

國立交通大學

統計學研究所

碩士論文

藉由粒線體相關疾病之基因型及表徵型的網路
分析來預測疾病基因

Predict Candidate Genes by Network Analysis of
Genotypes and Phenotypes for Mitochondrion Diseases

研究生：陳俊睿

指導教授：盧鴻興 博士

中華民國九十六年六月

藉由粒線體相關疾病之基因型及表徵型的網路
分析來預測疾病基因

Predict Candidate Genes by Network Analysis of
Genotypes and Phenotypes for Mitochondrion Diseases

研究生：陳俊睿

Student : Chun-Jui Chen

指導教授：盧鴻興

Advisor: Henry Horng-Shing Lu

國立交通大學

統計學研究所

碩士論文

A Thesis

Submitted to Institute of Statistics

College of Science

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2007

Hsinchu, Taiwan, Republic of China

中華民國九十六年六月

藉由粒線體相關疾病之基因型及表徵型的網路 分析來預測疾病基因

研究生：陳俊睿

指導教授：盧鴻興 博士

國立交通大學統計學研究所



在遺傳疾病的研究上，基因與疾病之間的關係是我們感興趣的。其中我們更感興趣的是，是否仍有一些基因與特定疾病的關係被隱藏起來而未被發現或驗證。然而盲目的透過生物實驗的方法一一檢驗證其他的基因與疾病關連性，不僅曠日耗時，更需大量的金錢。根據文獻的回報，我們建立網路來連結各個基因與疾病。而從文獻裡，我們亦可估算出兩者之間的機率，藉以完成一個基因與疾病的網路，並建構演算法選出可能的致病基因。最後再利用交叉比對，決定模型的把關條件並且證明這個模型對於選取致病基因的可行性。

Predict Candidate Genes by Network Analysis of Genotypes and Phenotypes for Mitochondrion Diseases

Student : Chun-Jui Chen Advisor : Dr. Henry Horng-Shing Lu

Institute of Statistic
National Chiao Tung University



ABSTRACT

In the study of heritable diseases, we are interested in the relationship between genes and diseases. What we are more concerned about is if there are some hidden relationships which were not validated in literature reports. However, it costs time and money to clarify them one by one through biologic experiments. According to literature reports, we can not only build a network to connect genes and diseases, but also estimate the probability of this network. By this network, we can predict candidate genes which also cause diseases and are not observed. Finally, cross validation studies are carried out to decide thresholds of models and evaluate the performance of our methods proposed in the article. The results show that these new methods are promising.

Key words: Disease Genotype-Phenotype Network, Bayesian Network, Noisy OR Model, Candidate Genes, Leave-one-out Cross Validation, Mitochondrion Diseases.

誌謝

首先感謝盧老師這兩年多來的照顧，沒有老師的耐心的指導，我沒有辦法順利的完成這篇文章；並且讓我在碩士班期間有機會到芝加哥大學交流，一探更高的學問殿堂。也要感謝應數系的許元春教授，沒有他在我徬徨時給予我幫助，就沒有現在的我。感謝統計所所長陳鄰安教授，在我出國前還替我張羅補助，跟相關的事項。最後感謝所有統計所老師的教導，跟所辦郭小姐及溫先生的幫忙，讓我在統計所期間能有一個充實、美好的學習環境。

再來感謝我的女朋友，佳蕊，謝謝妳這六年來的陪伴。沒有妳的支持我沒有力量走下去，我知道今後妳不會在我的生命中缺席，妳的體貼、溫柔都是我得以有所成就的助力。還有我的父母、姐姐、弟弟，雖然我一個人隻身在外，但你們對我的關心讓我不論身在何處都不會感到孤單。

我會永遠記得408，409的午餐時間，記得和永在、阿Q、阿淳、益銘、建威、柯董在午後的球場，和小米互相吐嘈，和侑峻分享心事，和益通睡前的話家常，和泰賓學長的暢飲，409眾美女的廚藝，408眾美女的牌技，打game的夜晚，男人幫的燒烤……，有太多美好的回憶了。隻字片語道不盡我心中對大家的不捨，離別在即，還是得說再見，希望大家各奔前程之後都能在各自的領域嶄露頭角。期待他日共剪西窗燭，再話交大夜雨時。

在此，僅以此篇論文獻給所有關心我的人。



陳俊睿 謹誌于
國立交通大學統計研究所
中華民國九十六年六月

Content

Chapter 1. Introduction	1
1.1 Motivation	1
1.2 Data format	2
Chapter 2. Bayesian networks	3
2.1 Definition of Bayesian networks	3
2.2 Constructing Bayesian networks	3
2.3 Noisy OR.....	5
Chapter 3. Algorithms for candidate gene prediction.....	7
3.1 Constructing disease genotype-phenotype network.....	7
3.2 Quickscore	8
3.3 Algorithms for prediction of candidate genes	10
Chapter 4. Leave-one-out Cross Validation	15
4.1 Steps for leave-one-out cross validation.....	15
4.2 Evaluating performances of two algorithms	16
4.3 Threshold	19
Chapter 5. Conclusion and Discussion	25
References	26

Chapter 1. Introduction

1.1 Motivation

We all know that there are strong relationships between genotypes and phenotypes. The diseases we are interested in are resulted from several genes mutate or express abnormally. But exploring these appearances is depending on tremendous biologic experiments which cost lots of money and time. However, we want to develop a cheaper and easier method through a great deal of literature reports.

Using Bayesian network has several advantages. First of all, it can represent how we infer from this data by graphic structure. The network structure is built through the causality determined by domain knowledge. In our case, we believe that variety of genotypes caused different phenotypes.

Furthermore, Bayesian network is a well tool to predict unknown events by new evidences. For example, once we obtain information about patients' heritable diseases, it is possible to predict candidate genes which might cause the patients get those diseases.

In our case, there are eleven diseases, deficiencies, about mitochondrion. According to reported literatures, we have the associations between genes and diseases, including those deficiencies. Besides, we further wonder if there are still other genes which are relating with deficiencies, but not reported in these literatures. We will use these data to construct a network, and then design algorithms for detecting those hidden genes.

Finally, we do leave-one-out cross validation that we will take one relationship each time to predict the relationship took out. Through this study, we can decide thresholds and evaluate the performances of algorithms.

1.2 Data format

GENE	FEATURE	PMID/GENE-FEATURE	PMID-GENE
DLD	AKDH-deficiency	6	16
DLST	AKDH-deficiency	3	4
OGDH	AKDH-deficiency	2	3
AASS	Dehydration	1	5

Our data format looks like above table. There are 9407 relationships in this format.

GENE: there are total 174 genes in our file.

FEATURE: there are 502 features including 11 special features associating with mitochondrion.

PMID/GENE: PMID is an acronym for PubMed Identifier which is a unique number assigned to each PubMed citation. And PMID/GENE means the number of articles associating with a specific gene.

PMID/GENE-FEATURE: This means the number of articles associating with a specific relationship between genes and features.

There are two files with the same format. One File-ALL is about all genes and features; another File-PD is about only the deficiency genes and features (The data files are collected by Dr. Curt Scharfe at Stanford University.)

Chapter 2. Bayesian networks

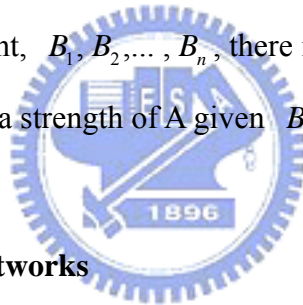
2.1 Definition of Bayesian networks

Not every network belongs to Bayesian network. There should be some properties which Bayesian networks are composed of (Finn V. Jensen, 2001).

Definition 1.1

Bayesian networks should consist of the followings:

1. A set of random variables and a set of directed edges between variables.
2. Each variable has a finite set of mutually exclusive states.
3. The variables with directed edges form a directed acyclic graph, what we call DAG. (“Acyclic” means there is no directed path $A_1 \rightarrow A_2 \rightarrow \dots \rightarrow A_n$ in which $A_1 = A_n$)
4. To each variable A with parent, B_1, B_2, \dots, B_n , there is a attached conditional probability $P(A | B_1, \dots, B_n)$, which means a strength of A given B_1, B_2, \dots, B_n .



2.2 Constructing Bayesian networks

Constructing a Bayesian network should follow some rules concerned with the structures and parameters. Those rules are also the typical characters of Bayesian networks (Finn V. Jensen, 2001).

Intuitively, we construct networks by causalities which mean the procedures of inferring events. We can represent the procedures of inferences in Bayesian networks by using structures of d-separation.

For example, we want to construct the relationship between pregnancy, hormonal state and urine test. Generally, pregnancy will affect the hormonal state, and then hormonal state further has a impact on the urine test. Once we know the hormonal state, condition of pregnancy won't influence the result of urine test any more. So the network representing the relationship

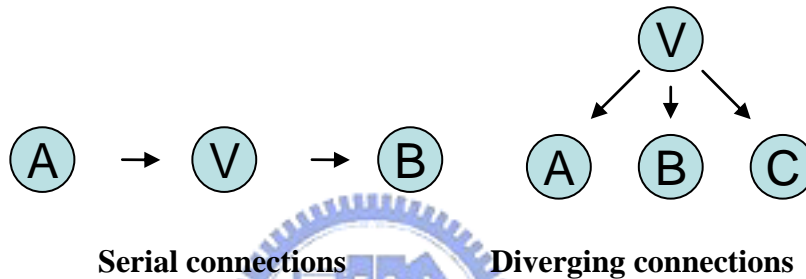
is:



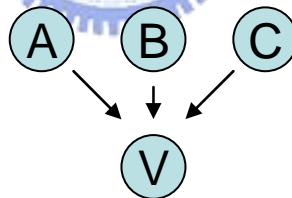
Definition 2.1(d-separation)

Two distinct variables A and B in a directed network are d-separated if all paths between A and B have an intermediate variable V such that either:

1. The connection is serial or diverging and V has received evidences.



2. The connection is converging, and neither V nor any of V's descendants have received evidences.



Converging connections

According to this definition, if A and B are d-separated, then changing the certainty of A can't affect the certainty of B. Combining d-separated with conditional independence, we can reduce the parameters in Bayesian network.

For example,



$A: \{a_1, a_2\}$, $B: \{b_1, b_2\}$ and $C: \{c_1, c_2\}$. If we don't use d-separated, then $P(C|A, B)$

includes 8 situations. Through involving d-separated, $P(C | A, B) = P(C | B)$, it will only include 4 situations. Furthermore, we can apply this character to calculate the joint probability by Theorem 3.2.

Theorem 2.2 (chain rule)

Let a Bayesian network be over $U = \{A_1, A_2, \dots, A_n\}$. Then, we can get the joint probability $P(U)$ which is the product of all probabilities specified in this Bayesian network.

$$P(U) = \prod_i P(A_i | pa(A_i))$$

Where $pa(A_i)$ is the parent set of A_i .

2.3 Noisy OR

Bayesian networks require conditional probabilities as their parameters. If each variable has two values, and one variable has m parents, then this variable requires a 2^m conditional probability. The larger m is, the more difficult computation becomes. Moreover, even m is small, information about the conditional probability in which one variable is given m variables is difficult to obtain. It is much easier to get the conditional probability in which one variable is just given other one. So, noisy OR model can not only reduce the computational complexity, but need information which is much easier to get (Richard E. Neapolitan, 2004).

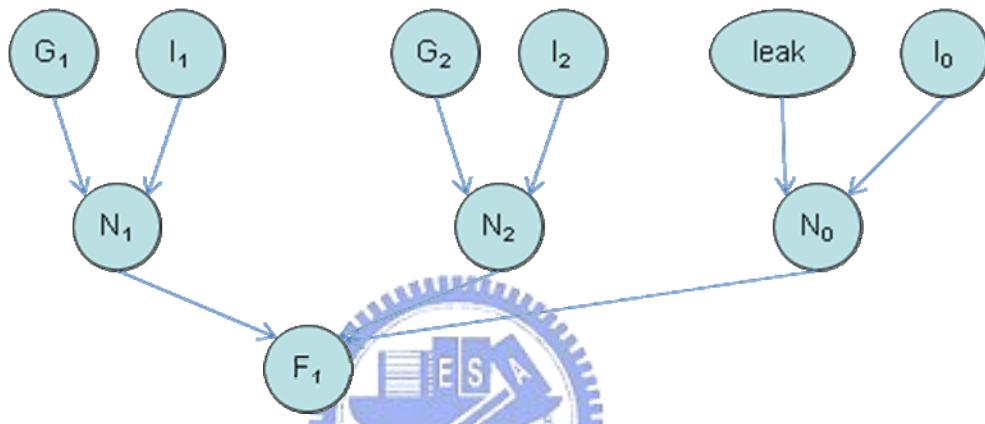
There are three assumptions in noisy OR model:

1. Causal inhibition: The assumption indicates that there are some inhibitive mechanisms which can inhibit a cause to affect their descendants when the cause is active. Only when cause is active and the inhibitive mechanism is turned off at the same time, the cause has impacts on his descendant.
2. Exception independence: The assumption mentions that the mechanism which inhibits one

cause is independent of others which inhibit other causes.

3. Accountability: The assumption points that an effect is valid if at least one of its parents (courses) is present and its inhibitive mechanism is turned off. So, all causes that are not observed but have impacts on effects should be gathered into a cause which is called “unknown” or “leak”.

For example: G: gene, I: inhibitor, N: intermedium, F: feature, leak: leakage.



In above model, I means inhibitors, G means genes and F means features.

$$\begin{aligned}
 P(N_1 = on \mid I_1 = off, G_1 = on) &= 1 \\
 P(N_1 = on \mid I_1 = off, G_1 = off) &= 0 \\
 P(N_1 = on \mid I_1 = on, G_1 = on) &= 0 \\
 P(N_1 = on \mid I_1 = on, G_1 = off) &= 0
 \end{aligned}$$

According to assumption 1, the feature will be present when any N is present. So , we have:

$$P(F_1 = off \mid N_j = on \text{ for some } j) = 0$$

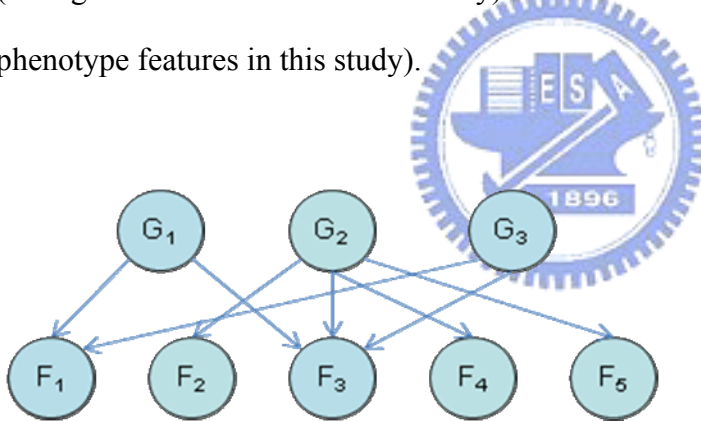
By accountability, we also have:

$$P(F_1 = off \mid N_j = off \text{ for all } j) = 1$$

Chapter 3. Algorithms for candidate gene prediction

3.1 Constructing disease genotype-phenotype network

We construct a mitochondria disease network by the noisy-OR model, which is widely applied for the construction of quick medical reference (QMR) networks for diagnostic assistance (Miller et al., 1986, Shwe et al., 1991, Middleton et al., 1991). QMR networks were constructed by the relationship between diseases and phenotypes in literature. As the information of genotypes become available in this study, we can investigate the relationships between genotypes and phenotypes directly. The noisy-OR model is a bi-partite graphical model as illustrated in the following figure. The top level of the graph contains hidden nodes (like gene deficiencies in this study) and the bottom level contains finding nodes (like phenotype features in this study).



The dependences between three gene deficiencies and 5 phenotype features are modeled via a noisy-OR model. For example, the deficiencies of G1 are associated with feature F₁ and F₃. The feature F₃ is associated with the deficiencies in G₁ or G₃. The association is modeled by the modeling probabilities to describe the noisy patterns.

In order to construct the noisy-OR model for this study, we will use the following estimates with the Bayesian network toolbox in Matlab that is available at <http://bnt.sourceforge.net/>. We also need some parameters to complete this model as follows:

1. $\text{inhibit}(i,j)$ = inhibition probability on the $G_i \rightarrow F_j$ arc and it is estimated by $\text{inhibit}(i,j) = 1 - \text{ratio}(i,j)$, where

$$\text{ratio}(i,j) = \frac{\text{no. of PMIDs for } G_i \rightarrow F_j}{\sum_j \text{no. of PMIDs for } G_i \rightarrow F_j}$$

2. $\text{leak}(j)$ = inhibition probability on the $\text{leak} \rightarrow F_j$ arc and it is estimated by $\text{leak}(j) = 1$ for every j . That is, the feature F_j is only associated with genes considered in this study and there is no leak.
3. $\text{prior}(i)$ = prior probability for the existence of gene deficiency in G_i and it is estimated by

$$\text{prior}(i) = \frac{\text{no. of PMIDs for } G_i}{\sum_i \text{no. of PMIDs for } G_i}$$

3.2 Quickscore

When we have constructed this model, we also need methods to infer the posterior probabilities. Generally, most algorithms for inferring probability of one gene given a set of observed diseases are exponential time-complexity, $O(2^n)$ where n is the number of genes. However, our total number of genes is 174. So, most algorithms are infeasible in our model. Instead of those algorithms, we use quickscore which can reduce the time-complexity to $O(nm^-2^{m^+})$ where m^- means number of diseases without present and m^+ means number of diseases with present (Heckerman, David. 1989).

Suppose that there are n genes which can cause feature F_j to be present. We can get that:

$$\begin{aligned} \text{inhibit}(i,j) &= P(F_j^- | \text{only } G_i^+) = q_{ij} \\ P(F_j^+ | \text{only } G_i^+) &= 1 - q_{ij} \end{aligned} \tag{1}$$

Where q_{ij} denotes the inhibit probability of the feature F_j given gene G_i , F_j^+ and F_j^- denote the presence and absence of feature F_j , G_i^+ denotes the presence of gene G_i .

Besides the genes we have known, there are still some other genes or other factors causing the

features. We lump them into one unknown cause that we named leak and assume that the leak is always presence (prior probability of leak equal to 1). The inhibit probability of leak is:

$$\begin{aligned} \text{leak}(j) &= P(F_j^- | \text{leak}) = q_{0j} \\ P(F_j^+ | \text{leak}) &= 1 - q_{0j} \end{aligned}$$

By the noisy OR model's assumptions, we can get the probabilities of feature given multi-genes easily.

$$\begin{aligned} P(F_j^- | G_1^+, G_2^+, \dots, G_n^+, \text{leak}) &= \prod_{i=1}^n P(F_j^- | G_i^+) = q_{1j} \cdot q_{2j} \cdot \dots \cdot q_{nj} \cdot q_{0j} \\ P(F_j^+ | G_1^+, G_2^+, \dots, G_n^+, \text{leak}) &= 1 - P(F_j^- | G_1^+, G_2^+, \dots, G_n^+, \text{leak}) \end{aligned}$$

Let H be a set of genes and H⁺ be the set of present genes in H. Calculating the probability of

$$\begin{aligned} P(F_j^- | H) &= \prod_{G_i \in H^+} P(F_j^- | G_i^+) = \prod_{G_i \in H^+} \text{inhibit}(i, j) = \prod_{G_i \in H^+} q_{ij} \\ P(F_j^+ | H) &= 1 - P(F_j^- | H) \end{aligned}$$

Also considering the "leak",

$$P(F_j^- | H \cup \text{leak}) = P(F_j^- | \text{leak}) \prod_{G_i \in H^+} P(F_j^- | G_i^+) = q_{0j} \prod_{G_i \in H^+} q_{ij} \quad (2)$$

If we want to know the posterior probability, we can compute that by the result of quickscore algorithm.

$$\begin{aligned} P(F(j)^+, F(j)^-) &= \sum_{F' \in 2^{F(j)^+}} (-1)^{|F'|} K \prod_{i=1}^n \left\{ \left[\prod_{F \in F' \cup F(j)^-} P(F^- | G_i^+) \right] P(G_i^+) + P(G_i^-) \right\} \\ P(G_i^+ | F(j)^+, F(j)^-) &= \frac{P(G_i^+, F(j)^+, F(j)^-)}{P(F(j)^+, F(j)^-)} \end{aligned} \quad (3)$$

Where F(j)⁺ and F(j)⁻ denotes a set of presences features and a set of absences features for similar features of deficiency feature F_j, P(G_i⁺) denote the prior(i), 2^{F(j)⁺} is a power set of F(j)⁺.

$$K = \begin{cases} 1 & \text{without leak} \\ \left(\prod_{F \in F' \cup F^-} P(F^- | \text{leak}) \right) & \text{with leak} \end{cases}$$

And we can get the probability of $P(G_i^+, F(j)^+, F(j)^-)$ by setting $P(G_i^+) = 1$ in equation (3),

then

$$P(G_i^+, F(j)^+, F(j)^-) = \sum_{F' \in 2^{F(j)^+}} (-1)^{|F'|} K \prod_{F \in F' \cup F(j)^-} P(F^- | G_i^+) \times \prod_{i' \neq i} \left\{ \left[\prod_{F \in F' \cup F(j)^-} P(F^- | G_{i'}^+) \right] P(G_{i'}^+) + P(G_{i'}^-) \right\}$$

3.3 Algorithms for prediction of candidate genes

In File-PD, there are specific gene deficiencies associated with the features of protein deficiencies. Using the association relationship in File-PD, we can search similar features and candidate genes in File-ALL that are associated with every protein deficiency F_j in File-PD as follows.

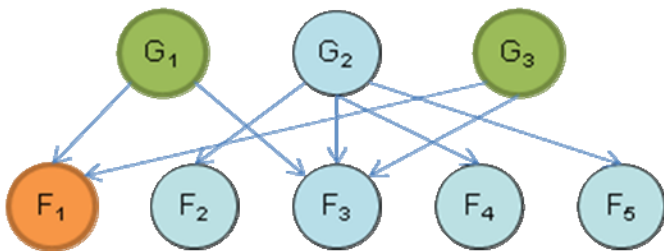


Algorithm 1 for candidate genes:

Step 1. Find the parent set of known gene deficiencies associated with one protein deficiency

F_j in File-PD: $H(j) = \{G_i \text{ in File-PD that is the parent of protein deficiency } F_j\}$.

For example:

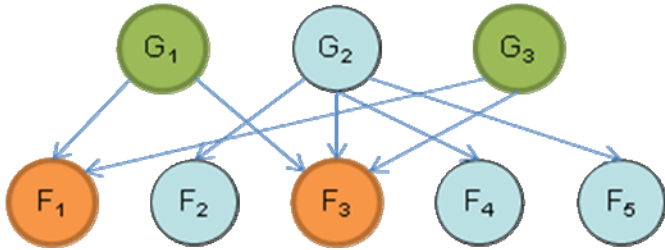


$H(1) = \{G_1, G_3\}$

Step 2. Find the set of similar features in File-ALL whose parents include $H(j)$:

$F(j) = \{F_m \text{ in File-ALL and the parents of } F_m \text{ include } H(j)\}$.

For example:

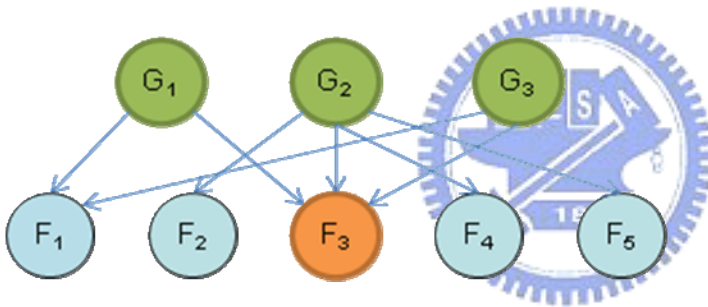


$$F(1) = \{ F_1, F_3 \}$$

Step 3. Find the set of associated genes for every similar feature, F_k in $F(j)$, in File-ALL that are associated with every protein deficiency F_j in File-PD as follows:

$$A(j,k) = \{ G_i \text{ in File-ALL such that } G_i \text{ is the parent of } F_k \text{ in } F(j) \}.$$

For example:

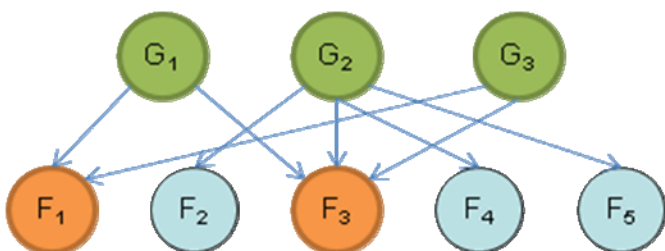


$$A(1,2) = \{ G_1, G_2, G_3 \}$$

Step 4. Find the set of all associated genes as the union of all sets of associated genes for all similar features as follows: $A(j) = \bigcup_{\{k: F_k \text{ in } F(j)\}} A(j,k)$, where the union is over all similar feature

F_k in $F(j)$.

For example:



$$A(1)=\{ G_1, G_2, G_3\}$$

Step 5. Calculate the hit rate of every associated gene as follows: $r(G_i) =S(G_i)/|F(j)|$, where

$$S(G_i) = \sum_{F \in F(j)} I\{G_i \in \text{Par}(F)\} \quad (G_i \in A(j), \text{ and } |F(j)| = \text{number of the set } F(j)).$$

For example:

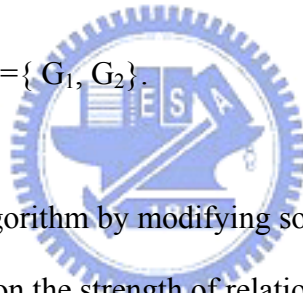
$$S(G_1)=1, S(G_2)=1/2=0.5, S(G_3)=1$$

Step 6. Find the set of candidate genes that include the associated genes with hit rates that are no less than a threshold, like threshold = 0.6 as follows:

$$C(j) = \{G_i \text{ such that } r(G_i) \geq \text{threshold and } G_i \in A(j)\}.$$

For example:

$$\text{Let threshold be 0.6. Then } C(1)=\{ G_1, G_2\}.$$



We also provide another algorithm by modifying some detail. It will add the probability of feature given genes, and focus on the strength of relationships between genes and futures.

Algorithm 2 for candidate genes:

Step 1. Find the parent set of known gene deficiencies associated with one protein deficiency F_j in File-PD: $H(j) = \{G_i \text{ in File-PD that is the parent of protein deficiency } F_j\}$.

Example: $F_j = \{AKDH\text{-deficiency}\}$ and $H(j) = \{DLD, DLST, OGDH\}$.

Step 2. Compute the conditional probability $(P(F_m|H(j)))$ computed by (2)) of all features in File-ALL.

Example: If we want to compute $P(\text{TCA-intermediates-elevated} | \text{DLD, DLST, OGDH})$

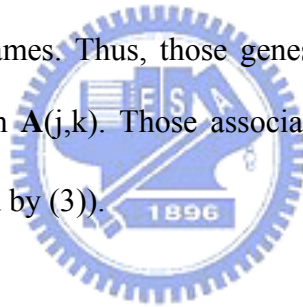
OGDH)=1-inhibit(DLD, TCA-intermediates-elevated)*inhibit(DLST,TCA-intermediates-elevated)*inhibit(DLST, TCA-intermediates-elevated).

Step 3. Find the set (F(j)) of similar features, $F(j)=\{F_m \mid P(\text{feature} \mid H(j)) \geq P(F_j \mid H(j)) \text{ and } F_m \neq F_j\}$.

Step 4. Find the set of associated genes for every similar feature, F_k in F(j), in File-ALL that are associated with every protein deficiency F_j in File-PD as follows:

$A(j,k) = \{G_l \text{ in File-ALL such that } G_l \text{ is the parent of } F_k \text{ in } F(j)\}$.

Example: $F_k = \{\text{TCA-intermediates-elevated}\}$ and $A(j,k) = \{*\text{DLD}, \text{FH}, *\text{DLST}, *\text{OGDH}, \text{BCS1L}, \dots, \text{SCO1}\}$, where those three genes in $H(j) = \{\text{DLD}, \text{DLST}, \text{OGDH}\}$ are marked with * in the front of gene names. Thus, those genes are associated genes after excluding those three genes in $H(j)$ from $A(j,k)$. Those associated genes are ordered in a decreasing order of $P(G_i \mid F_j)$ (computed by (3)).



Step 5. Find the set of all associated genes as the union of all sets of associated genes for all similar features as follows: $A(j) = \bigcup_{\{k: F_k \text{ in } F(j)\}} A(j,k)$, where the union is over all similar feature

F_k in F(j).

Example: $A(1,1) = \{\text{HADHA}, \text{PDHA1}, \text{OAT}, \dots, \text{UQCRB}, \text{PARL}\}$, $A(1,2) = \{\text{DMPK}, \text{HD}, \text{ATP7B}, \dots, \text{HTRA2}, \text{ME2}\}$, ..., $A(1,7) = \{*\text{DLD}, *\text{DLST}, *\text{OGDH}\}$. And the $A(1) = \{*\text{OGDH}, *\text{DLST}, *\text{DLD}, \text{SURF1}, \text{SLC25A19}, \dots\}$. The set of A(1) is the union of A(1,1), A(1,2), ..., A(1,7).

Step 6. Calculate the hit rate of every associated gene as follows: $r(G_1) = S(G_1)/|F(j)|$, where

$$S(G_1) = \sum_{F \in F(j)} I\{G_1 \in \text{Par}(F)\} \quad (G_1 \in A(j), \text{ and } |F(j)| = \text{number of the set } F(j)).$$

Example: $S(\text{GCDH} \in A(1))=5$, and $|F(1)|=6$. The hit rate is $r(\text{GCDH}) = 0.833333333$.

Step 7. Find the set of candidate genes that include the associated genes with hit rates that are no less than a threshold, like threshold = 0.6 as follows:

$$C(j) = \{G_1 \text{ such that } r(G_1) \geq \text{threshold and } G_1 \in A(j)\}.$$

Example: $C(1)=\{ \text{SURF1, SLC25A19, SLC25A15, ..., AASS}\}$.



Chapter 4. Leave-one-out Cross Validation

Cross-validation (CV) studies are performed to determine the threshold and empirical hit rates in prediction of candidate genes. Beside this, we also can compare the performance for those two methods in above chapters by ROC method and other strategies.

4.1 Steps for leave-one-out cross validation

In this section, we will show the procedures of how to do cross validation and their results.

Steps for cross-validation:

Step 1.

One association relationship between G_i and F_j in File-PD is removed from File-ALL each time. We thus generate new data sets, data1, data2, ..., dataN, where N is the number of total relationship rows existing in File-PD.

Example 1: The file of data1 is the File-ALL removing the relationship about DLD and AKDH-deficiency.

Step 2.

We apply the steps for finding candidate genes in Algorithm 1&2 using the file that has removed the relationship of G_i and F_j . Then, we obtain the hit rate for a relationship between F_j and G_i .

Example 1 (continued): From the file of data1, we obtain the hit rate of $r(DLD)= 1$.

From the CV results, most cases have hit rates that are at least 0.6 except a few of cases. Then, the overall average of all hit rates in algorithm 1 is **0.8912** and the overall average of all hit rates in algorithm 2 is **0.9279**. We also try other strategies to further compare these two results of leave-one-out cross validation.

4.2 Evaluating performances of two algorithms

There are 109 relationships in File-PD. So we have 109 results of cross validation in each cross validation. First, we compare two methods' AUC (area under ROC curve, let y be power and x be type I error). There are 109 AUC in each method.

Table 1. Information about AUC of two algorithms under 109 cross validations

Method\AUC	Median	Mean	Max	Min
CV for alg. 1	0.7114094	0.7640562	1	0.6354839
CV for alg. 2	0.8001468	0.8131531	0.9795322	0.5941176

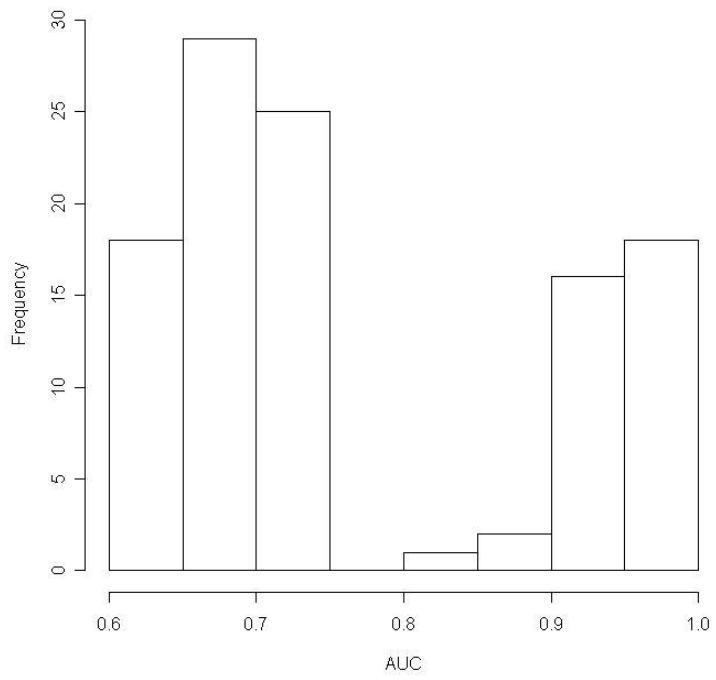


Fig 1. Auc of alg. 1

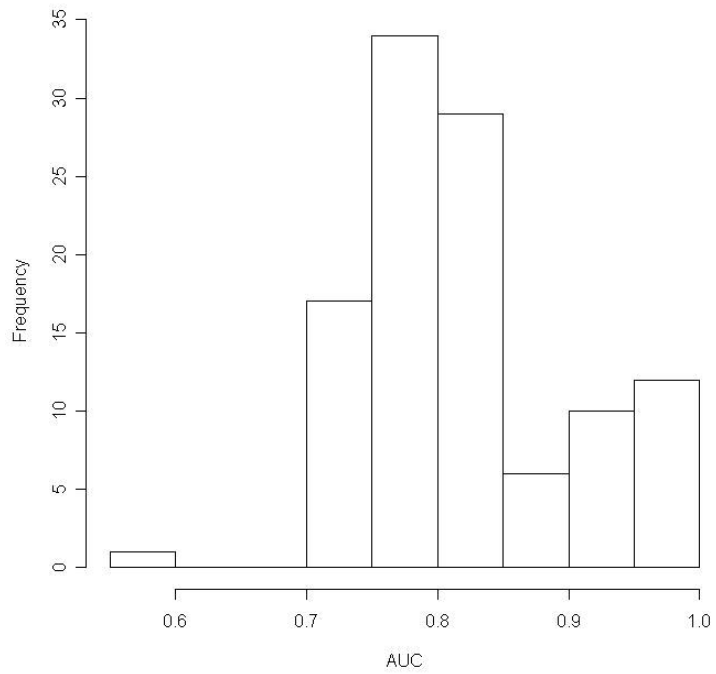


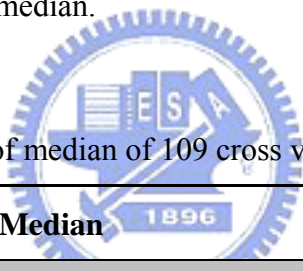
Fig 2. Auc of alg.2

From table 1, we can find that the performance of algorithm 2 is better than algorithm 1. But there is one questionable point that why we can decide type I and type II error without the knowledge about which gene truly associates with deficiencies. Furthermore, we are going to compare the ratio that we define as:

$$\text{covering ratio} = \frac{\text{covering rate of known genes associated with a deficiency}}{\text{the number of candidate genes}}$$

Each protein deficiency contains some gene deficiencies. The numerator of ratio indicates that gene deficiencies predicted by different thresholds divides the total number of gene deficiencies associated with this protein deficiency. By this definition, the larger ratio is, the better performance is. Because we hope the model can have a high covering rate and select less number of genes. Let cut points from 0.01 to 1. In each cut point, we have 109 ratios and we summarize those results by median.

Table 2. The maximum values of median of 109 cross validation under 100 thresholds



Method\Max	Median
CV for alg. 1	0.009009009
CV for alg. 2	0.01397516

From table 2, the performance of algorithm 2 is still better than algorithm 1 in ratio. The conclusion of comparison algorithm 1 and 2 should be that algorithm 2 has a better performance than algorithm 1.

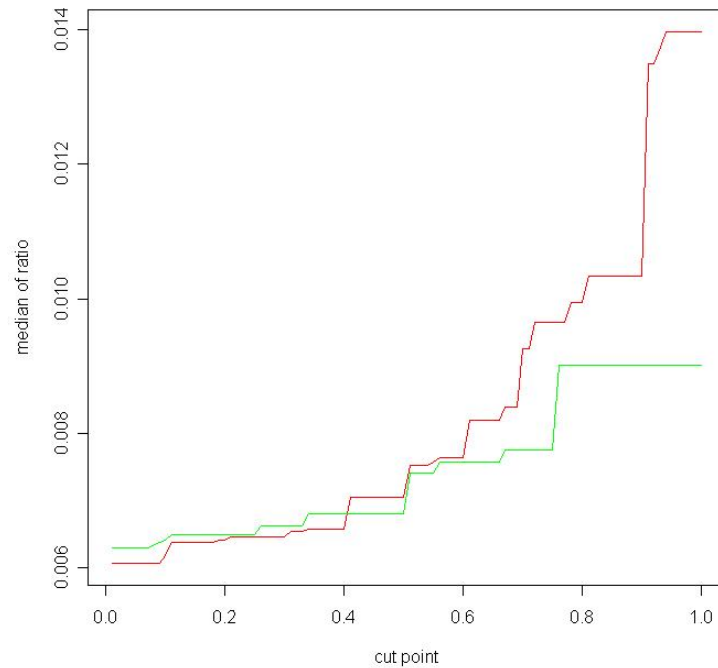


Fig 3. Ratio of alg. 1(green) & alg. 2(red)



4.3 Threshold

In section 4.3, we talk about that we need to decide thresholds for those two algorithms. Here, we have three strategies to choose thresholds. First is controlling type I and type II error by arithmetic mean, second is finding the shortest distance by geometric mean and final is through observing the jump of the covering rate.

First, let us control the type I and type II error by their arithmetic mean. There are 100 cut points from 0.01 to 1. Each point also contains 109 arithmetic means and we summarize information by taking average.

Table 3. Thresholds and their corresponding arithmetic means in each algorithm

Method\Arithmetic mean	Min	Cut points
CV for alg. 1	0.2513621	0.97, 0.98, 0.99, 1
CV for alg. 2	0.2582092	0.79, 0.8

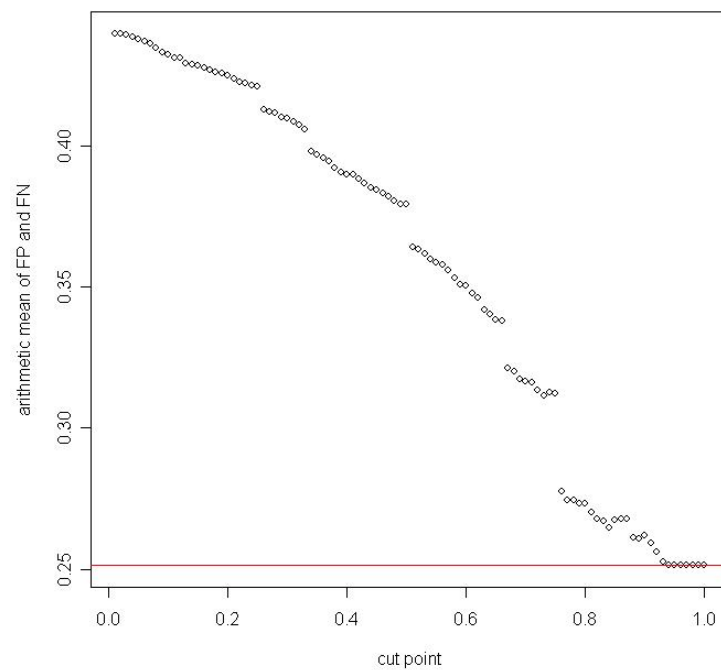


Fig 4. Arithmetic mean of type I and type II error VS. cut point for alg. 1

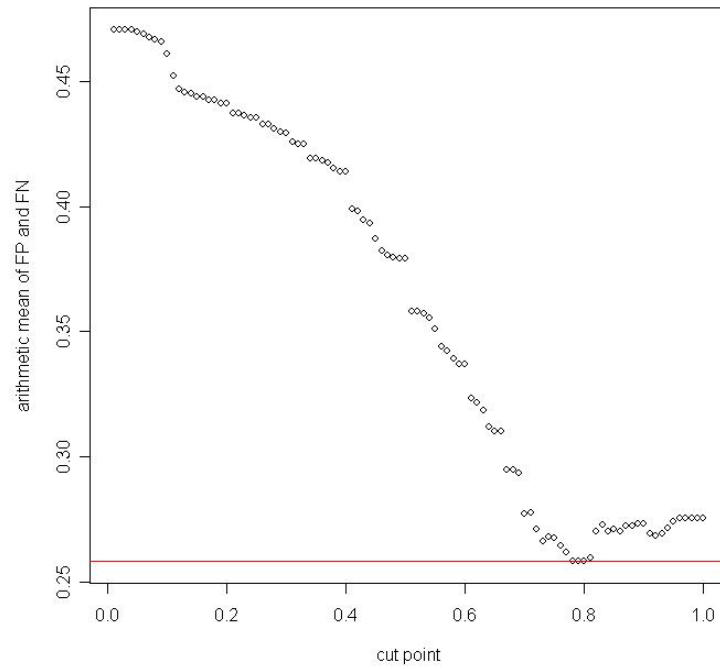


Fig 5. Arithmetic mean of type I and type II error VS. cut point for alg. 2

Second, we find the shortest distance calculated by geometric mean of false positive and false negative.

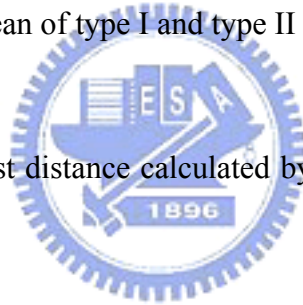


Table 4. Thresholds and their corresponding geometric means in each algorithm

Method\Distance	Min	Cut point
CV for alg. 1	0.4580215	0.97, 0.98, 0.99, 1
CV for alg. 2	0.3869187	0.89, 0.9

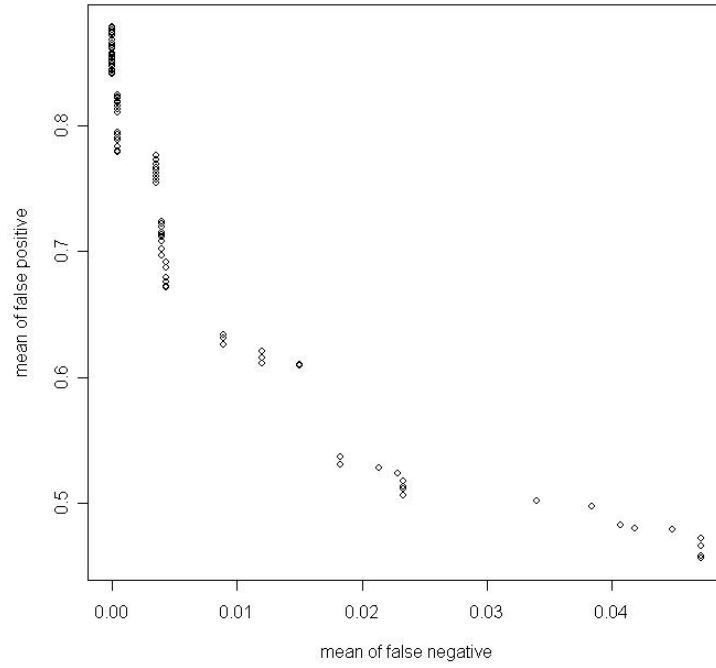


Fig 6. Geometric mean of type I and type II errors VS. cut point for alg. 1

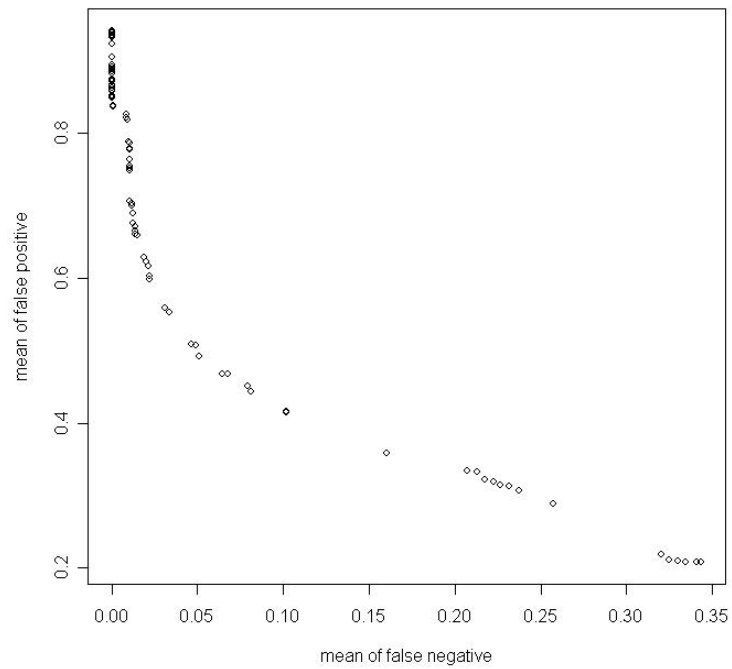


Fig 7. Geometric mean of type I and type II errors VS. cut point for alg. 2

Finally, we try to observe the change of covering rate, and find out the cut point which will make the covering rate change rapidly. This will mean that using the point might attain a lower number of candidate genes and an appropriate covering rate.

Table 5. Thresholds and their corresponding jumping ranges in each algorithm

Method	Cut point	Difference with next point
CV for alg. 1	0.84	0.01070336
CV for alg. 2	0.9	0.0633653

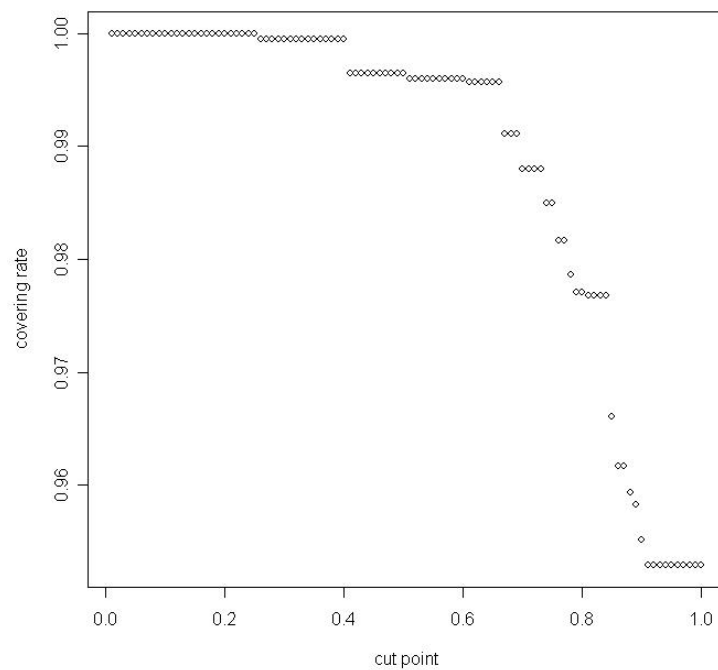


Fig 8. Covering rate vs. cut point for alg. 1

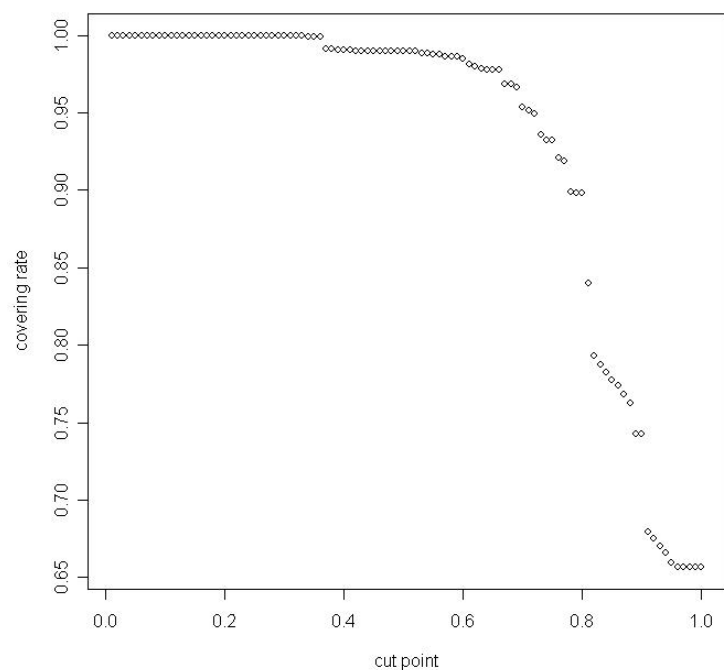


Fig 9. Covering rate vs. cut point for alg. 2



Table 6. Information about thresholds and corresponding number of candidate genes in two alg.

conclusion\method	Arithmetic mean	Geometric mean	Covering rate
Cut point (alg. 1)	0.84	1	1
Candidate gene	94.9266	86.3578	86.3578
Cut point (alg. 2)	0.8	0.9	0.9
Candidate gene	80.26606	57.70642	57.70642

Chapter 5. Conclusion and Discussion

First, after deciding the threshold, we can construct a gene-disease deficiency qmr-like model. But it still lacks information to estimate model parameter. Maybe we can gather more data, but it is not easy because our data are depending on literature reports. New data need new PubMed publishes. So the next research orientation should be working on new method for tuning parameters of predicted models.

Second, in our noisy-OR model, we ignore the “leak” and assume its inhibiting probabilities 1. But this assumption seems questionable, because this indicates that there are no any factors which will affect the diseases, excluding those genes we have known from literature reports. In order to complement this model, estimation of “leak” is a subject which we can keep working on. Although we have tried some statistic methods, the results are still so unconvincing.

Furthermore, no matter which genes we choose, the most important thing is that we need a golden standard to compare with the results of our algorithms. If there are biological and experimental validations, then our results will be more persuasive.

References

- [1] Jensen FV. Bayesian Networks and Decision Graphs. Springer. 2001.
- [2] Heckerman D. A tractable inference algorithm for diagnosing multiple diseases. Proceedings of UAI. 1989:174–181.
- [3] Miller R, Masarie FE, Myers JD. Quick medical reference (QMR) for diagnostic assistance. MD Comput. 1986 Sep-Oct;3(5):34-48.
- [4] Middleton B, Shwe MA, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. II. Evaluation of diagnostic performance. Methods Inf Med. 1991 Oct;30(4):256-67.
- [5] Neapolitan RE. Learning Bayesian Networks. Prentice Hall. 2004.
- [6] Shwe MA, Middleton B, Heckerman DE, Henrion M, Horvitz EJ, Lehmann HP, Cooper GF. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. I. The probabilistic model and inference algorithms. Methods Inf Med. 1991 Oct;30(4):241-55.