope	19 days	Fri 8/8/14	Wed 9/3/14
3	Identify Aims	11 days	Mon 9/1/14	Mon 9/15/14
4	Indentify Objectives	11 days	Mon 9/1/14	Mon 9/15/14
5	Submission draft Terms Of Reference	0 days	Tue 9/16/14	Tue 9/16/14
6	Evaluation Of draft Terms of Reference	10 days	Tue 9/16/14	Mon 9/29/14
7	Submission of Final Terms of Reference	0 days	Mon 10/6/14	Mon 10/6/14
8	**Literacture Review**	**104 days**	**Fri 8/8/14**	**Wed 12/31/14**
9	Conducting a literacture search	104 days	Fri 8/8/14	Wed 12/31/14
10	Ciriticall analysis the project with existing projects	22 days	Mon 9/1/14	Tue 9/30/14
11	Write Literacture Review report	13 days	Wed 10/1/14	Fri 10/17/14
12	Submission of Literacture Review	0 days	Mon 10/20/14	Mon 10/20/14
13	**Requirement Analysis**	**29 days**	**Wed 10/15/14**	**Sat 11/22/14**
14	Condation online surveys	13 days	Wed 10/15/14	Fri 10/31/14
15	Identify main requirements	12 days	Sat 11/1/14	Sat 11/15/14
16	Writing requirement specification document	6 days	Sat 11/15/14	Fri 11/21/14
17	Submission of requirement specification	0 days	Wed 10/22/14	Wed 10/22/14
18	**Designing the prototype**	**71 days**	**Tue 10/28/14**	**Tue 2/3/15**
19	Design the system architecture	55 days	Tue 10/28/14	Sat 1/10/15
20	Submission on interim report	0 days	Mon 2/2/15	Mon 2/2/15
21	**Implementing the prototype**	**66 days**	**Tue 11/25/14**	**Tue 2/24/15**
22	Find sutable technology to implementation	11 days	Tue 11/25/14	Tue 12/9/14
23	Setting up deveopment enviroment	3 days	Tue 12/9/14	Thu 12/11/14
24	Development	55 days	Tue 12/9/14	Sat 2/21/15
25	Prototype report and demostration	0 days	Mon 2/23/15	Mon 2/23/15
26	**Testing**	**24 days**	**Mon 12/15/14**	**Thu 1/15/15**
27	White Box Testing(Unit Testing)	9 days	Mon 12/15/14	Thu 12/25/14
28	Black Box Testing	16 days	Thu 12/25/14	Thu 1/15/15
29	**Evaluation**	**28 days**	**Thu 1/15/15**	**Sat 2/21/15**
30	Showcasing prototype for potential users	3 days	Thu 1/15/15	Sun 1/18/15
31	Analysing the feedback	3 days	Thu 1/15/15	Sun 1/18/15
32	Enhance Prototype based in the feedback	27 days	Thu 1/15/15	Fri 2/20/15
33	**Final Report**	**187 days**	**Fri 8/8/14**	**Mon 4/27/15**
34	Preparing final report	171 days	Fri 8/8/14	Fri 4/3/15
35	Submission of draft project report	0 days	Mon 4/6/15	Mon 4/6/15
36	Submission of bound copies of final project report	1 day	Mon 4/27/15	Mon 4/27/15

Table 2 Activity schedule

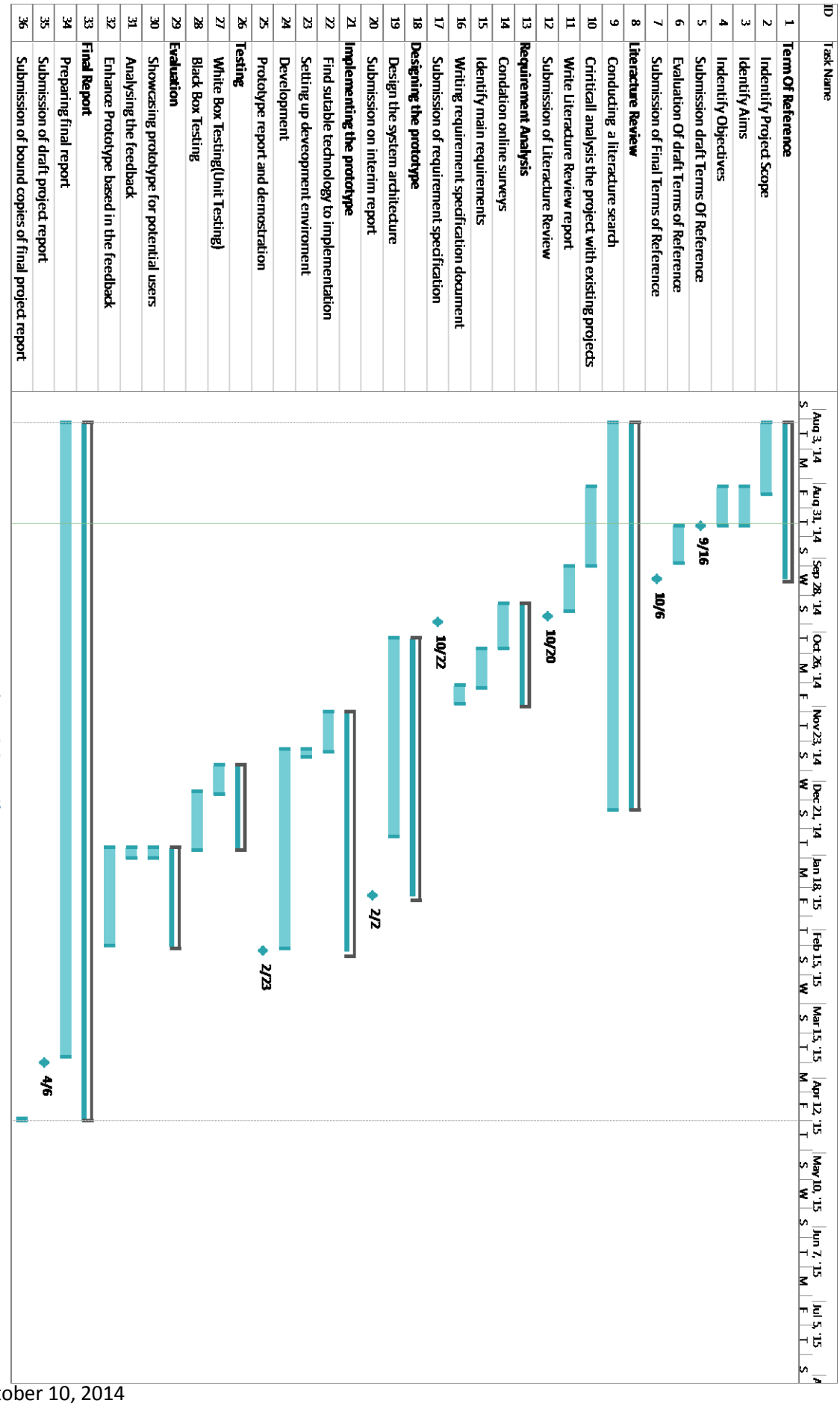

Figure 3 Gantt Chart