INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER (UOW)

BEng/BEng.(Hons) in Software Engineering

Final year Project 2014/2015

Literature Review

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

Document revision history

Date	Description	Author
10/10/14	Initial draft literature review	Lahiru Manohara
15/10/14	Corrected draft literature review	Lahiru Manohara
22/10/14	Final corrected literature review	Lahiru Manohara

Document approvals

Date	Approvals
12/10/14	Mr Achala Aponsu
18/10/14	Mr Achala Aponsu
22/10/14	Mr Achala Aponsu

Table of Contents

1. Introduction

The diseases are main threat of the human life. Inherit factor acts main role of the diseases paradigm. Advancement of technology and science contributes, to identify inherited diseases based on DNA (deoxyribonucleic acid). There are various methods introduces in many researches to identify inherited disease based on DNA such as genetic variations in DNA sequences (Jiaxin et al. 2010), prioritize candidate genes and identify inherited diseases(Xie et al. 2012) and Single nucleotide based genetic mapping method for complex diseases (Dawy et al. 2008).

It is a challenging task to identify inherited diseases without applying bioinformatics techniques and algorithms (Tranchevent et al. 2010).Finding the inherited disease gene from human genome is one of the most major task in bioinformatics research (Xu and Li 2006). Data mining techniques in genetic studies are helps to identify inherited diseases (Fiaschi et al. 2009).

The review is willing to examine following areas, in order to better understanding of the DNA contribution of inherited diseases.

1. SNPs (Single Nucleotide Polymorphism).
2. nsSNPs (Non-synonymous single nucleotide polymorphisms).
3. Gene ontology for inherited genes.
4. Protein-protein interaction network.

Evaluation will be discussed through advantages and disadvantages of previous work to find appropriate solution for identify inherited diseases based on DNA system.

2. SNPs (Single Nucleotide Polymorphism) based approaches for identify inherited diseases

SNPs are genetic variants in single bases of DNA sequence (Jiaxin et al. 2010). Many researchers carried out significant out come on SNPs analysis based approaches. For these researches different traditional classification methods and novel data mining techniques are applied to classification problem. Also explains effective of identifying the disease associate with SNPs and performance of the different algorithms.

Jiaxin et al. (2010), He and Jiang (2012) assert that SNPs in protein coding area are further categorized into synonymous SNPs whose occur will not alter encoded amino acids and nonsynonymous SNPs(nsSNPs) whose present will alter encoded amino acids. In protein level, nsSNPs supplies amino acid substitutions, which potentially cause protein system and functions, and further affects human diseases (Yates and Sternberg 2013). They believed nsSNPs that are associated with the diseases from group of candidate nsSNPs as a binary classification problem. To solve this problem they used different classification algorithms.

Jiaxin et al. (2010) assert that LogitBoost ensemble learning approach achieve the best performance among all the other method such as AdaBoost, Random forest, L2boosting , stochastic gradient regression with two common classification method decision tree, and support vector machines. The five ensemble learning approaches are and two classification methods are briefly explain as follows.

1) **AdaBoost**

   Adaboosting (adaptive boosting) is one of the most popular boosting algorithms. This algorithm characterized by its adaptive changes to fit sample weights during the boosting process, according to the weighted classification error identify from the last training. In each iteration, a sample-based “weak” classifier is constructed, and the weights of each incorrectly classified sample are become grater (or other way, the weights of the each correctly classified sample are become smaller), as a result driving new classifier to focus more on those “hard” samples, Otherwise classifier not correctly classified with previous weakly classified samples. This is working with sequence of weighted samples, and then final classifier defined to be a linear combination of the classifiers from each stage. Adaboost algorithm can be combined with many other learning algorithms to gain its performance as a Meta algorithm. It has been largely used with decision tree is “the best off-the-self classifier in the world” (Vladimir 1992). So this research takes the advantage of the decision tree as the basic “weak” classifier (Jiaxin et al. 2010).

2) **LogitBoost**

   Logiboost is an improved version of the AdaBoost algorithm. The difference between Adaboost and Logiboost is that Logiboost confidence on the binomial log-likelihood as a loss function, which is a more natural basis in binary classification than the exponential basis underlying the Adaboost algorithms. Compared with Adaboost, LogitBoost is more strongly work in noisy problems, easy to implement and does not require tuning and model or kernel selection like neural networks or support vector machines. LogitBoost can work with Logit models, decision stump, or decision trees (Jiaxin et al. 2010).

3) **Random Forests**

   Random forest is a boosting method implemented for vote most popular tree after grown many classification trees. In large data set this method will show outstanding efficiency, few parameters to be adjusted, no over-fitting problem, fast computational speed, and a strong ability of anti-noise characteristic. Also, Random forests have a build in method to estimate the importance of features. This method is usefully to prioritize the features by their importance and reduce the feature set in order to improve the computational complexity (Jiaxin et al. 2010).

4) **L_2booting**

   L_2boosting is gradient boosting for optimizing arbitrary loss functions where component based linier models are make a use of based leaners. It is shown batter performance with stumps (tree with two terminal nodes) and other more common competitors, particularly when the predictor space is multi-dimensional. In addition, L_2boosting is working with both regression and classification problems. It shows competitive performance classification relative problems like LogitBoost (Jiaxin et al. 2010).

5) **Stochastic gradient regression**

Stochastic gradient is regression prediction method. This method uses regression tree as base learner. The optimization of the gradient descent, stochastic gradient regression utilizes the pseudo-residuals result from negative gradient of loss function to set up iterative regression tree. According to idea of Bagging, this algorithm randomly selects part of the pseudo-residual to make regression tree instead of the whole pseudo-residuals. This model can be liner combination of some regression trees (Jiaxin et al. 2010).

6) **Support Vector Machine**

Support Vector Machine also known as Support Vector Network is machine learning method for two-group classification problems. The idea of the SVM is input vector is mapped to infinite dimensional feature space. In this feature space a liner decision environment is build. Special properties of the decision environment ensure high generalization ability of how to learn a machine. The idea behind the support vector network is previously implemented some method for the restricted case where training data can be separated without errors, extend this result to non-separable training data (Cortes and Vapnik 1995).

7) **Decision tree**

Decision tree uses to solve complex decisional problems with significant uncertainty. This method applies to scenarios which specified decision alternative cannot be predicted with certainty, when making a decision a lot of different factor taken as inputs, may be useful to reduce uncertainty in decision by collecting additional information, and the decision maker's behavior is going on to risk taking can impact the relative advantageous of different alternatives (Breiman et al. 1999).

Figure 1 and Figure 2 shows the performance of the algorithms. Four evaluation criteria are accuracy of the prediction (ACC), the area under receiver operating characteristic (ROC) curve (AUC), the balanced error rate (BER) and Mtatthew's correlation coefficient (MCC). Generally, best classification method displays the smaller BER, larger ACC and MCC. The result shows the best method is Logitboost algorithm (Jiaxin et al. 2010).

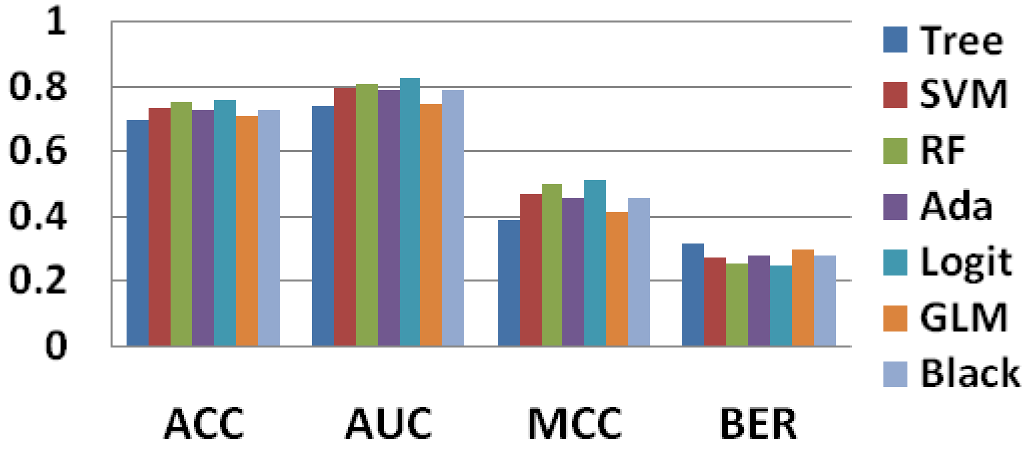


Figure 1. Four evaluation criteria (Jiaxin et al. 2010)

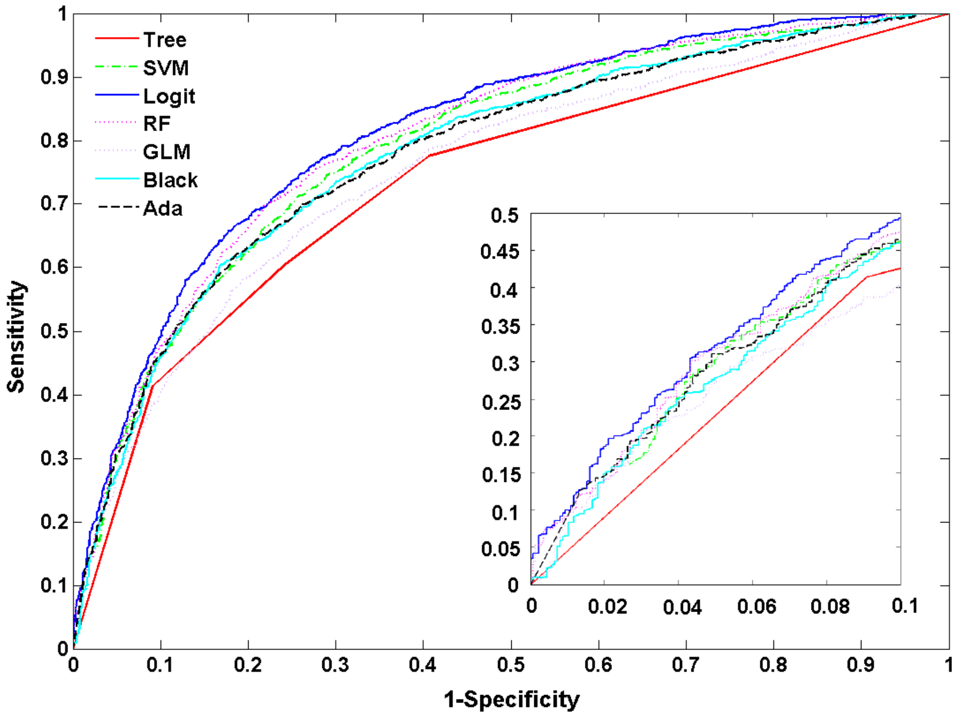


Figure 2. Performance of the 7 approaches (Jiaxin et al. 2010)

Rui and Jiaxin (2011) Explains different approach to find nsSNPs associated with disease. The method is priotizie candidate nsSNPs depends on the integrated use of two sequence conservation features and a domain-domain interaction network. In order to predict possible effects of nsSNPs, bioinformatics method such as PolyPhen (Ramensky et al. 2002), SIFT (Ng and Henikoff 2003), KBAC (Liu and Leal 2010) and MSRV (Jiang et al. 2007) has been proposed. These methods and ensemble learning methods are binary classification methods. But Rui and Jiaxin (2011) believes classification methods provide limited information to beneficiation of diseases associated nsSNPs. And also they argue multiple sequence alignment method is limiting to detect disease associated nsSNPs. But Jiaxin et al. (2010) used multiple sequence alignment method (MAS) to extract conserved protein domains the query protein sequence and purely relying on it. To overcome this limitation, in Rui and Jiaxin (2011) introduce a method, integrate two conservation properties of amino acids and a domain-domain interaction network to calculate association score. Domain-domain interaction method briefly describe as follows.

1. Data source
   The proposed approach depends on the integrated use of tree categories of data 1). Annotations of nsSNPs extracted from the Swiss-Prot database (Consortium 2010), 2) annotation of the protein families and structural domains extracted from Pfam database (Finn et al. 2006), and 3) a domain-domain interaction network obtain from the DOMINE(Raghavachari et al. 2008) and the InterDom database (Ng et al. 2003).

2. Calculation of similarity score between nsSNPs
   Following three similarity scores measure from a pair of two nsSNPs.
   First equation corresponds to probability of occurrence of the original amino acid (P_{org}).

   $$Sim_{org}(a,b) = 1 - \left| p_{org}(a) - p_{org}(b) \right|$$

   Second equation corresponds to probability of occurrence of the substituted amino acid (p_{sub}).

   $$Sim_{sub}(a,b) = 1 - \left| p_{sub}(a) - p_{sub}(b) \right|$$

   Third equation corresponds to calculation of diffusion kernel of the domain-domain interaction network

   $$\mathrm{Sim}_{\mathrm{DDI}}(\mathrm{a,b}) = \mathrm{K}_{\mathrm{DDI}}(\mathrm{a,b}),$$

3. Prioritization of candidate nsSNPs
   After calculating single pairwise similarity measure of nsSNPs apply it to prioritization a set of candidate nsSNPs done according to the “guilt-by-association principle” (Altshuler et al. 2000).Calculation of the association score for a candidate nsSNPs as the mean similarity value between the nsSNP and all seed nsSNPs, as

   $$A(c) = \frac{1}{S(d)} \sum_{s \in S(d)} Sim(c,s)$$

   A(c) is the association score and S(d) the set of seed nsSNPs from query disease d.

4. Integrating of multiple ranks
   Applying “guilt-by-association” method to individual data source, it is given multiple ranking lists. Resorting purpose the Stouffer’s Z-score method for combine the ranks and obtain a single ranking list. Inverse normal transformation of the ranks in the list to contain corresponding Z-score, as

   $$\mathrm{Z}_{\mathrm{i}}^{(\mathrm{K})} = \Phi^{-1}\left(1 - \frac{r_i^{(k)} + 0.5}{\max(r_i^{(k)}) + 1}\right)$$

   Then obtain a combined Z-score by adding up their corresponding Z-score, as

   $$Z_i^{(k)} = \sum_{k=1}^{m} \frac{z_i^{(k)}}{\sqrt{m}}$$

There are certainly several limitations in the proposed proof-of-concept search. This approach currently resort the Pfam database to extract conserved protein domain for candidate nsSNPs because known protein domain variation are limited. To overcome the problem can be use any other sequence alignment tool such as BLAST or PSIBLAST to extract sequence conservation features (Altschul et al. 1997). Rui and Jiaxin (2011) And Jiaxin et al. (2010) both mention in their researches, mutation in other genome regions such as transcriptional-factor binding sites or promoter regions may also cause to human diseases.

Those researches can be concluded as follow. The binary classification solution which is applied to identify diseases associated with nsSNPs, the Logitboost is the more accurate classification algorithm. But it is fully depends on the multiple sequence alignment. Combination of the multiple sequence analysis and domain-domain interaction method is best method to identify nsSNPs associate with diseases. Further enhancement can be added to domain-domain interaction model. There are other data source can be used to evaluate functional similarity between two genes and their products. These data source contain gene expression profile, gene ontology annotations, protein-protein interaction, and many others (Rui and Jiaxin 2011).

3. Combination of SNPs and other methodologies

Fiaschi et al. (2009) assert that SNPs and other clinical contribute for the inherited diseases. The general framework is proposed to solve that problem and apply it find pre-eclampsia, a progressive disorder which occurs during pregnancy and soon after the birth, affecting both the mother and their babies (Roberts et al. 1989). Liangcai et al. (2008) argue mutated risk genes and jointly genetic and environmental factors, which are thought to be key importance (Risch and Merikangas 1996).Proposed method is apply for analysis risk pathway of the bipolar disorder(BD) and proved this method is correct (Hirschfeld et al. 2003). Wu et al. (2014) believes detecting association between human genetic variant and their phenotypic involvement is a significant problem in understanding genetic bases human-inherited diseases. Most of the current system predicts association between nsSNPs and diseases based on features obtain from only protein sequence and/or structure information, and not given details about which specific disease in nsSNPs.

Fiaschi et al. (2009) explains case and control study method. This method is creates two different groups among the population. One group is labeled as ‘case’ (people with diseases or a condition) and the other group is labeled as ‘controls’. Case control analysis in SNPs studies widely used sub-class algorithm of the decision tree. In the study three algorithm taken in consideration: ID3 (Breiman et al. 1999), ADTree (Freund and Mason 1999) , and C4.5 (Quinlan 1993). Figure 3 shows the explanation of the methodology step by step. Wu et al. (2014) argue nsSNP associated diseases are binary classification problem. Because new approached has been introduce to predict association between nsSNPs and diseases based on multiple similarity networks and diseases phenotype similarity networks. The basic assumption of the proposed method is that nsSNPs associated with phenotypically similar diseases would have similar properties. Therefore, some disease and a query nsSNPs, Can be calculate a predictive score that indicates the stability of association between the diseases and the nsSNP by measuring the similarity between the query nsSNP and nsSNPs, that are some diseases identified that have significant phenotypic overlap with the diseases under study.

Further evaluate combination of methodologies, Liangcai et al. (2008) proposed method explains in Figure 4. This approach is different from other two apaches explained before in this literature. Method uses to analysis disease association of the SNPs and environmental factors in the same KEGG pathways. “KEGG pathway is the network of molecular wiring diagrams of interaction and reaction, it is provide a reference knowledge base from linking genome to metabolic processes by mapping genes to REACTION and INTERACTION” (Kanehisa and Goto 2000). There are steps in the calculate diseases risk in particular approach. First step is SNP significance analysis and the corresponding risk calculation. Data are count of case and control for each and calculate risk ratio from relationship between the SNP and complex diseases. Second step is “KEGG pathway reconstruction and the reconstructed network attribution analysis. KEGG pathway is a network where a node represents metabolite and one edge represents some enzyme or a gene cluster”. Third step is “SNPs screening and SNPs to reconstructed network G’ mapping” (Hoh and Ott 2003). Fourth step is “Calculate the two integrated measurements of RS scoring and prioritize the pathway”. This method not only focusing on genetic factors which contain in the study of relationship between multiple genes and the diseases, but also the metabolic environment factors by researching the relationship between genes and pathways. The measurements algorithm of SPAM (A SNP pathway based Association Method) can be summarized as follows.

$$RS(D,P_i)=\sum_{j=1}^{N}\left\{d(GS_j,p_i)*\frac{1}{M}\sum_{\substack{k=1\\ g_k\in GS_j}}\max Risk(g_k,D)\right\}$$

Or

$$RS(D,P_i)=\sum_{j=1}^{N}\left\{d(GS_j,p_i)*\frac{1}{M}\sum_{\substack{k=1\\ g_k\in GS_j}}[1-\min p(g_k,D)]\right\}$$

“where RS(D, p_i) represents the relationship scoring between pathway p_i and the disease D, N is the number of gene clusters on pathway p_i and d(GS_j, p_i) demonstrates the complexity of gene cluster GS_j on pathway p_i. m is the number of the genes that p<0.05. M is the count of all the genes on pathway p_i.”

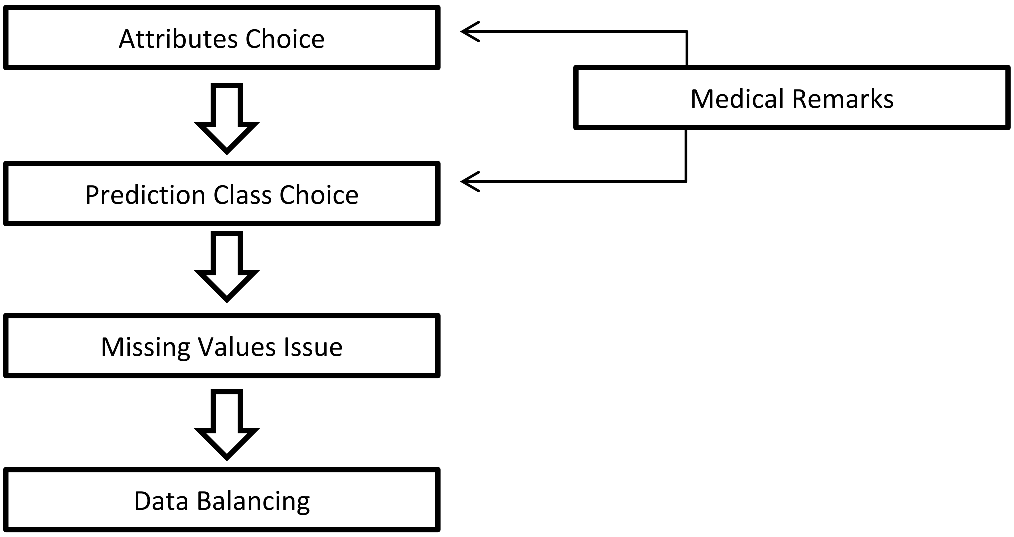


Figure 3. Sequence of steps of following in the pre-processing of a general dataset composed of medical and SNPs attributes. (Fiaschi et al. 2009)

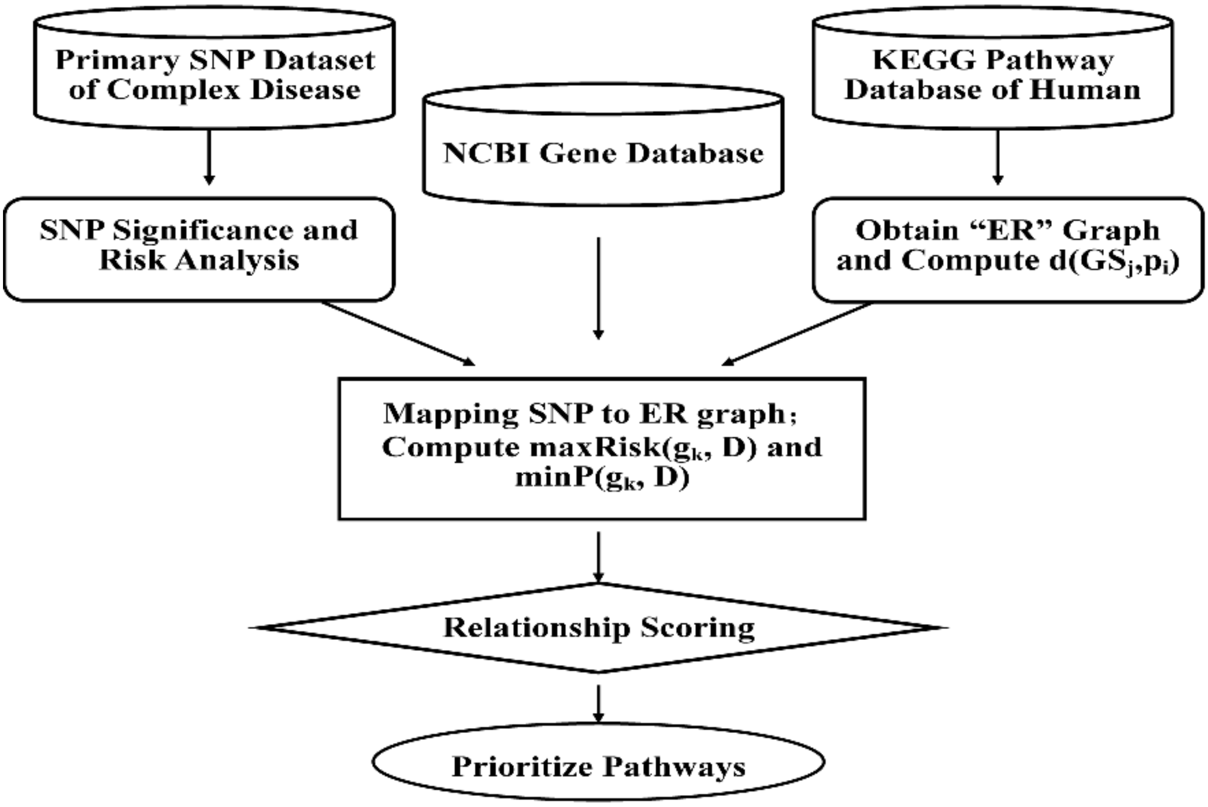

Primary SNP Dataset of Complex Disease
NCBI Gene Database
KEGG Pathway Database of Human
SNP Significance and Risk Analysis
Obtain "ER" Graph and Compute d(GS_j,p_i)
Mapping SNP to ER graph; Compute maxRisk(g_k, D) and minP(g_k, D)
Relationship Scoring
Prioritize Pathways

Figure 4. A SNP pathway based Association Method (SPAM) (Liangcai et al. 2008)

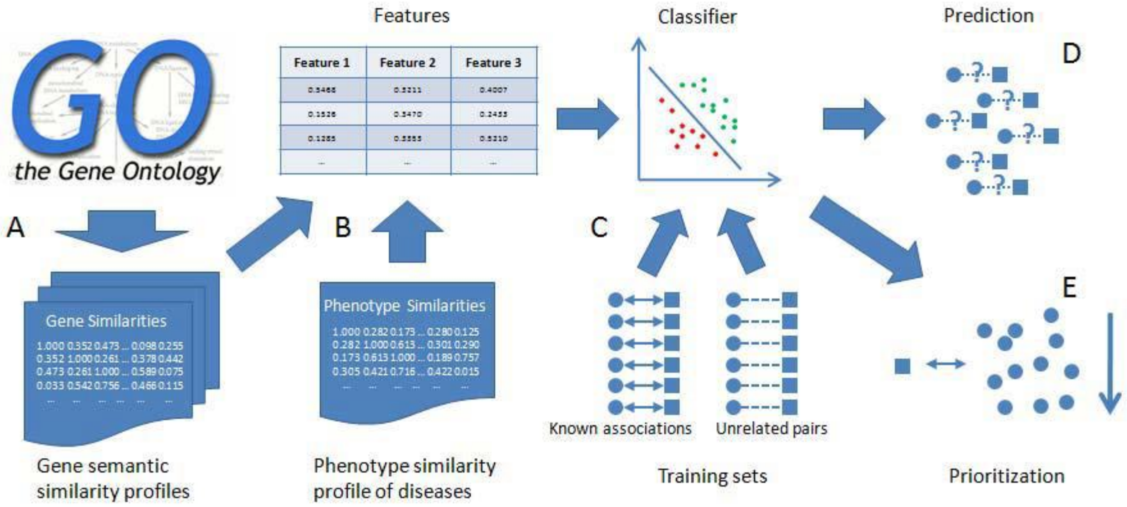
Fiaschi et al. (2009), Wu et al. (2014) and Liangcai et al. (2008) explains in their researches, combination of SNPs analysis and other methods are more effective. Wu et al. (2014) believes proposed approach can used on both disease with partially known genetic based or unknown genetic bases. Also explains some limitation on their research, only getting nsSNPs data from one data based for analysis query protein sequence. To overcome this limitation will use other multiple sequence alignment methods, such as PSI-BLAST (Altschul et al. 1997) or COBOLT (Papadopoulos and Agarwala 2007). Feature selection method is not implementing in this approach. So further study, can be implement a feature selection method to gain the performance of the proposed solution. Fiaschi et al. (2009) suggests family history related data helps to identify and gain performance of the proposed solution. And also agree with Wu et al. (2014), the feature selection is a affect their solutions. Liangcai et al. (2008) and Wu et al. (2014) both mention in their researches some other genome region such as the transcriptional and promoter regions may also be associated with human diseases.

Those researches can be concluded as follows. The combination methods are express and explain more about inherited diseases other than analyzing only SNPs. KEGG pathway is useful to analysis other environment factors with SNPs data. The phenotype data is explains the more about given genotype data associated with disease. But there are some point in these researches can be affect to performance and accuracy of the proposed solution. Fiaschi et al. (2009) mentions their solution transforming categorical data to Boolean one according to the threshold value. Selection of the threshold can be huge effect to the accuracy of the solution. Wu et al. (2014) mentions in the proposed solution Canberra distance (Emran and Ye 2001) algorithm is the best to calculate distance between pairwise nsSNPs. But some situation same distance in two similar pair can be in data set and it is hard to classify which is the nearest one to associate with diseases.

4. Gene prioritization and identify inherited diseases.

Xie et al. (2012) believes prioritizing candidate genes and finding the best candidate whose gene contributing to disease is most important problems of genomics. Finally, proposed method uses identify common heritable disorders such as autism (Kim et al. 2000), schizophrenia (Jingchun et al. 2008) and diabetes. He and Jiang (2012) explains binary classification formulation proposed in their approach has been used to prioritize the candidate genes. Tranchevent et al. (2010) mentions identifying the most relevant candidate genes is a challenging and time consuming task. Most of the time a biologist would have to go manually through the list of candidates and need to find it is relevant or not. There are 19 computational approaches has been evaluated by literature review. Some approaches are related to identify inherited disease from prioritizing the candidate genes methods as well. A similarity between all these methods are used 'guilt-by-association' concept. "The most relevant candidates will be the ones that are similar to the genes already known to be linked to the biological process of interest" (Smith and Eyre-Walker 2003), (Goh et al. 2007), (Jimenez-Sanchez et al. 2001).

Xie et al. (2012) proposed a method to prioritize candidate disease gene based on Support Vector Machine (SVM) and ontology association. Existing feature-based gene prioritization tool uses multiple data types for gene annotation and feature extraction. This approach is based on the generally used assumption that genes contributed with similar phenotypes tend to have similar features. The steps of analysis are enrichment analysis and feature extraction, feature-based gene prioritization, and statistical validation of the result of analysis. But He and Jiang (2012) argues identify disease gene from the angle of prioritizing candidate gene can give only a sorted list of the candidate genes. It is hard to count whether highest ranked genes are most relevant with the diseases or not. This method is useful when small set of positive samples about relationships between genes and the query diseases, usually ignores such information that a large amount of genes are not contribute with any disease, as a result can be used as negative samples. To solution for these limitations, binary classification uses to formulate the question of identify disease genes. More specifically, He and Jiang (2012) argue that genes and related diseases can be used as positive samples, and disease gene pair sampled at random can serve negative samples. The overview of proposed solution shows in Figure 5. Tranchevent et al. (2010) argues predicting the function of a gene or its involvement in a genetic condition are interrelated problems. Further explain, several gene function prediction methods has certainly been applied to disease gene prioritization with enough performance (Zhang et al. 2006). According to Tranchevent et al. (2010) "knowledge, none of the existing gene function prediction methods includes disease-specific data". Figure 6. Shows general methodology of the prioritizing candidate genes and Table 1. "Shows an overview of the input data required by the tools as well as the output produce".



Feature 1	Feature 2	Feature 3
0.3466	0.5211	0.4007
0.1526	0.3470	0.2433
0.1285	0.5555	0.5210
...	...	...

Figure 5. Work flow of the proposed method (He and Jiang 2012).

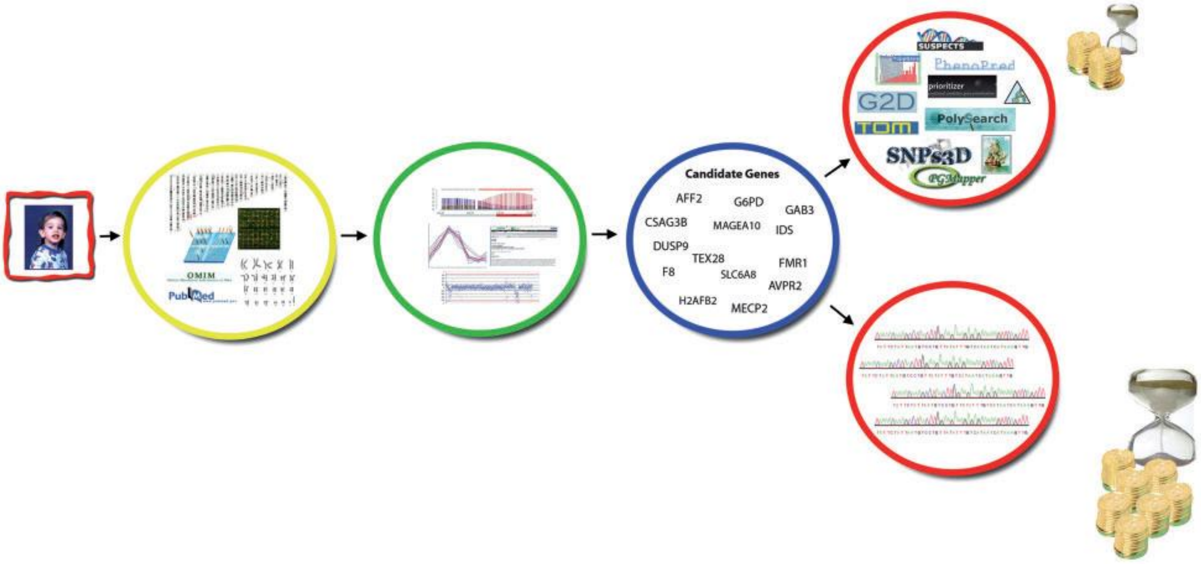
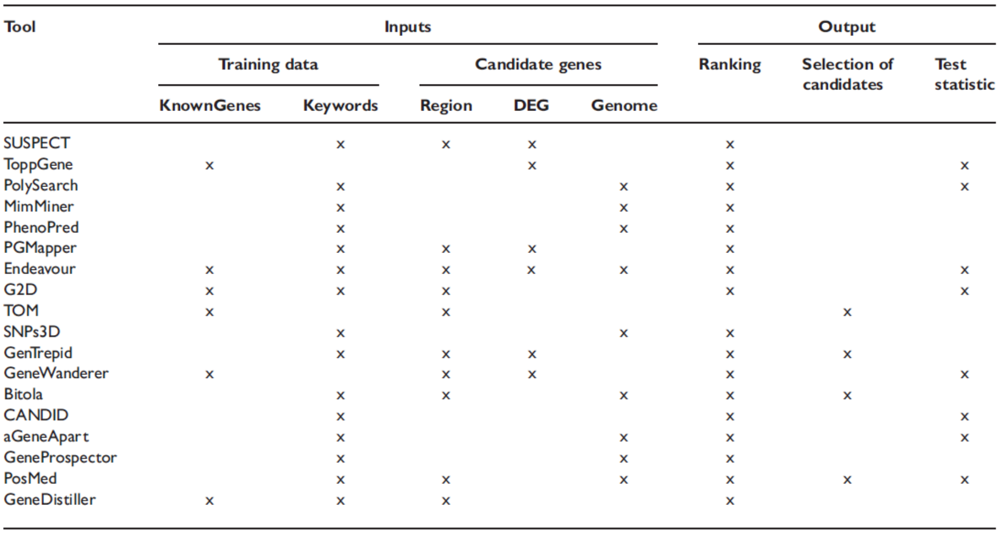


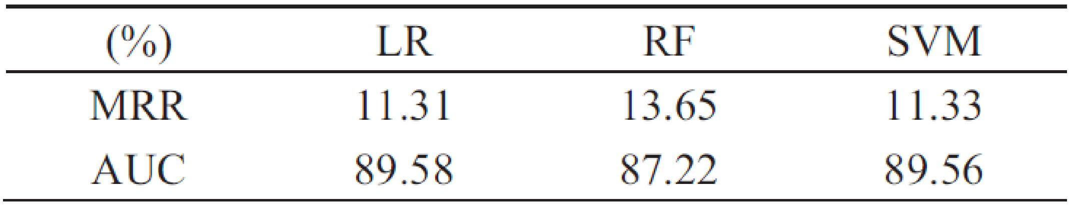
Figure 6. Overview of gene prioritization tools (Tranchevent et al. 2010).



Tool	Inputs					Output		
	Training data		Candidate genes			Ranking	Selection of candidates	Test statistic
	KnownGenes	Keywords	Region	DEG	Genome			
SUSPECT		×	×	×		×		
ToppGene	×			×		×		×
PolySearch		×			×	×		×
MimMiner		×			×	×		
PhenoPred		×			×	×		
PGMapper		×	×	×		×		
Endeavour	×	×	×	×	×	×		×
G2D	×	×	×			×		×
TOM	×		×				×	
SNPs3D		×			×	×		
GenTrepid		×	×	×		×	×	
GeneWanderer	×		×	×		×		×
Bitola		×	×		×	×	×	
CANDID		×				×		×
aGeneApart		×			×	×		×
GeneProspector		×			×	×		
PosMed		×	×		×	×	×	×
GeneDistiller	×	×	×			×		

Table 1. Description of the inputs needed by the tools and the outputs produced by the tools (Tranchevent et al. 2010).

According to He and Jiang (2012) result comparing different algorithms such as Liner Regression, Random Forest and Support Vector Machine. According to Table 2. result suggest the effectiveness of the proposed method in the prioritization candidate genes, according to Mean rank ratio (MRR) and Area under the ROC curve (AUC) logistic regression and support vector machine can achieve higher performance than the random forest. Xie et al. (2012) also agree support vector machine based gene prioritizing is preform significant average accuracy. Support vector machine is working on the unbalance training data set as well. Tranchevent et al. (2010) comparison study about gene prioritizing tool that are available. These tools are following various type of input, data source used during the process of prioritization and the desired output. Tis study recommends tool called G2D to find the relation to inherited disease a combination of data mining on biomedical databases and gene sequence analysis (Perez-Iratxeta et al. 2002).



(%)	LR	RF	SVM
MRR	11.31	13.65	11.33
AUC	89.58	87.22	89.56

Table 2. Result of 10-fload cross validation experiments for prioritizing candidate genes (He and Jiang 2012).

Those researches can be concluded as follow. Xie et al. (2012) and He and Jiang (2012) both agreed with performance of the support vector machine to apply gene prioritizing process. Also agreed with prioritization method and other methods such as similarity profile of human disease analysis and enrichment based feature selection process perform well in identify inherited diseases. He and Jiang (2012) believes feature selection method and integrating more data sources uses with proposed solution can build a better prediction ability. Tranchevent et al. (2010) comparison study is more important to classify which tool is most relevant to research about gene prioritization areas such as identifies inherited diseases. Also help to identifies type of input, data source used during the process of prioritization and the desired output.

5. Protein – Protein interaction network and inherited disease

Xu and Li (2006) believes the availability of human genome-wide protein-protein interaction (PPIs) open a path for discovering inherited disease genes by topological features in PPIs network. Jingchun et al. (2008) agree to protein-protein interaction network useful to identify inherited diseases called Schizophrenia (*What Is Schizophrenia?* 2014). Furthermore, studies of proteins and their interaction are important to understand their dynamic roles within a cell. Mapping the schizophrenia genes into the whole human interaction network and then extract their sub networks make clear on the cellular mechanisms and biological processes related to the inherited disease. Xu and Li (2006) argues “exploiting sequence based feature and found that for many of them there are significant differences between genes underlying human hereditary disease and those not known to be involved in diseases”. “Associated with a human disease preferentially interacted with other disease-causing genes significantly indicating heritable disease-genes might share some topological features in the PPIs network, whereas the non-disease genes do not” (Gandhi et al. 2006).

The proposed method steps are human PPIs dataset were downloading from Online Predicted Human Integration Database (OPID) (Brown and Jurisica 2005), building training samples, define features set, classification algorithm and validation. K-nearest neighbors (KNN) algorithm used for classification. The KNN algorithm is a simple but powerful non parametric classification algorithm (Franke et al. 2006). The performance of the KNN classifier shows in Figure. According proposed method “if a gene shares more neighbors with known hereditary disease-genes in the PPIs network is should be more likely be a disease-gene too” (Xu and Li 2006). Jingchun et al. (2008) used the same method which (Xu and Li 2006) proposed in their research. According to KNN based approach, Jingchun et al. (2008) find the properties of schizophrenia gene in human protein-protein network.

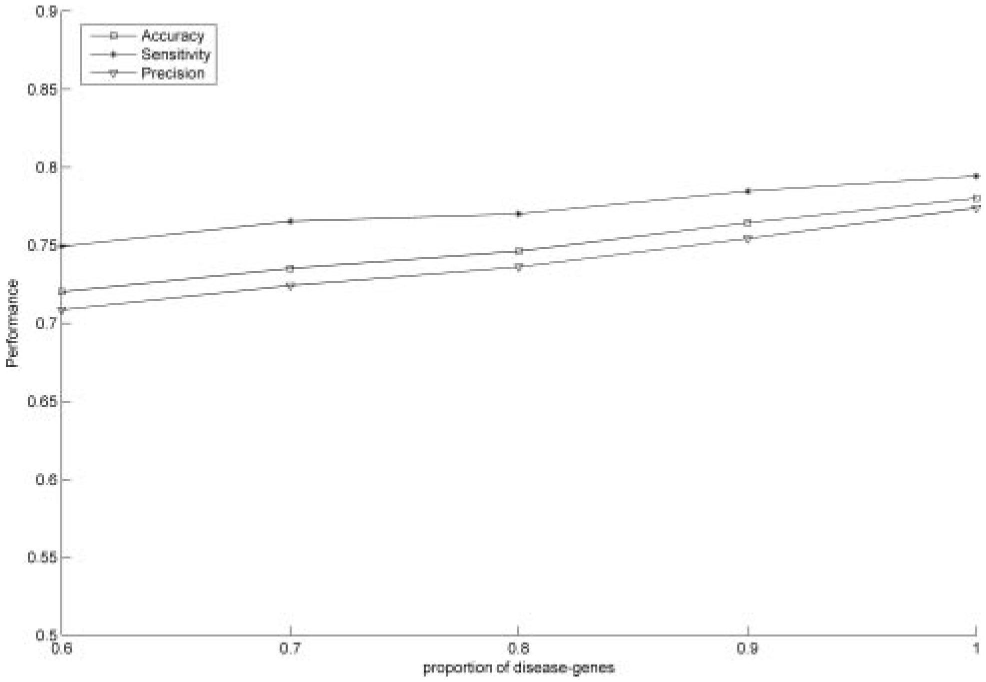


Figure 7.Performance of KNN classifier with adding new disease-genes (Xu and Li 2006).

Those researches can be concluded as follow. PPI network is mostly important for identify heritable disease. Jingchun et al. (2008) used PPI network method for identify schizophrenia inherited diseases. Xu and Li (2006) “proposed a KNN based classifier to explicitly and directly exploit the recent PPIs data”. This proposed method work with different human PPIs and achieved higher accuracy. “With increasing quantity and quality of human interaction and phenotypic data, the performance and utility of this approach in facilitating biologists to detect novel disease-genes should improve even further”.

6. Literature Summary

Most of the inherited diseases are based on the DNA variation called SNPs, gene which segment contain in DNA and protein-protein interaction network. Most of the researches are in inherited disease domain related to SNPs (Jiaxin et al. 2010),(Wu et al. 2014). Combination of the SNPs and the other methodologies is useful to identify the inherited diseases (Fiaschi et al. 2009), (Liangcai et al. 2008),(Rui and Jiaxin 2011). Gene prioritization is a another best approach for the identify inherited diseases (He and Jiang 2012), (Xie et al. 2012). Human protein-protein interaction (PPIs) is a new opportunity for discovering inherited diseases (Xu and Li 2006), (Jingchun et al. 2008).

Most of the identify SNPs related to inherited diseases are binary classification problem (Rui and Jiaxin 2011), (Jiaxin et al. 2010), (Wu et al. 2014). Jiaxin et al. (2010) proved in their research Logitboost is best method to classify SNPs related to inherited disease. But some other most commonly used algorithms are Support Vector Machine, Decision tree and the Random forest (Fiaschi et al. 2009).Gene prioritization methods are used scoring based mathematical model to calculate which genes are truly associate with the inherited diseases (He and Jiang 2012), (Tranchevent et al. 2010). KNN classifier is most significant algorithm for analysis protein-protein interaction network based inherited diseases (Xu and Li 2006).

Some researches were used different cross validation method to prove accuracy, efficiency and the appropriation of the proposed method. Some proposed method were implemented as general framework for identify inherited diseases (Liangcai et al. 2008), (Fiaschi et al. 2009). Researchers are also proposed some method for identifying specific inherited diseases such as neuropsychiatric disorders and Schizophrenia (Jingchun et al. 2008), (Xie et al. 2012). General framework based methods and other specific type inherited diseases exploration system are valuable for further study of the Identify Inherited Diseases based on DNA system(IIDDNA).

7. Bibliography

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25*(17), pp. 3389-3402.

Altshuler, D., Daly, M. and Kruglyak, L. (2000) Guilt by association. *Nat Genet, 26*(2), pp. 135-137.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1999) *Classification and Regression Trees,* CRC Press, New York.

Brown, K. R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics, 21*(9), pp. 2076-2082.

Consortium, T. U. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research, 38*(suppl 1), pp. D142-D148.

Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning, 20*(3), pp. 273-297.

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.

Emran, S. M. and Ye, N. (2001) Robustness of canberra metric in computer intrusion detection. in *Proc. IEEE Workshop on Information Assurance and Security, West Point, NY, USA*: Citeseer.

Fiaschi, L., Garibaldi, J. M. and Krasnogor, N. (2009) A framework for the application of decision trees to the analysis of SNPs data. in *Computational Intelligence in Bioinformatics and Computational Biology, 2009. CIBCB '09. IEEE Symposium on*. pp. 106-113.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Research, 34*(suppl 1), pp. D247-D251.

Franke, L., Bakel, H. v., Fokkens, L., De Jong, E. D., Egmont-Petersen, M. and Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics, 78*(6), pp. 1011-1025.

Freund, Y. and Mason, L. (1999) The alternating decision tree learning algorithm. in *ICML*. pp. 124-133.

Gandhi, T., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S. and Periaswamy, B. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics, 38*(3), pp. 285-293.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabási, A.-L. (2007) The human disease network. *Proceedings of the National Academy of Sciences, 104*(21), pp. 8685-8690.

He, P. and Jiang, R. (2012) Integrating multiple gene semantic similarity profiles to infer disease genes. in *Control Conference (CCC), 2012 31st Chinese*. pp. 7420-4725.

Hirschfeld, R. M. A., Lewis, L. and Vornik, L. A. (2003) Perceptions and impact of bipolar disorder: How far have we really come? Results of the National Depressive and Manic-Depressive Association 2000 survey of individuals with bipolar disorder. *Journal of Clinical Psychiatry, 64*(2), pp. 161-174.

Hoh, J. and Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat Rev Genet, 4*(9), pp. 701-709.

Jiang, R., Yang, H., Zhou, L., Kuo, C. C. J., Sun, F. and Chen, T. (2007) Sequence-Based Prioritization of Nonsynonymous Single-Nucleotide Polymorphisms for the Study of Disease Mutations. *The American Journal of Human Genetics, 81*(2), pp. 346-360.

Jiaxin, W., Wangshu, Z. and Rui, J. (2010) Comparative study of ensemble learning approaches in the identification of disease mutations. in *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on*. pp. 2306-2310.

Jimenez-Sanchez, G., Childs, B. and Valle, D. (2001) Human disease genes. *Nature, 409*(6822), pp. 853-855.

Jingchun, S., Leng, H. and Zhongming, Z. (2008) Schizophrenia Genes: Characteristics of Function and Protein Interaction Networks. in *BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on*. pp. 437-441.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research, 28*(1), pp. 27-30.

Kim, J. A., Szatmari, P., Bryson, S. E., Streiner, D. L. and Wilson, F. J. (2000) The Prevalence of Anxiety and Mood Problems among Children with Autism and Asperger Syndrome. *Autism, 4*(2), pp. 117-132.

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. in *Natural Computation, 2008. ICNC '08. Fourth International Conference on*. pp. 34-38.

Liu, D. J. and Leal, S. M. (2010) A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet, 6*(10), pp. e1001156.

Ng, P. C. and Henikoff, S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research, 31*(13), pp. 3812-3814.

Ng, S.-K., Zhang, Z., Tan, S.-H. and Lin, K. (2003) InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Research, 31*(1), pp. 251-254.

Papadopoulos, J. S. and Agarwala, R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics, 23*(9), pp. 1073-1079.

Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2002) Association of genes to genetically inherited diseases using data mining. *Nat Genet, 31*(3), pp. 316-319.

Quinlan, J. R. (1993) *C4. 5: programs for machine learning,* Morgan kaufmann.

Raghavachari, B., Tasneem, A., Przytycka, T. M. and Jothi, R. (2008) DOMINE: a database of protein domain interactions. *Nucleic Acids Research, 36*(suppl 1), pp. D656-D661.

Ramensky, V., Bork, P. and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Research, 30*(17), pp. 3894-3900.

Risch, N. and Merikangas, K. (1996) The Future of Genetic Studies of Complex Human Diseases. *Science, 273*(5281), pp. 1516-1517.

Roberts, J. M., Taylor, R. N., Musci, T. J., Rodgers, G. M., Hubel, C. A. and McLaughlin, M. K. (1989) Preeclampsia: An endothelial cell disorder. *American Journal of Obstetrics & Gynecology, 161*(5), pp. 1200-1204.

Rui, J. and Jiaxin, W. (2011) Integrating sequence conservation features and a domain-domain interaction network to detect disease-associated nsSNPs. in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*. pp. 262-267.

Smith, N. G. C. and Eyre-Walker, A. (2003) Human disease genes: patterns and predictions. *Gene, 318*(0), pp. 169-175.

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P. and Moreau, Y. (2010) A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*.

Vladimir, V. (1992) Principles of Risk Minimization for Learning Theory. pp. 831--838.

What Is Schizophrenia?, (2014) Available: http://www.nimh.nih.gov/health/topics/schizophrenia/index.shtml 2014].

Wu, J., Yang, S. and Jiang, R. (2014) Inferring non-synonymous single-nucleotide polymorphisms-disease associations via integration of multiple similarity networks. *Systems Biology, IET, 8*(2), pp. 33-40.

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.

Xu, J. and Li, Y. (2006) Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics, 22*(22), pp. 2800-2805.

Yates, C. M. and Sternberg, M. J. E. (2013) The Effects of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs) on Protein–Protein Interactions. *Journal of Molecular Biology, 425*(21), pp. 3949-3963.

Zhang, P., Zhang, J., Sheng, H., Russo, J. J., Osborne, B. and Buetow, K. (2006) Gene functional similarity search tool (GFSST). *BMC bioinformatics, 7*(1), pp. 135.