Abstract

Title of Thesis：Discovering the gene and disease relationship from biomedical literature database by applying the information retrieval technology

Author：Ming Chin Lin

Thesis advised by ： Yu-Chuan Li, Professor

Thesis co-advised by ： Chien-Yeh Hsu, Associate Professor

With the development of biomolecular technology, there is getting more and more information derived from genome research. Besides, the microarry was introduced to allow people study genome wide pattern of gene expression profile, the scientists have the opportunity to study the function of genes.

At the same time, the functional genomic research also bring a great impact to clinicians which usually study single gene or study disease at biochemistry level. In traditional, the MEDLINE always is the major resource for clinicians research. Recently, the explosion amount of the genomic related research bring for clinicians is too complicated to understand it. For examples, when talking about one disease, there are approximate over ten thousand of articles and hundred genes in it. It is almost impossible for clinicians to digest the knowledge. So it is urgent that there must be some computational tools developed to help clinicians observing the gene and disease relationship

In this research, we focus on constructing the probabilistic model of gene and disease relationship. By using two models to represent the knowledge from biomedical literature database, we can compare the two models in system performance and

precision.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Significance

Nowadays, the research interest is shifting from gene sequence to functional genomic research. In the functional genomic era, the most important thing is to observe how genes act in our human body [1]. Since microarray was introduced to allow people study genome wide pattern of gene expression profile, the scientists have the opportunity to study the function of genes [5].

At the same time the clinicians also mention about the important of the functional genomic research, because they have chance to observe the patient at genetic level [9]. It is a breakthrough thought for modern medicine research, especially in cancer research. Nowadays, the cancer is the major death reason in developed countries and how to detect or understand cancer is always the most important issue for clinicians. Hopefully, due to the development the molecular biology, the biomedical research brings a lot of information to clinical research, for examples, describing disease in gene model.

But the problem is that how clinicians use this information to help them in clinical cancer research. For example, when clinicians are interesting in "Breast cancer disease" and genomic level pathological mechanism of diseases, usually clinicians will use MEDLINE to search the "Breast cancer" related articles. But in fact, when querying in MEDLINE by using "Breast cancer" as a MeSH term keyword, there are approximate 60 thousand of articles derived and there are more than thousands of symbols appearing in it. As we know, in molecular medicine research, we hope we

can detect disease in genetic level, but it is merely impossible for human to read about more than 10 thousand articles, and then have the insight about genes and disease relationship without any other computational tools help. Besides, until right now, there are approximate more twenty thousand genes are discovered, and it is difficult for clinicians to recognize all the symbols in it [18].

Further more, there are too many genes describing in biomedical literature, it is very difficult for clinicians to describing which gene is more important and which is less important.

Therefore, for clinicians, it is very important that to develop a tool to capture the gene related information in biomedical literature database and then use the information to build up probabilistic model describing genes and disease relationship.

1.2 Motivation

In this research, we focus on how to capture the information from biomedical literature database, and use the information to build up the gene-disease probabilistic model. First, the information retrieval (IR) technology is developed by computer science for a long time, and recently there are many studies talking about how to apply IR in biomedical research [2, 11, 13, 15, 16].

Second, we have to find out that how to build up the probabilistic model. In medical informatics research, there is well-developed medical decision system to help clinicians to make diagnosis in clinical problem. The most popular examples is the use of the EKG monitor. Nowadays, the EKG monitor using in the hospital which can make a precise diagnosis from patient's clinical data collecting from EKG lead. The capability of computational medical decision support is approved in many fields, such as EKG use, clinical differential diagnosis and many others.

In medical decision theory, Bayesian decision theory is a fundamental statistical

approach to the problem of medical decision. To answer the medical decision problem, usually we can use Bayesian formula to calculate the probability of states.

Further more, if we have a complex medical question, we can draw a causal relationship graph to represent the knowledge. Usually this graph is so-called causal network or simply belief nets by mean of Bayesian belief net. They take the topological form of a directed acyclic graph (DAG) (Figure 1), where link is directional and there is no loop.

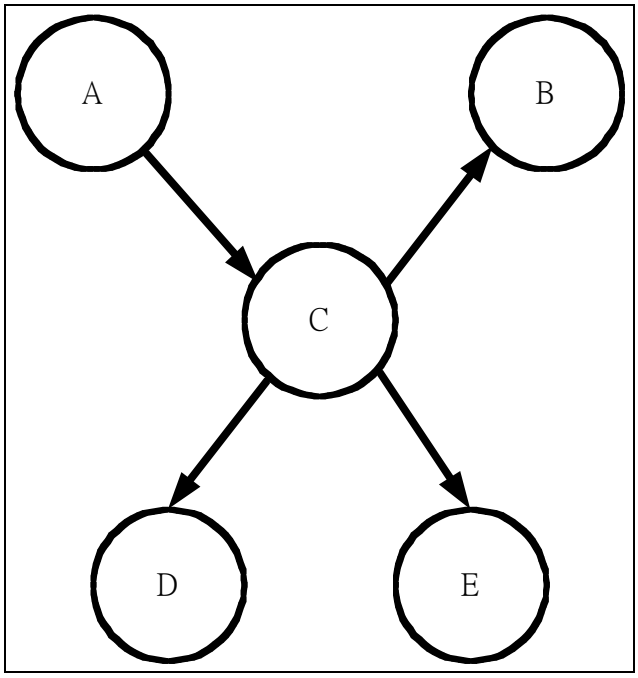

Figure 1 Directed acyclic graph (DAG)

From the knowledge of molecular technology, we can know that there are a lot of relationship between gene and gene and gene and disease. So we are interesting that is the implicit biomedical knowledge from biomedical literature database can be represented by Bayesian formula to give as an inside of gene - disease relationship. But the problem is that, how to build up the causal network of gene and disease. In the traditional clinical problem, usually the causal relationship is built by clinical experts, and it is too complicated problem for human being to construct this

gene-disease causal map. So some automation tools are needed to for this. Another problem is that how to build up the conditional probability table of gene-disease causal network. Fortunately, there are a lot of biomedical literature database available on Internet. So we try to use that information to build up our network.

1.3 Research purpose

In this research, we want to capture the gene – disease relationship from biomedical literature database. So it is very important that how to collect the articles, which are describing the gene and disease relationship.

In PubMed database, there is a lot of effort done for the article categorization. Especially, the Medical Subject Heading (MeSH) made a great contribution in this area. So in this research that we believe one article belongs to one MeSH term, the articles are that term related. For example, one article belongs to "Asthma MeSH term"; we can say that this article is asthma related. To extend this, we believe that if one gene appears in one article, then the gene is related to the MeSH term of that article (Figure 2).

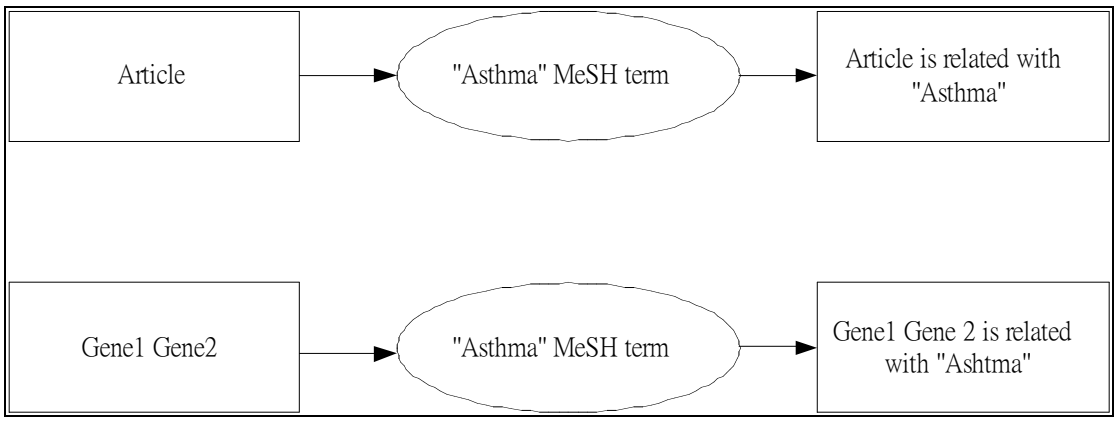

Figure 2 The relationship of articles and MeSH term

In the other hand, we can draw a new concept map (Figure 3)

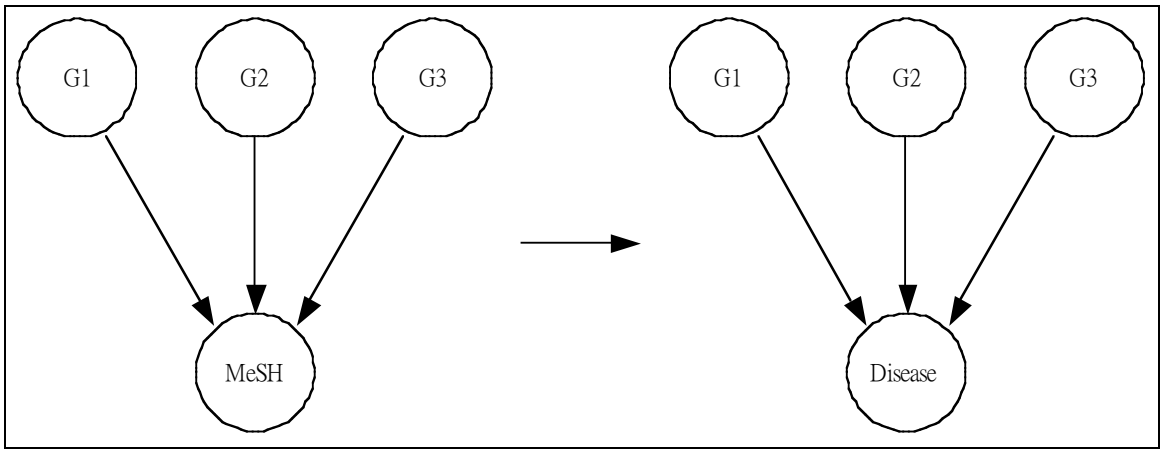

Figure 3 Illustrate the gene, MeSH, and disease relationship

To expend this, we can draw two causal networks to represent this causal relationship.

In this research, we focus on two topics:

A. Representing disease categories and gene relationship by two different causal networks from biomedical literature database (Figure 4, 5).

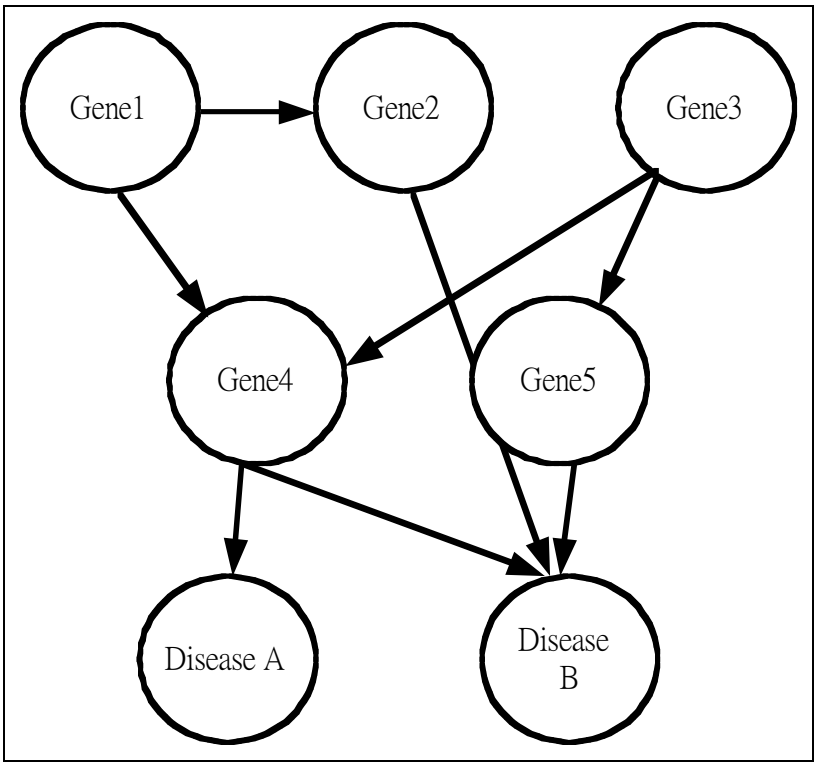

Figure 4 "Structural Learning model" model of gene-disease relationship.

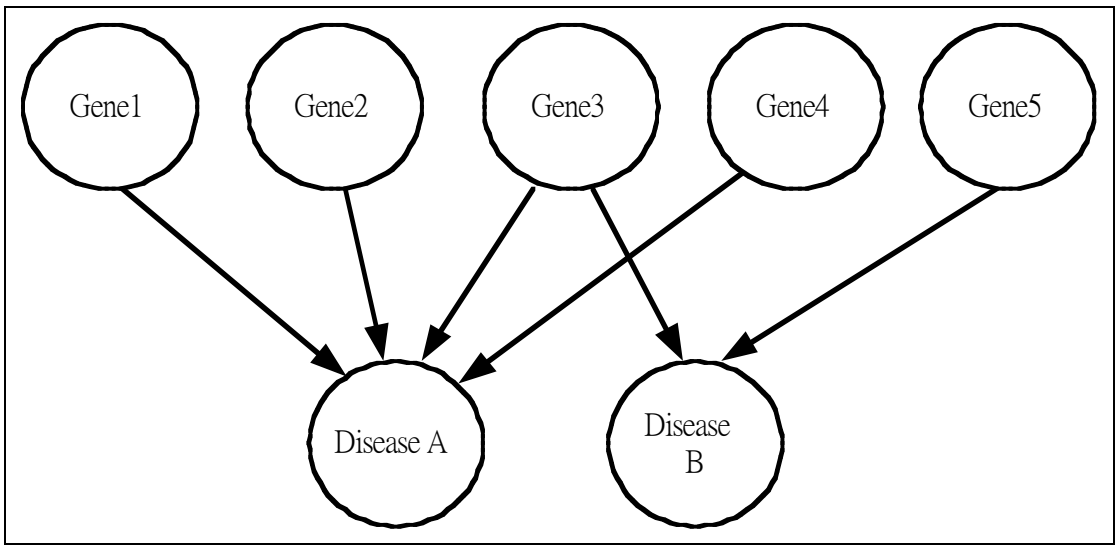

Figure 5 "Causal model" of gene-disease relationship.

B. If we can draw the causal network and calculate the individual disease

probability, we can use the genes which occurrence among in articles to determine the article disease category

In this research, we try to represent the implicit knowledge from biomedical literature, especially the gene and disease relationship. From above, we explain that we will construct two models "Structural Learning model" and "Causal model" to illustrate the gene disease relationship.

In the other hand we are interesting the probabilistic relationship of the two models. So we also have to calculate the probability state of the models. Because in this research, we focus on the building of the network, we would the commercial software "HUGIN researcher" to help us calculate the probability and the Structural of models. To reduce the complexity, in "Causal model", we can separate it to two individual causal models (Figure 6, 7). Then we can calculate the probability disease by disease.

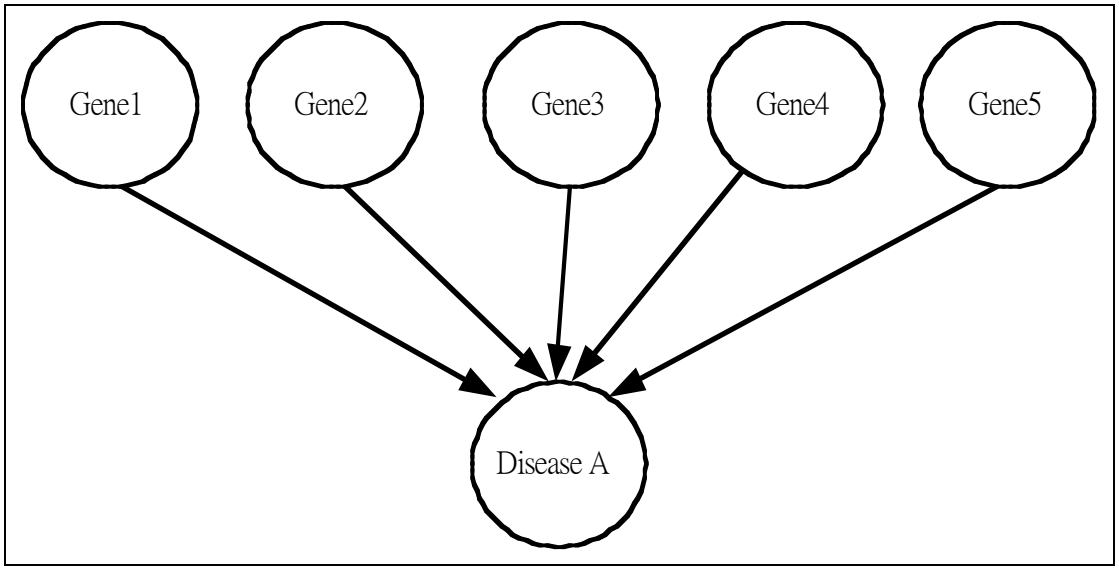

Figure 6 "Causal gene model" of gene-disease relationship for disease A

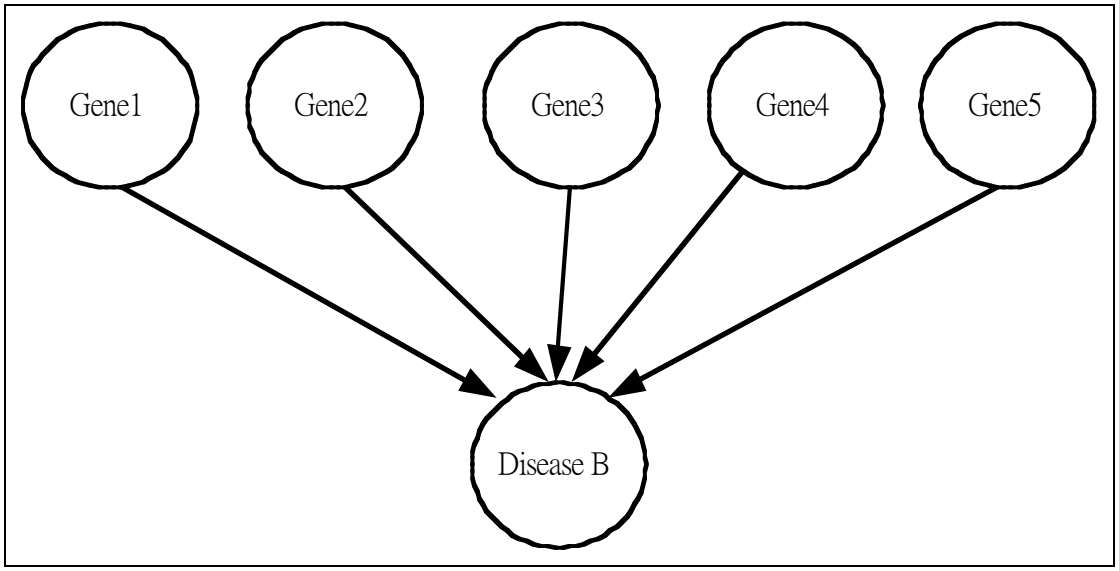

Figure 7 “Causal model” of gene-disease relationship for disease B

It is the same in “Structural Learning model”; we can build up two models as following (Figure 8, 9).

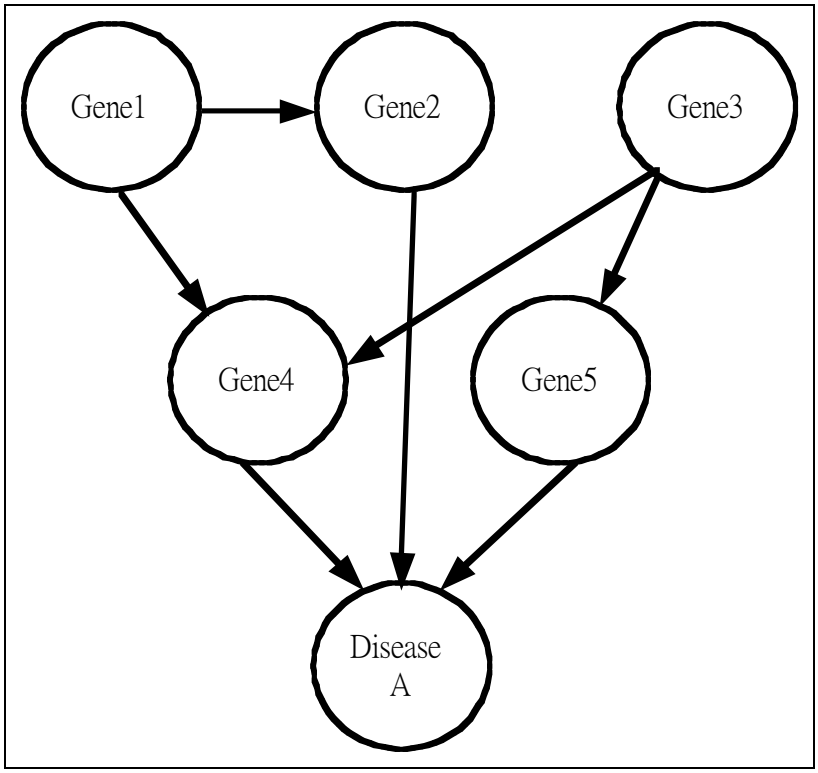

Figure 8 “Structural Learning model” of gene-disease relationship for disease A

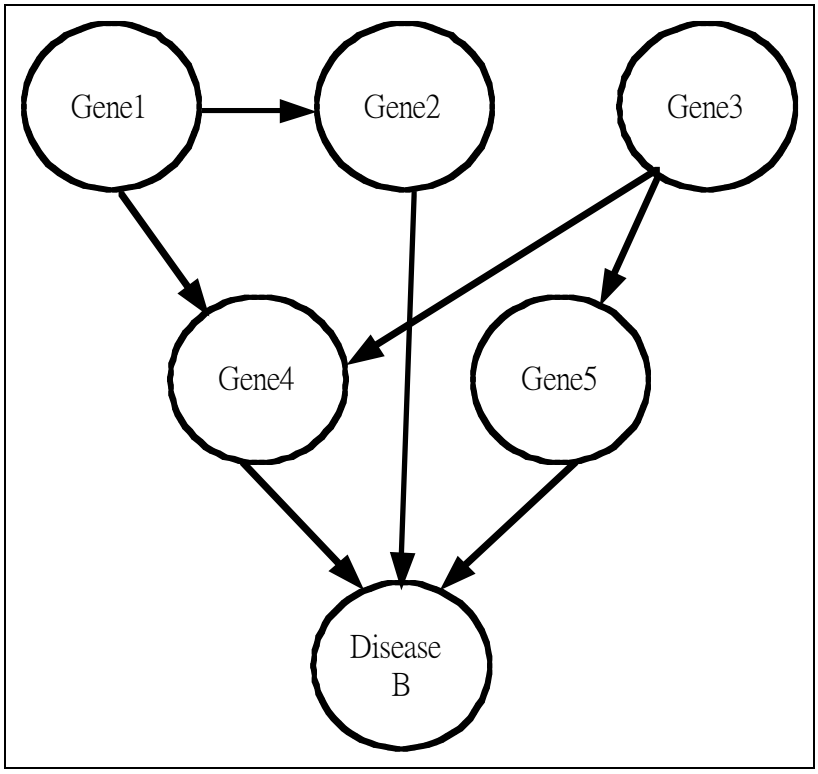

Figure 9 “Structural Learning model” of gene-disease relationship for disease B

By using the above causal network models we can represent the gene-disease knowledge from biomedical literature.

To reach those two purposes, we design the research as following (Figure 10)

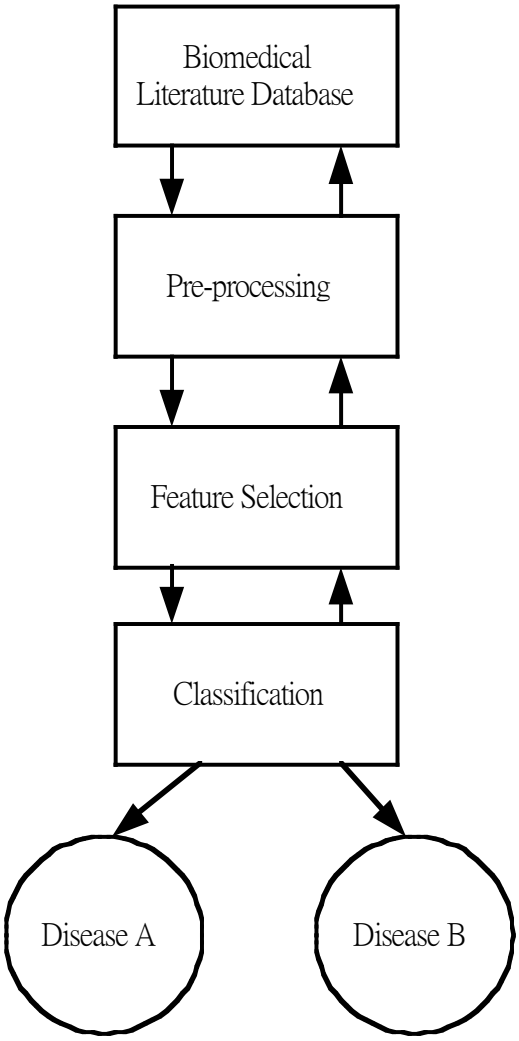

Figure10. The research flow

We try to build up two gene-disease causal networks to represent the biomedical knowledge from literature database, and than using cross validation to validate the network. If the network can tell from two diseases categories correctly, we can say the model is suitable for classification.

Chapter 2

Literature Review

With the development of molecular biology, there is getting more and more biomedical knowledge database available on Internet. Nowadays, it is very convenience for scientists to search information through public databases via Internet, for examples, "Medline" which is founded by National Library of Medicine (NLM) in United States. By using "Medline" this kind of public literature databases, the users can search the related topics by specified disease states, chemical substance, syndrome, and so on. Usually, if the user query one term in Medline without any limitation and then there are more than thousands of articles derived. In practice, the information directly derived from search engine is too huge and complicated for human reading.

To solve this problem, using computers to analyze those articles is very helpful for users to extract valuable information from biomedical literature database. But, reading in the biomedical research field, the specialized knowledge in specified domain is required. For example, there are approximate10 thousand kinds of gene names and thousand kinds of genes in the biomedical database. Even though for human, to understand all of the knowledge is almost impossible. On the other hand, the high computing power of computers provides the new chance for us to dig the implicit information from those biomedical databases. Especially, nowadays, the enormous amount and variety are too complicated for human to understand. Therefore, to develop the refined tools for helping analyze the biomedical knowledge is an urgent task in bioinformatics research.

Fortunately, in the past 10 years, the previous researchers provide possible research direction, methodology and resource on information retrieval [2, 4, 6, 7, 8, 10]. We can follow their research result to extend their work. In the following article, the previous related research would be discussed.

2.1 Building public database

In the last decade, the investigators developed several huge and centralized biomedical databases. With those databases, it is more convenience to exchange the research result between scientists, for example, GDB(http://gdbwww.gdb.org/) Genome Database (USA), OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM) Online Mendelian Inheritance in Man (USA) ,GENATLAS (http://www.dsi.univ-paris5.fr/genatlas/) ,GeneCards(http://bioinformatics.weizmann.ac.il/cards/) integrated database and so on.

2.2 Construction of lexicon

The first step of information retrieval is to build up the lexicon. In 1998, Proux proposed detecting gene symbols and names in biological texts, their study group use lexical analysis and contextual analysis to search the gene symbol and build up the gene synonym maps [17]. In the other hand, there is the Human Genome Organisation (HUGO http://www.gene.ucl.ac.uk/hugo/) providing the official gene symbol and centralized database. Therefore, the papers published in recent year usually contain the official name provided by HUGO.

The method of construction lexicon:

- Manual
- Statistic and frequency based
- Natural Language Processing (NLP)
- Semantic
- Rule-based

2.3 Analyzing article collection

In the traditional databases, like PubMed, it only provides simple keyword search tool and the search result is sometimes enormous and complicated. In the post-genomic era, the research scientists usually face hundred of genes from thousand of articles. It is almost impossible for users to extract the relationship of genes. According to this, there are several related search focusing on computer-aided information retrieval.

In 1999, Blaschke and their colleagues proposed:"**Automatic extraction of biological information from scientific text: protein-protein interactions**" [4]. They use Bayesian model to describe the PIP related articles and PIN non-related articles by using discriminating keywords. In 2000, Rindflesch proposed:"**EDGAR extraction of drugs, genes and relations from the biomedical literature**"[19]. They used the keywords "neoplasms AND cells AND gene AND drug AND resistance AND mechanism" querying in PubMed and 383 abstracts related to anti-tumor drug resistance are derived. They used hierarchical clustering method to illustrate the relationship of discriminating keywords between articles. They can derive the advance information from articles without reading any articles in details.

In 2001, Jenssen and their study group proposed: "**A literature network of**

human genes for high-throughput analysis of gene expression" [12]. They proved their hypothesis that "The pair of genes co-occurrence in the same articles implies that they are functional related". Using this, they can illustrate the gene relationships graph.

From above articles, there are several advance databases available, for examples, MedMiner "http://discover.nci.nih.gov/textmining/filters.html", GeneCards http://bioinfo.weizmann.ac.il/cards/ [18], PubGene http://www.pubgene.org/ [18].

2.4 Introduction the "Causal model" and "Structural learning model"

There are two models used in this project to illustrate the gene and disease relationship. One is called "Causal model" and another is called "Structural Learning model".

In "Causal model" all the related genes are connected to diseases node directly, and there is no other link between gene and gen. In "Structural Learning model", all the genes and diseases have possible link to each other. In directed acyclic graph (DAG), (Spirtes, Glymour and Scheines, 1993; Pearl, 1995, 2000; Swanson and Granger, 1997) establish a **PC algorithm [22]**, which performs a series test of conditional independence test on the sample and use the result to build the directed acyclic graph. The PC algorithm is wildly used in many fields, such as finance, mathematic and bioinformatics research to find the causal relationship between variables. At following paragraph, we will explain how PC algorithm works.

First, we focus on only two nodes X and Y. There are three situations between X an Y.

X—Y (Undirected edge)

X→Y (Directed edge)

X←>Y (bi-direct)

Step1: if there are N variables, we can draw a graph as following, and link all the nodes together. (Figure 11)

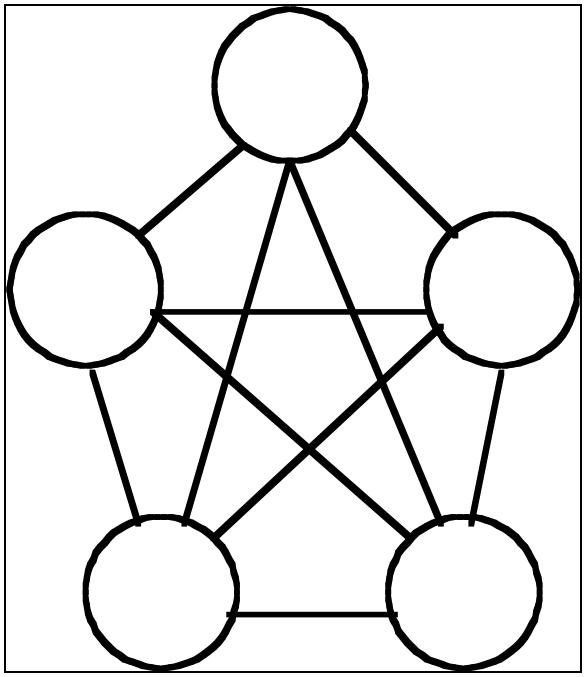

Figure 11 Undirected edge graph

Step2: PC algorithm removed the links from the complete undirected graph by first checking for unconditional correlation between pairs of nodes (Figure 12).

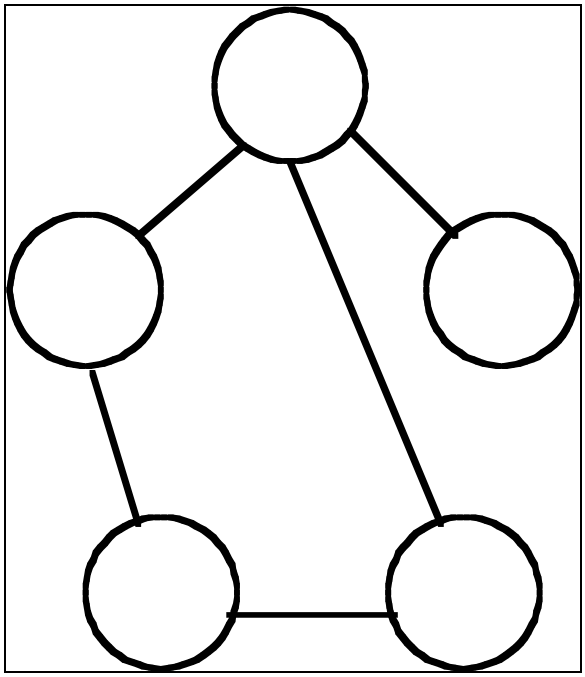

Figure 12 Undirected edge graph after removing the links

Step3: the links between two nodes in undirected graph have to be examined to

decide the direction of link. But there is no enough information to decide the direction of the link, the additional information from other node is required. So we can draw a triplet node as following.

In the figure, only state I and II, the link between X and Y can be decided. We can not prove X and Y link direction with Z information, either can not when reversing all the link in the triplet. So we just discuss the state I and II.

How to determine state I or II is based on observing the relationship of Y and Z. In state I, Y→ X ←Z, we can observe the unconditional association of Y and Z is zero and the conditional association in non-zero. And in state II, Y← X →Z, we can observe the unconditional association of Y and Z is non-zero and the conditional association is zero.

Step 4: If we decide X and Y belong to which state, then we can decide the casual direction of the X and Y. In figure, we can observe the relationship of X, Y and Z (Figure 13, 14).

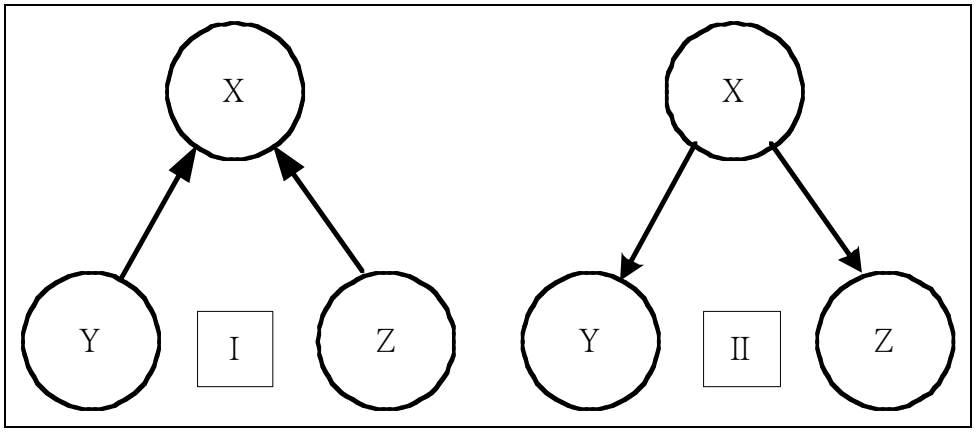

Figure 13. State I and II show directed edge graph

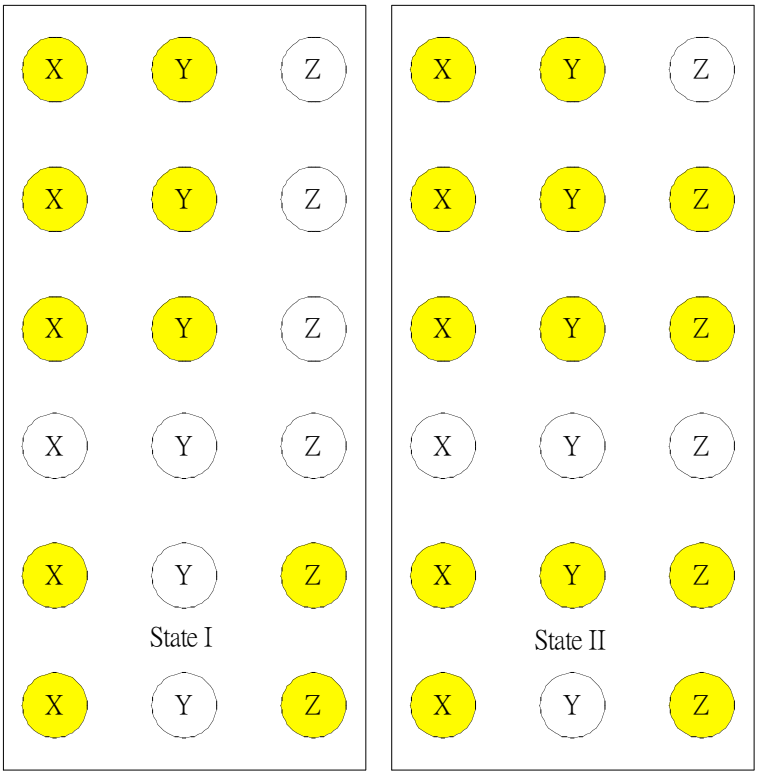

Figure 14 The relationship of X ,Y and Z

Chapter 3

Material and method

In this research, we try to represent the gene-disease relationship from biomedical literature database. Nowadays, the MEDLINE is the major biomedical literature database all over the world, so we choose the WEB edition of MEDLINE "PubMed" as our literature database resource. To apply the information retrieval technology on biomedical literature, we design a research flow as following figure (Figure 15).

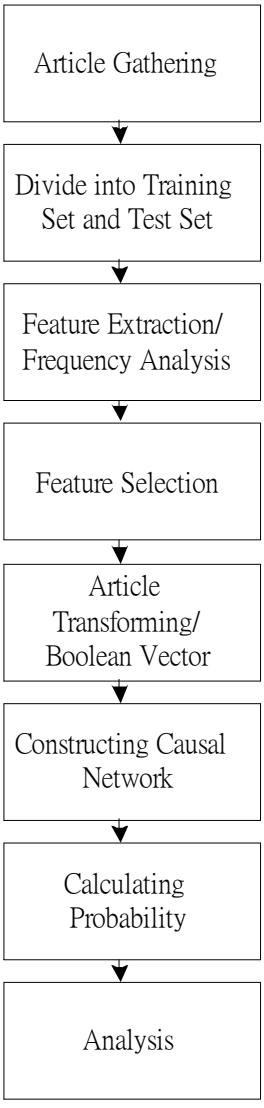

Figure 15 The research method

3.1 Article gathering

We collect the biomedical literatures from "PUBMED http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed ", which is constructed under the National Center for Biotechnology Information. Nowadays, the contents of web pages become the most popular and easy way to exchange our information. Especially, from personal computer to enterprise server, we all can exchange our information through web page. So in this project, we use the most popular technology WWW to collect articles instead of using Z.39 protocol, which is used in the past.

3.1.1 Search strategy

In this research, we chose three diseases categories "Dermatitis, Atopic", Asthma", "Breast Cancer".

First, we use MeSH browser (http://www.nlm.nih.gov/mesh/2002/MBrowser.html) to find the MeSH term of three diseases (Table 1, 2).

Table 1 MeSH term tree of Asthma and Dermatitis, Atopic

Immunologic Diseases [C20]

Hypersensitivity [C20.543]

Hypersensitivity, Immediate [C20.543.480]

Anaphylaxis [C20.543.480.099]

Conjunctivitis, Allergic [C20.543.480.200]

Dermatitis, Atopic [C20.543.480.343]

Food Hypersensitivity [C20.543.480.370] +

Respiratory Hypersensitivity [C20.543.480.680]

Alveolitis, Extrinsic Allergic

[C20.543.480.680.075] +

Aspergillosis, Allergic Bronchopulmonary [C20.543.480.680.085]

Asthma [C20.543.480.680.095] +

Hay Fever [C20.543.480.680.425]

Rhinitis, Allergic, Perennial [C20.543.480.680.791]

Urticaria [C20.543.480.904] +

Table 2 MeSH term tree of Breast Neoplasms

Neoplasms [C04]

Neoplasms by Site [C04.588]

Abdominal Neoplasms [C04.588.033] +

Anal Gland Neoplasms [C04.588.083]

Bone Neoplasms [C04.588.149] +

Breast Neoplasms [C04.588.180]

Breast Neoplasms, Male [C04.588.180.260]

Mammary Neoplasms [C04.588.180.520]

Mammary Neoplasms, Experimental [C04.588.180.525]

Phyllodes Tumor [C04.588.180.762]

The PUBMED website provided a powerful search tool for users to explore the wanted information. In the hyperlink PUBMED tutorial, http://www.nlm.nih.gov/bsd/pubmed_tutorial/m2016.html , they build up a very

detail manual. In the following articles, we will introduce the "Search Strategy" used in this project. To reach this purpose, we have to apply "Boolean Logic" to decide our boundary of our article set.

3.1.2 Introduction of "Boolean logic"

AND

Using the AND operator, all the rules must be true, then the articles will be collected.

For example: Querying "Asthma AND Atopic dermatitis"

OR

Use the OR operator to retrieve documents that contain at least one of the specified search terms.

Use OR when you want to pull together articles on similar subjects.

For example: Querying "Asthma OR Atopic dermatitis"

NOT

Use the NOT operator to exclude the retrieval of terms from your search.

For example: Querying "Asthma NOT Atopic dermatitis"

3.1.3 Advanced search

The PubMed is well-developed search engine for biomedical literature database. When using a query in PubMed directly, there are a lot of intelligent modifications done by the system itself. To restrict the research articles, we use the advanced search mode to make sure every query with the same rules. The following figures shows how we done for using PubMed advanced search.

Step1.To have more precise boundary result, we use “Limit” and “Detail” to control our query result. In this research, we restrict our query term matching the MeSH term in PubMed limit column. Because the MeSH term was done by a lot of experts manually, we use this as our golden standard to decide the collecting article category (Figure 16).

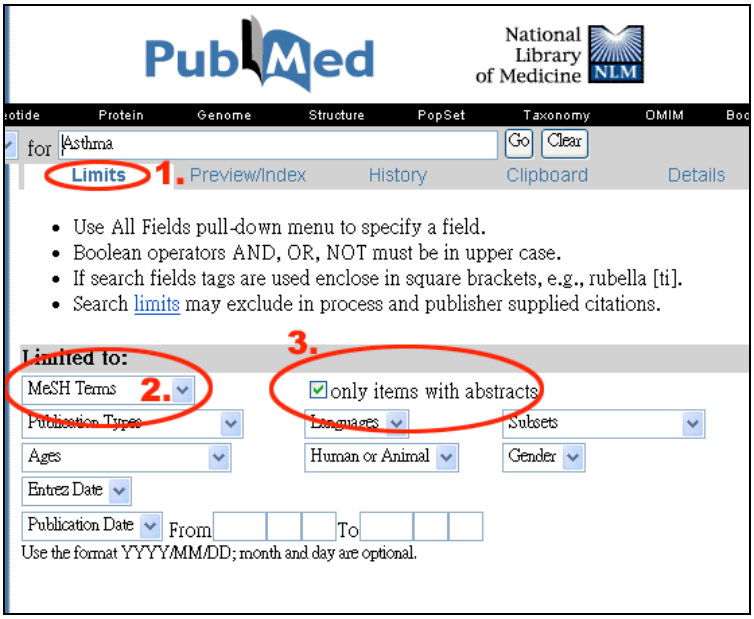

Figure 16 The “limit” screen capture. 1.limit option 2. selecting MeSH term 3.selecting the article if there is abstract

Step2. In PubMed, the system would modify the input query, so we would check the “detail” to make sure there is no wondering rule in it (Figure 17).

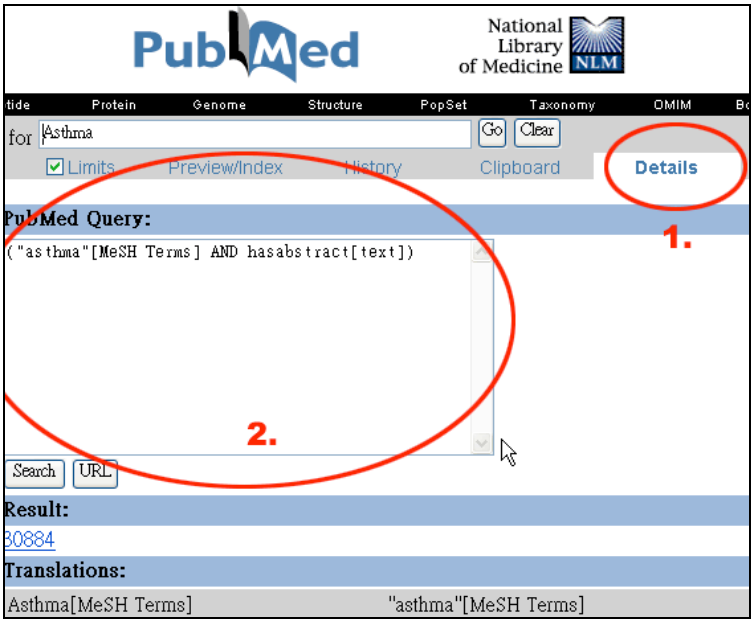

Figure 17 The detail screen snap. 1.Option of details 2. the query in detail

Step3. Selecting the output format (Figure 18).

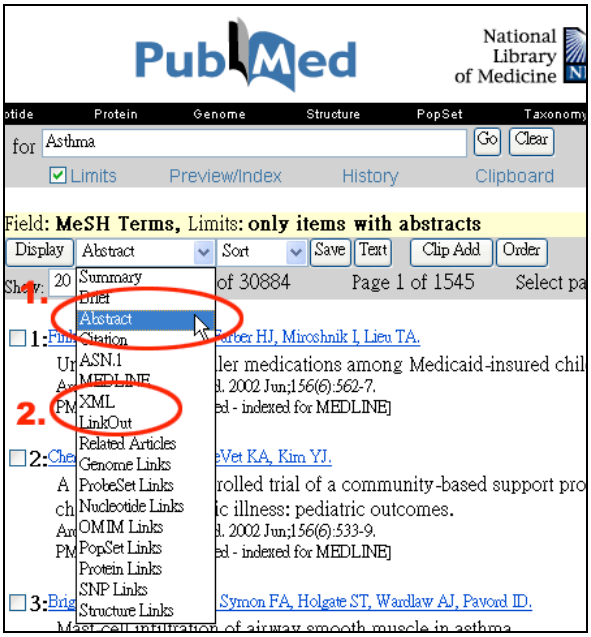

Figure 18 Selecting the output format

Step 4. In PubMed, there is a limitation that the maximum number of items that can be saved is 10000. So we have to use some process to overcome this limitation (Figure 19).

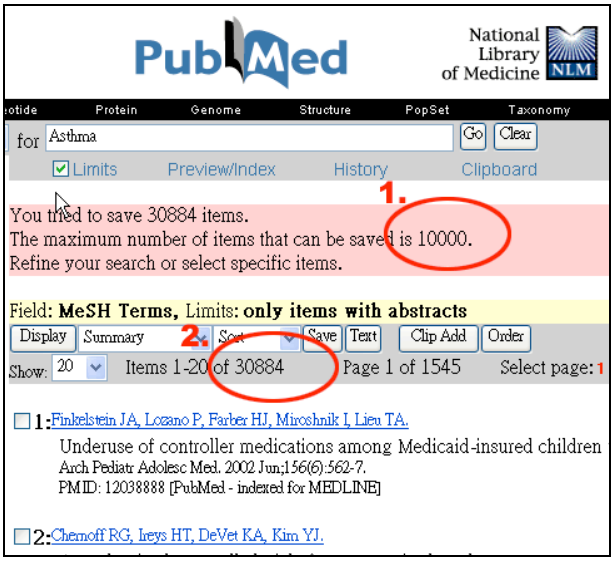

Figure 19 The alert screen snap of warning. The maximum number of download size is 1000.

Step5. In PubMed, the query can bi limited by "Publication Date", so we download the articles year by year (Figure 20).

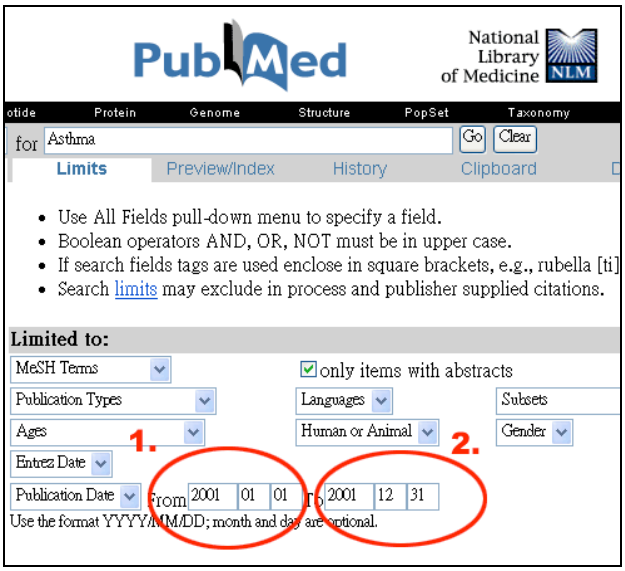

Figure 20 using publication to limit the article number

Step 6. After matching the limitation, we can save the result to local hard disk (Figure 21).

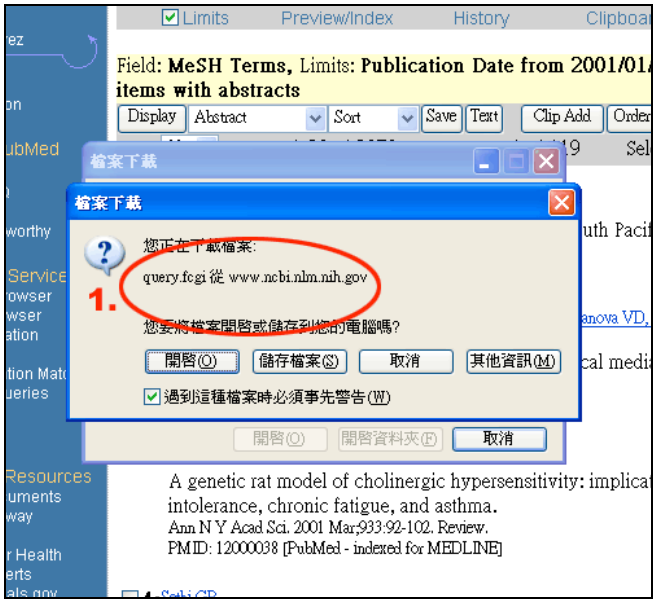

Figure 21 Save to local hard disk

Due to this research, we focus on two diseases category “Asthma”, ”Atopic dermatitis” and “Breast cancer”. We use the following rules to retrieve our result.

1. Collecting “Asthma related article”
2. Collecting “Breast cancer related article”
3. Collecting “Atopic dermatitis related article”

3.1.4 Stored in XML

Recently, the digital library research filed has well-developed technology to exchange the structural information. Especially, the XML structural language is compatible with several major applications, for examples, databases, WEB browse, and statistic tools. Fortunately, the PubMed provides an interface to let user saving the query result in XML format. It is very convenience for users put the result into different kind programs to analysis the data. But the problem is that the querying result saved

with XML format maybe several times larger than the original data, due to the XML tag (Figure 22).

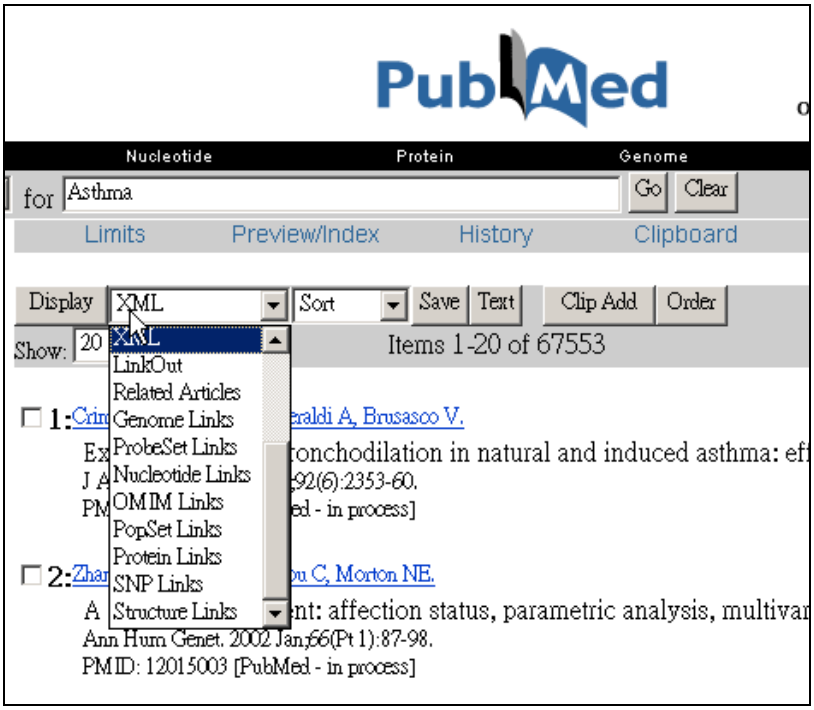

Figure 22 The option of XML

3.1.5 Stored in free-text format

In PubMed, they provide a free-text format to save the querying result.

3.2 Divide article set

In this research, all the articles derived from PubMed are divided into two parts randomly

- Training set – for building causal networks
- Test set – for evaluation

The proportion of training set to test set is 5:1.

3.3 Feature extraction

All the training sets are used to calculate the frequency and each gene symbols

occurs in article once, the frequency of gene is added one. But it is very difficult for computer or human to recognize all the gene names or symbols. Not mention about that one gene could have several different names.

Fortunately, the HUGO offers a detail synonyms map, which can help the researchers mapping synonymous map. In this project, the gene name and gene synonym map are adopted for controlled vocabulary.

The gene terms of PubMed article usually have several synonym terms. Until recent few years, there are official committees like (The Human Genome Organization, HUGO) offering the official gene symbols. Until right now, there are14,180 gene terms available on HUGO website. In the other hand, there are several study groups using lots of methods like syntactic parsing, processing of statistical and frequency information and rule based method to detect the genes names.

By using synonym map, we can derive more correct gene occurrence frequency. Then, we can have a gene-frequency histogram. And, using this rank, we can choose the most frequent gene name to be used our selecting features.

3.4 Feature selection

From above step, we can retrieve more than one hundred gene names. It is too much false positive gene names and too complex for building model. To reduce the false positive error and the complexity, we need perform some feature selection processing steps.

3.4.1 Frequency analysis

All the gene names occur in the article are counted in this research. If one gene name occurs three times in one article, and then the frequency number of that gene are added three. According to the Zipfs' Law, if the lower frequency of gene in histogram

means the gene is less important.

3.4.2 Manual check by human reading

From frequency analysis, we can capture the domain specific "term". There is another problem that some terms are truly related in that domain, but the terms are not actually gene name. So in final step, we have to perform manually check to select the proper genes to represent the feature. For each high-frequency term, we check weather the term is true gene name or not from randomly selected articles, which containing that term. If there are more than 80 percent of terms, which are mismatch appearing in random selected articles, then the term is marked with mismatching.
On the other hand, from the preliminary study, using twenty nodes is the maximum size for running the software. Finally, we select twenty genes from three diseases to represent the articles

3.5 Represent the article

In IR, there are a lot of methods to represent the vector space, for examples, Total Frequency Inverse Document Frequency (TFIDF) or Boolean vector. From the preliminary study, the Boolean vector has a better result and lower complexity. So we choose Boolean vector to represent the article, For example

Article1 {NO,NO,YES,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO}

Article 2 {NO,NO,NO,NO,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO}

Article 3 {NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,YES}

Article 4 {NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,YES}

3.6 Construct the causal network and conditional probability table

3.6.1 Building causal model

In “Causal model”, all the gene nodes are connected to disease node directly without any other interaction link. We use “HUGIN researcher 6.1” to build up the model, and we connect each node to disease node manually (Figure 23).

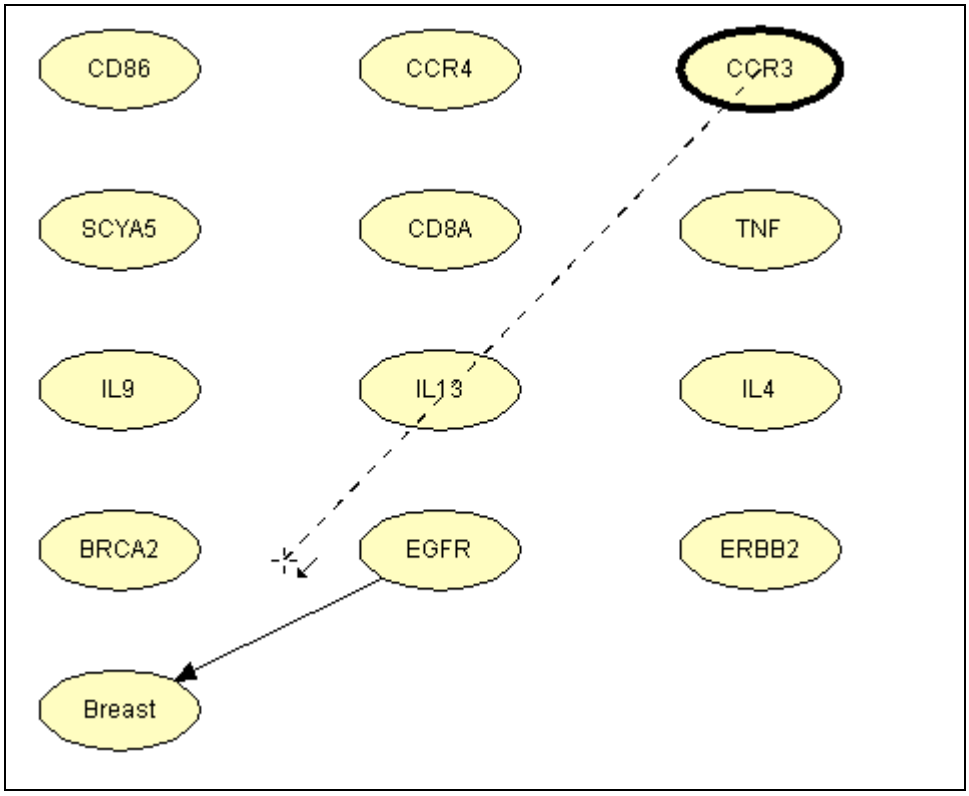

Figure 23 Manually building the “Causal model” in HUGIN researcher 6.1

3.6.2 Building “Structural Learning Model” network

In “Structural Learning model”, the PC algorithm is used to discover the relationship among all the gene nodes and disease node (Figure 24).

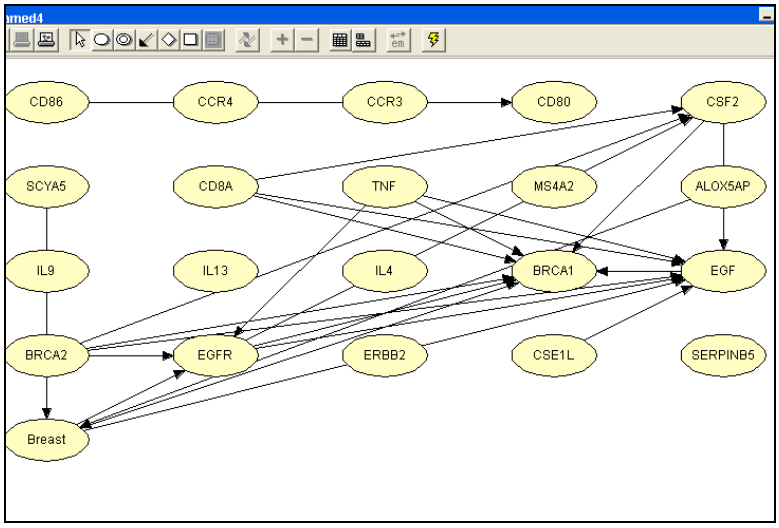

Figure 24 The result of "Structural Learning model"

3.6.3 Adding experience table

Besides constructing the causal network, it is very complicated problem to calculate the probability of Bayesian Network, so we add the learning dataset to the "HUGIN expert 6.1" to construct the conditional probability table (Figure 25).

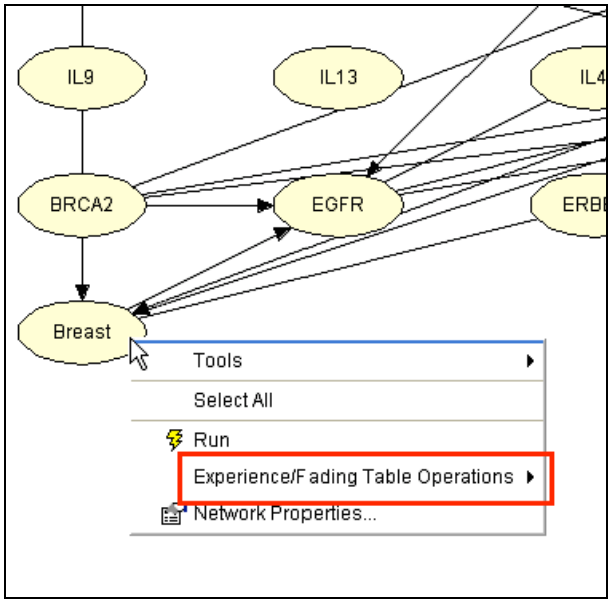

Figure 25 Adding the experience table to each nodes

3.6.4 Calculating the conditional probability

Because we use the "HUGIN researcher 6.1" to calculate the probability of disease state, we have to output the conditional probability table for every possible gene state combination. Then the data is imported to another program (MedScore), which is responsible to find the exam probability (Figure 26).

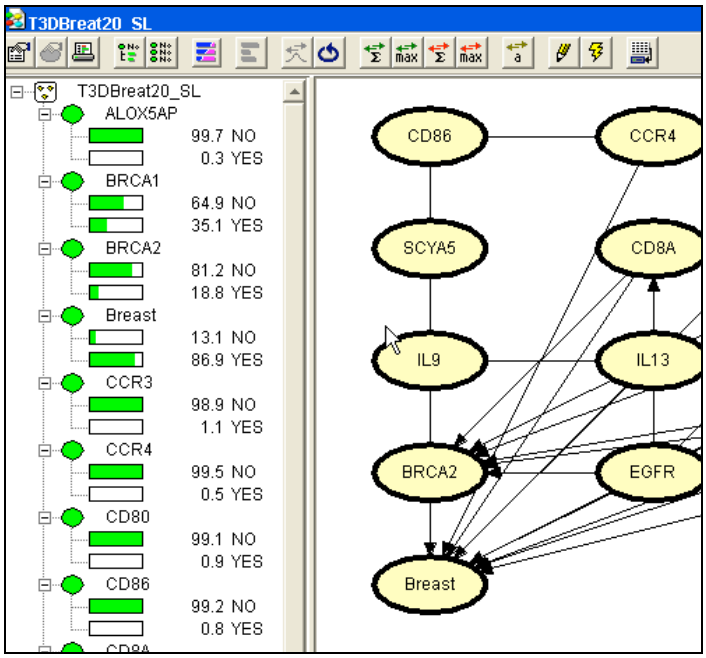

Figure26 Setting each node probability and calculate the disease node probability

3.7 Evaluation two models by test set

After constructing two models of three diseases, we use the test articles, which only contain the gene nodes information without the any disease state information to predict the disease state. All the test articles are collected by querying MeSH term. So we use the MeSH term of articles as our gold standard.

Chapter 4

System Design

Nowadays, the Information Retrieval technology in biomedical research is still at its infant stage. To bridge the gap from collecting the articles to calculate the articles probability, we develop four parts of programs to help this research. At the mean time, the "HUGIN Researcher 6.1" is used for constructing causal network (Figure 27)

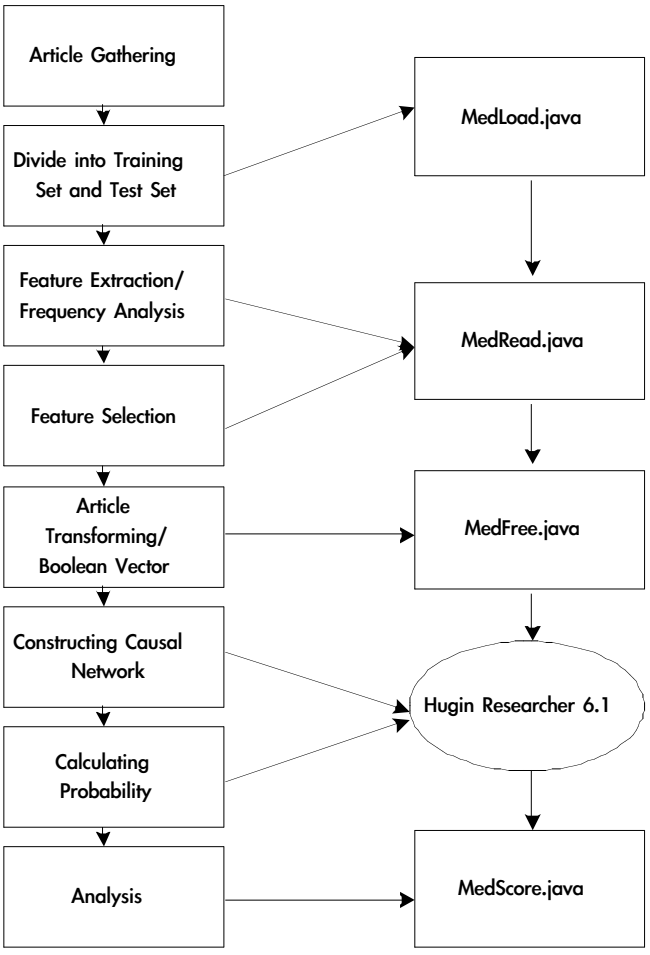

Figure 27 The design of system

4.1 Article gathering: Introduction of program: (MedLoad.java)

In program (MedLoad.java), it is responsible for parsing articles and dividing the articles into training set and test set. All the articles derived from PubMed in this research are more than 10 thousand ones, so we have to download the articles year by year. In the other hand, the articles, which are stored in free-text format are needed to parsed into individual article line by line.

After pre-processing of articles, we divided the collecting articles into training set and test set, therefore the training set is used to build up the probabilistic model, and test set is used for the evaluation.

4.2 Introduction of program: (MedRead.java)

In the program (MedRead.java), it is responsible for extracting the gene symbols within articles and counting the frequency of each gene symbols. The most important thing in this research is to recognize the gene symbols appearing in the articles. And, the HUGO organization provides a detailed gene name and alias. So this program is developed to capture gene name between articles.

4.3 Introduction of program: (MedFree.java)

In the program (MedFree.java), it is responsible for transforming the article into Boolean vector. After feature selection, all the articles is transformed into vector space, which can be read by computer.

4.4 Introduction of program: (HUGIN researcher 6.1)

In the program (HUGIN), it is responsible for building probabilistic model and calculating the probability of articles. At the same time, the probability of articles in

test set is calculated by HUGIN researcher and output the result into probability table. And then, we can use this table to find the suitable score for articles.

4.5 Introduction of program: (MedScore.java)

In the program (MedScore.java), it is responsible for find the matching score from probability table, which is derived from HUGIN researcher.

Chapter 5

Result and Analysis

5.1 Data set

In this research, we focus on the article from PubMed to retrieve the gene- disease relationship. Before this research, we have to define our article set. From the clinical domain knowledge, we choose "Asthma and Atopic dermatitis" two disease category which both belong to hypersensitive disease category. Besides, the "Breast cancer" disease category is selected due to the pathological pathway of Breast cancer is much different that further two diseases.

In PubMed database, the same query method is applied to all three diseases category. The query method is "Query by MeSH term". Then, we can derive three diseases category as following (Table 3)

Table 3 The result of query using three different MeSH term

MeSH term	Query result (number of article)
Atopic dermatitis	4166
Asthma	30911
Breast cancer	71424

5.2 Divide training set and test set

In this research, we hope we can retrieve the implicit gene-disease relationship from

biomedical database, and use this knowledge to build up a probabilistic causal network to predict disease category. So first, we have to divide our article set into two parts – training set and test set randomly. Training set is for build up causal network and the test set is for validating the network.

The proportion of training set to test set is approximate five to one.

5.3 Feature extraction

The first step in applying information retrieval technology is to extract valuable information to describe the knowledge. In this research, all the gene names, which appear in articles, are captured for frequency analysis.

Because the gene names of biomedical article in the past is not unique, so we adopt HUGO (Human Gene Nomenclature Database) as our controlled vocabulary. The Human Genome Organisation [HUGO, http://www.gene.ucl.ac.uk/hugo/] has already approved **14,500** active gene symbols and **11,124** literature aliases and **2,920** withdrawn symbols. [updated on Fri May 3, 2002]

But the problem is that, there are too many gene symbols and there is a lot of mismatching synonym terms. The extracting features by frequency analysis are needed to be further processing (Table 4).

Table 4 The number of extracting feature before feature selection

Disease category	Number of extracting features (Before feature selection)
Atopic Dermatitis	251
Asthma	476

Breast cancer	1123

5.4 Feature selection

Because we want to represent the articles in Boolean vectors, we have to extract the most discriminating feature to represent the article. In this research, we use genes names as the features. But there are too many mismatching gene symbols, we have to perform frequency analysis, background analysis and human check reading to make sure the correct and most discriminating gene symbols.

5.4.1 Manual check by human reading

We use HUGO gene symbols and withdrawn symbols and there are a lot of mismatching symbols. So finally, we have to check every high-frequency captured features to select correct gene name. It is interesting that there are some terms, which are disease-specific. For examples,"RNASE3", it is a specified protein in Asthma; which is high rank in frequency analysis but not appearing in controlled group (Table 5).

Disease	counts	HUGO name	Mismatching alias name
AsthmaRNASE3	742	RNASE3	eosinophiliccationic protein (ECP)
Asthma:COPD	692	COPD	COPD
Asthma:SIASD	470	SIASD	SD
Asthma:PTGFR	437	PTGFR	FP
Asthma:ADM	410	ADM	discovered endogenous vasorelaxing peptideisolated from pheochromocytoma
Asthma:AHR	339	AHR	airway hyperresponsiveness(AHR)
Asthma:PTPR	316	PTPRF	thelate Asthmatic response (LAR) to allergen

Table 5 The “Asthma” frequency table and mismatching alias names

5.4.2 Selection of gene names

In the final step, the most important thing is to select the most discriminating genes names among three diseases and reduce the number of genes. And the same time, we will choose the gene names, which have high rank in frequency analysis and which are not mismatching symbols.

In the other hand, from the preliminary study, using twenty nodes is the maximum size for running the software. Finally, we select twenty genes from three diseases to represent the articles.

Atopic Dermatitis

{CD86 SCCYA17 CCR4 CCR3 CD80 CSF2 SCYA5 CD8A TNF MS4A2 SPINK5}

ATHMA

{LTC4S ALOX5AP GLY16 IL9 IL16 IL13 SCYA17 IL4 CCR3 MS4A2UGB CSF2 ALOX5 IL4R SCYA5 HNMT CD8A TNF}

BREAST CANCER

{BRCA1 EGF BRCA2 EGFR CSF2 ERBB2 CSE1L SERPINB5 WWOX APC PRLR DDT CYP19 ESR1 FGF8 TSG101 FHIT FES COMT HMMR CDH1 RAD51 SNCG}

Selected feature

{ CD86,CCR4,CCR3,CD80,CSF2,SCYA5,CD8A,TNF,MS4A2,ALOX5AP,IL 9,IL13,IL4,BRCA1,EGF,BRCA2,EGFR,ERBB2,CSE1L,SERPINB5,}

5.5 Boolean vector transformation of articles

In this step, we use the selected features to construct the Boolean vector of articles. It

is important that we would eliminate the article which there in no any gene name in it. In table 6, we show the size of training set and test set (Table 6, 7).

Table 6 Number of articles (After transforming to Boolean vector)

MeSH term	Number of articles (After transforming to Boolean vector)	
	Training set	Test set
Atopic dermatitis	79	16
Asthma	236	47
Breast cancer	1799	373

Table 7 The examples Boolean vector of articles

Article1{NO,NO,YES,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO}
Article 2 {NO,NO,NO,NO,YES,NO,NO,YES,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO}
Article 3 {NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,NO,NO,NO,YES}
Article 4 {NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,NO,YES,NO,NO,NO,YES}

5.6 Construct the causal network model

We use HUGIN software to construct the “Structural model” and “Causal model” of Asthma, Atopic dermatitis and Breast cancer (Figure 28, 29, 30).

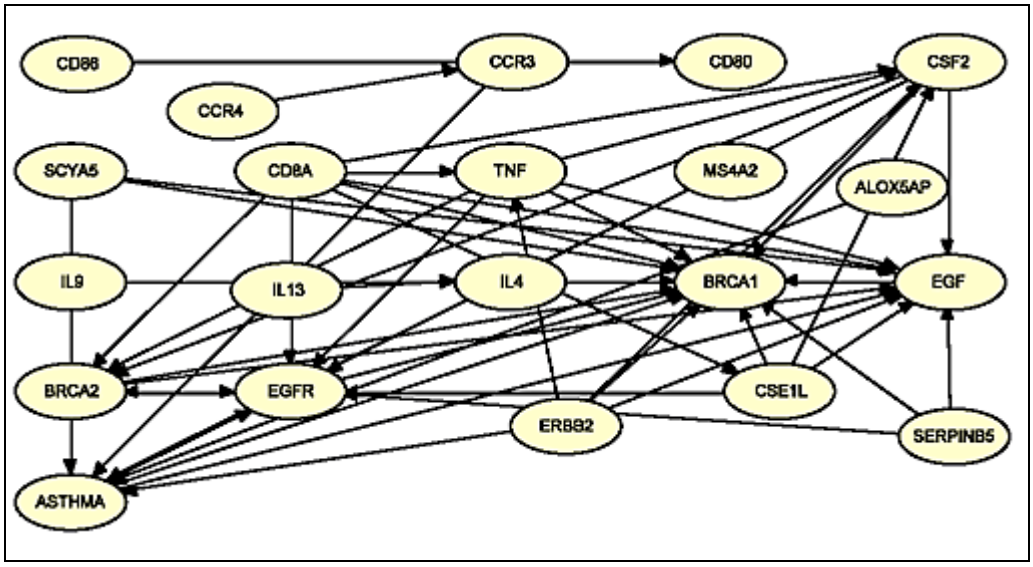

Figure 28 "Structural Learning model" of ASTHMA

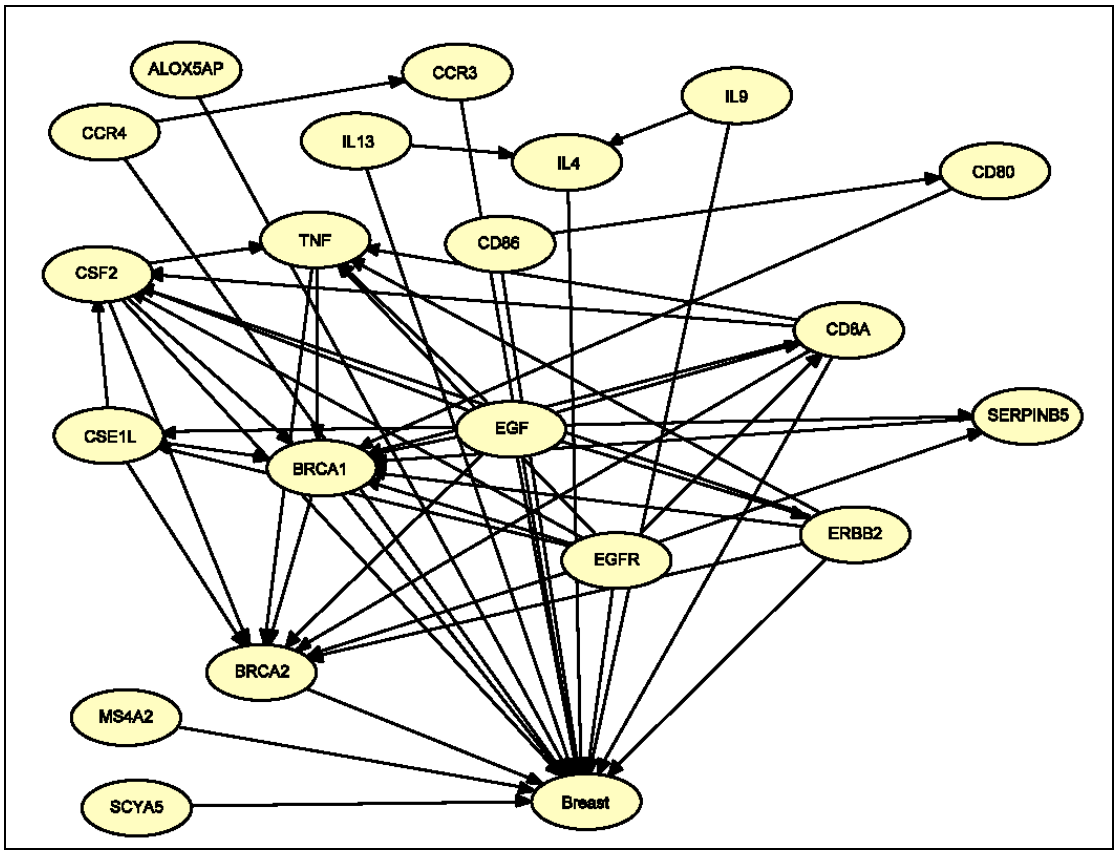

Figure 29 "Structural Learning model" of Breast Cancer

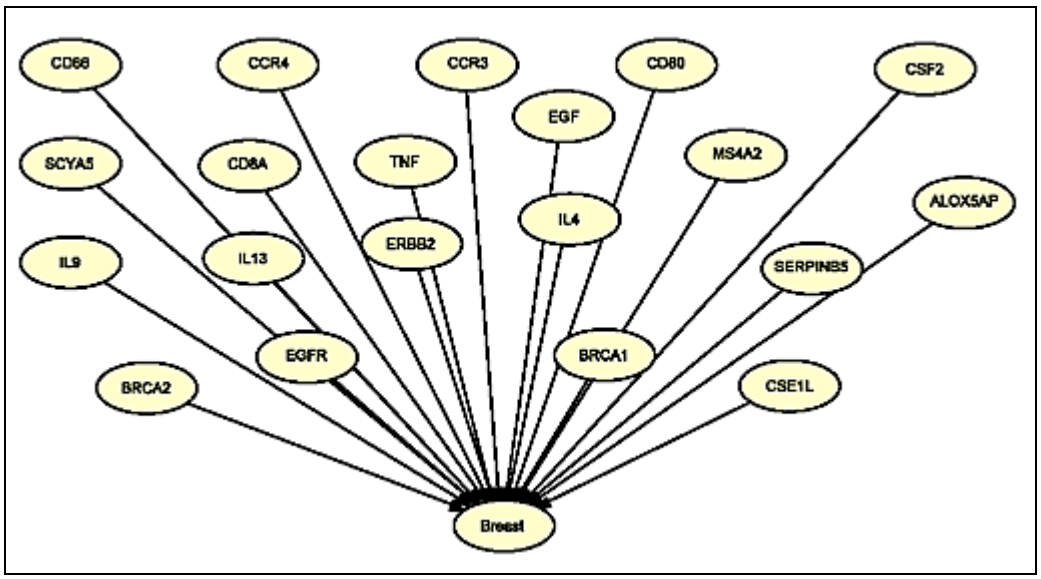

Figure 30 The "Causal model" model of Breast cancer

5.7 Calculating the probability table

To calculate the probability of each article, we use HUGIN software to construct the probability table. Then we can search the probability table to calculate all test set articles probability.

In the following figure, this is the initial state of "ASTHMA Structural Learning model". In each table of gene, we can observe each gene prior probability in our training set.

In this "ASTHMA Structural Learning model", we can observe that the ASTHMA prior probability is "11.6%" (Figure 31).

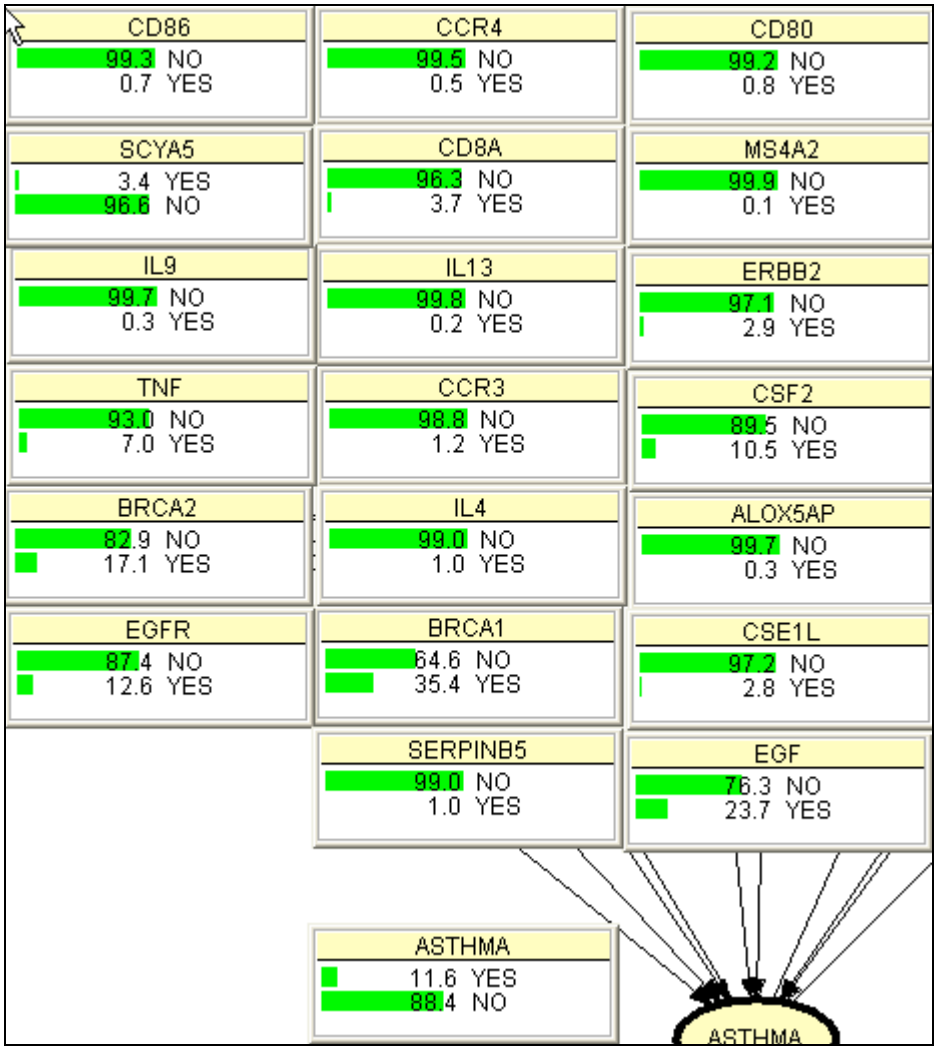

Figure 31 The prior probability of each node.

In the following figure, we set the "Gene SCYA5" state "YES 100%", we can observe that the Asthma probability becomes 74%. Because the SCYA5 is high rank of frequency analysis in Asthma, this result is compatible to our expectation (Figure 32).

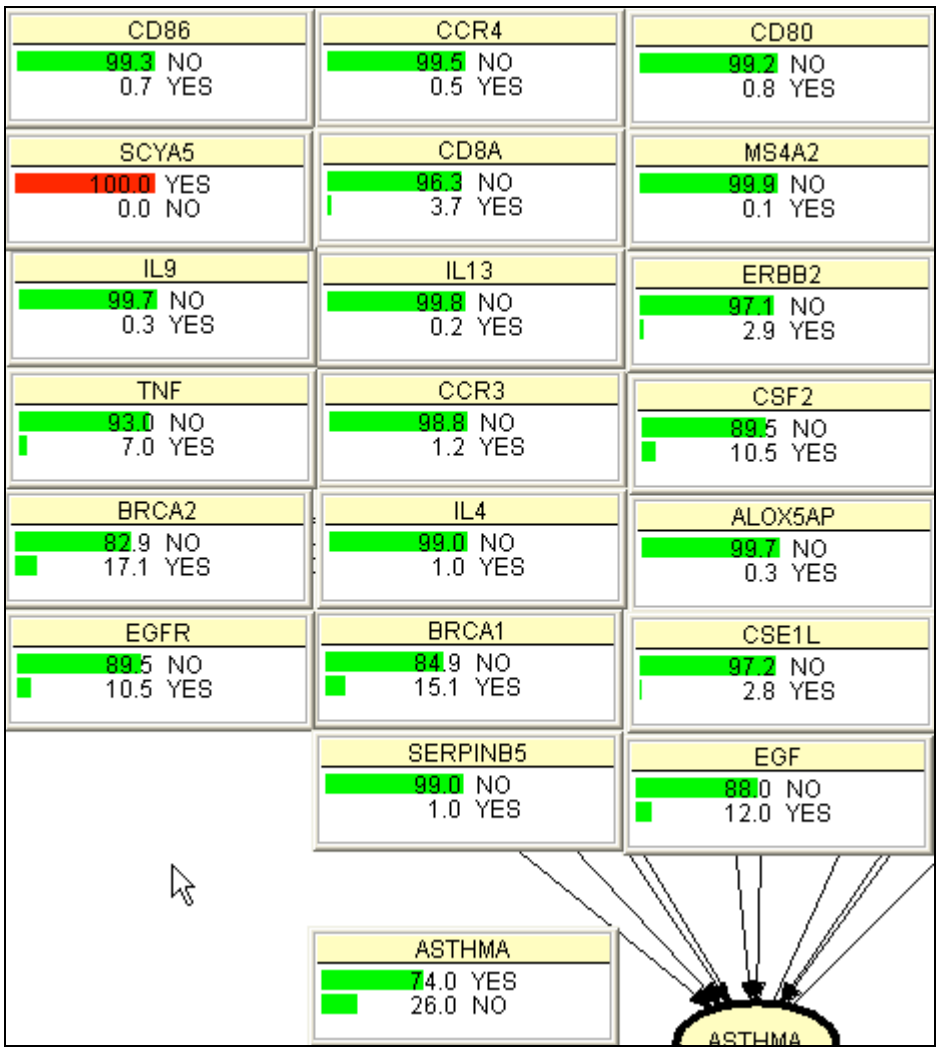

Figure 32 In Asthma "Structural Learning model", setting the "Gene SCYA5" state "YES 100%" and Asthma state "YES 74%"

Then we do another test. In the following figure, we set the "Gene BRCA1" state "YES 100%", we can observe that the Asthma probability becomes 1.3%. Because the BRCA1 is high rank of frequency analysis in "Breast cancer", this result is compatible to our expectation (Figure 33).

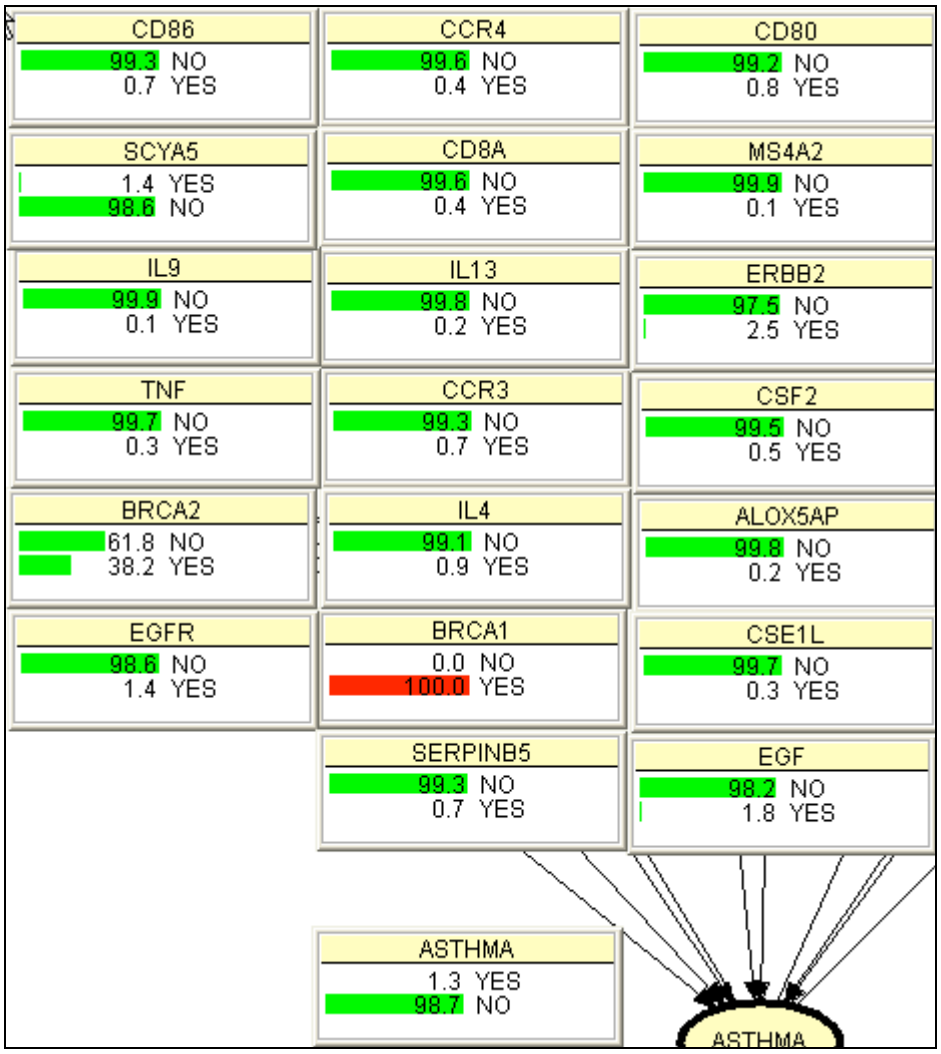

Figure 33 In Asthma "Structural Learning model", setting the "Gene BRCA 1" state "YES 100%", and get "ASTHMA" State "YES 1.3%"

After one gene test, we are interesting when test by two genes, is it still work on this model. In the following figure, we set the "Gene SCYA5 and CSF2" state "YES 100%", we can observe that the Asthma probability becomes 98.6%. Because both SCYA5 and CSF2 are high rank of frequency analysis in "Asthma", this result is compatible to our expectation (Figure 34).

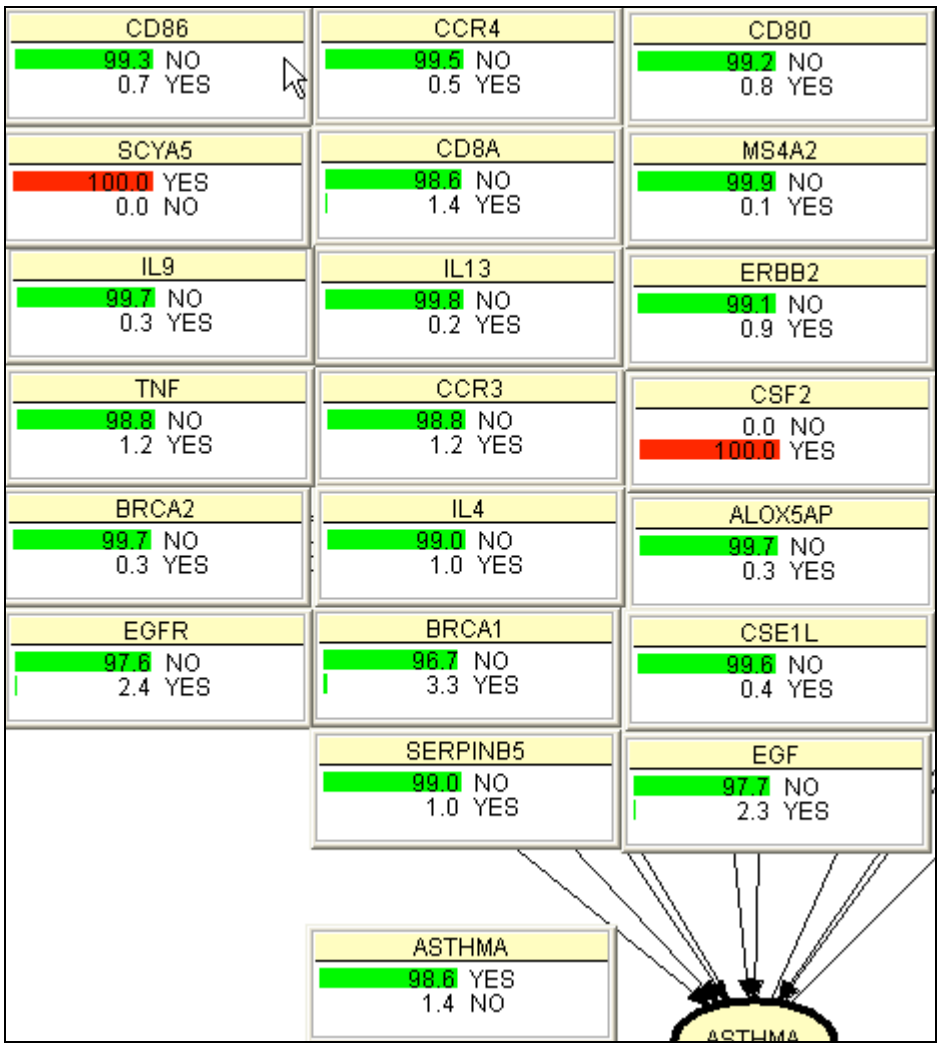

Figure 34 In Asthma "Structural Learning model", setting the "Gene SCYA5, CSF2" state "YES 100%", and get "ASTHMA" State "YES 98.6%"

Because the above experiment is compatible to our expectation, we calculate all the probability of test set. In the following, we will use the probability of each model to evaluate our model.

5.8 Analysis and evaluation

In this project, we build two different kinds of models for each disease to calculate the probability of disease. In figure 35 and 36, we illustrate the histogram of "Breast cancer related articles and not related articles probability in "Structural Learning" model and "Causal model". It is obvious that Breast cancer related articles have higher probability in both "Structural learning model" and "Causal model".

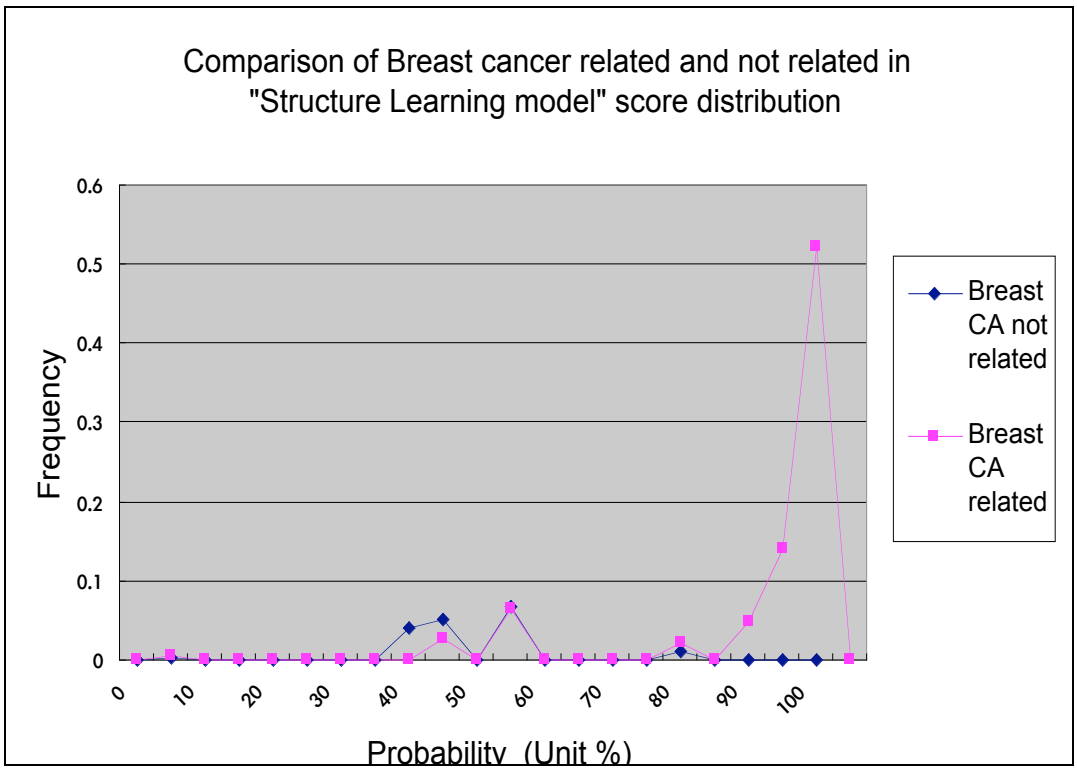

Figure 35. Histogram of "Breast cancer related articles" and "not related articles" probability in "Structural Learning model". The red line represents the "Breast cancer related articles" The peak of frequency appears at probability 90-100%. And the blue line represents the "Breast cancer not related article"; the peak of frequency appears at 40-50% probability.

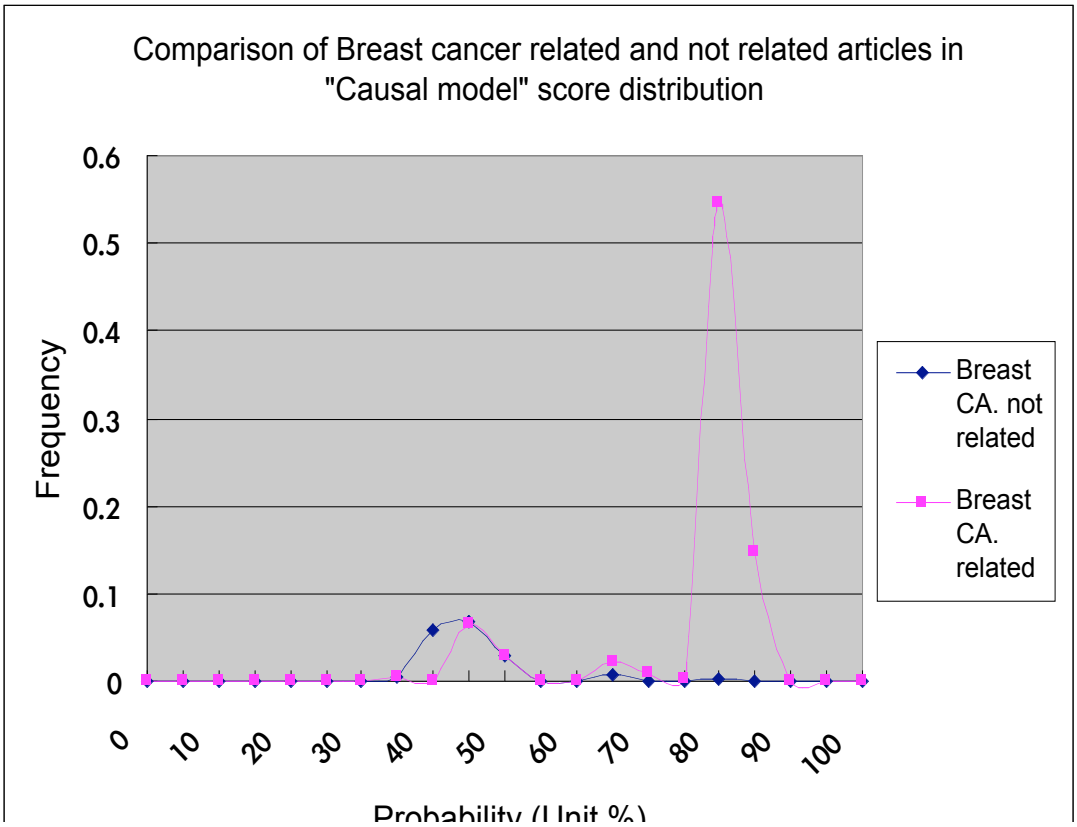

Figure 36 Histogram of "Breast related articles" and "not related articles" probability in "Causal model" breast model. The red line represents the "Breast cancer related articles" The peak of frequency appears at probability 80-90%. And the blue line represents the "Breast cancer not related article"; the peak of frequency appears at 40-50% probability

In figure 37 and 38, we illustrate the histogram of "Asthma related articles and not related articles probability" in "Structural Learning model" and "Causal model". It is obviously that asthma related articles have higher probability in both Structural learning model" and "Causal model"

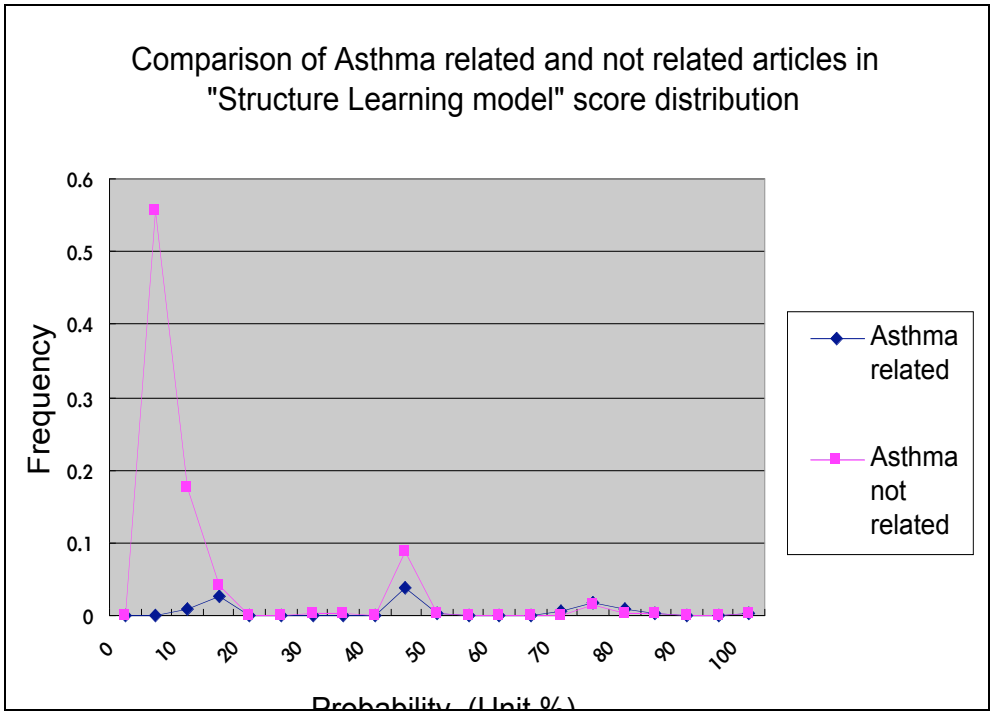

Figure 37. Histogram of "Asthma related articles" and "not related articles" probability in "Structural Learning model" Asthma model. The red line represents the "Asthma not related articles" The peak of frequency appears at probability 10%. And the blue line represents the "Asthma related article"; the peak of frequency appears at 40-50% probability

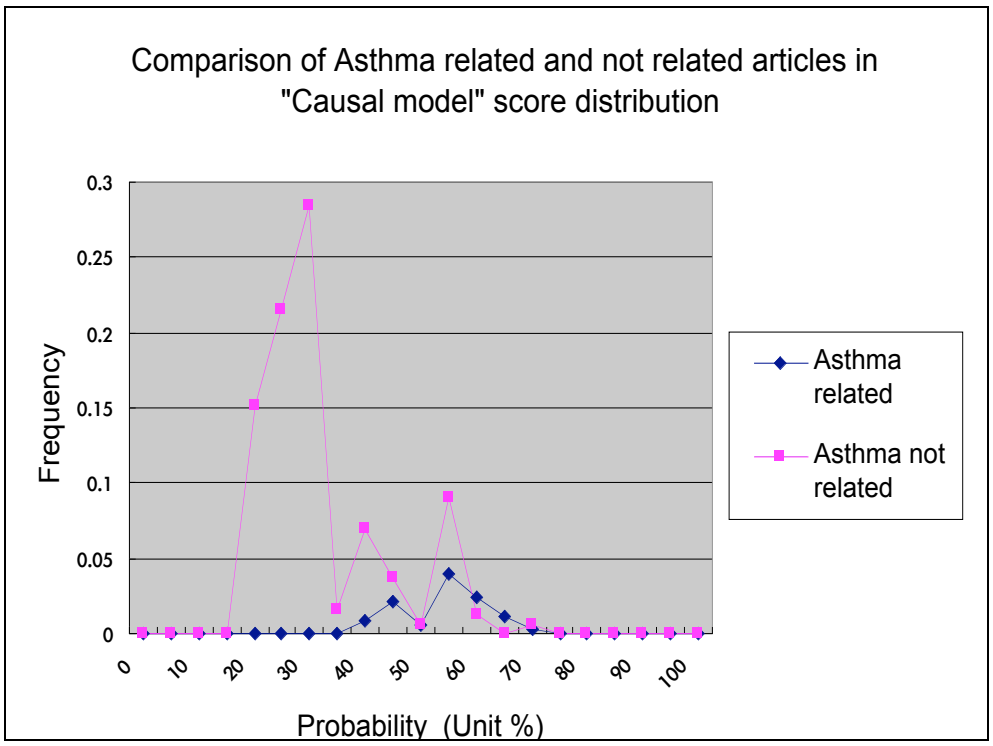

Figure 38. Histogram of "Asthma related articles" and "not related articles" probability in "Causal model" Asthma model. The red line represents the "Asthma not related articles" The peak of

frequency appears at probability 30-40%. And the blue line represents the "Asthma related article"; the peak of frequency appears at 50-60% probability

5.9 ROC curve

In this research, the ratio of training set and test set is appropriate five to one. The training set is used for building up the probability model and test set is used for evaluating the models. In ROC chart, the scores of the 463 test set are re-plotted to show the true positive rate and false positive rate of articles in different models.

In figure 39 and 40, the 463 articles from test set are re-plotted in "Structural Learning model" and "Causal model", which belong to breast cancer disease model. The two probabilistic models of breast cancer related articles are used to calculate the probabilistic score of the articles. In these two graphs, the breast cancer related probabilistic score of the 463 articles from test set are re-plotted to show the TPR and FPR of the prediction of breast cancer articles.

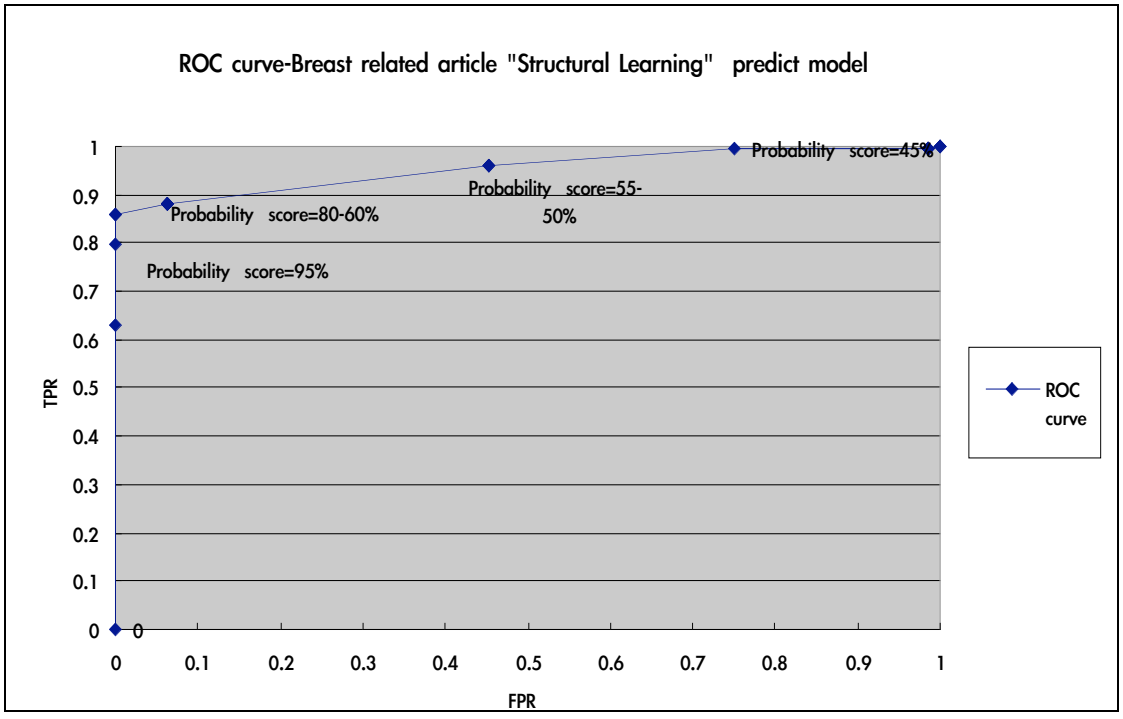

Figure 39 ROC curve of Breast related articles using "Causal model" predict model

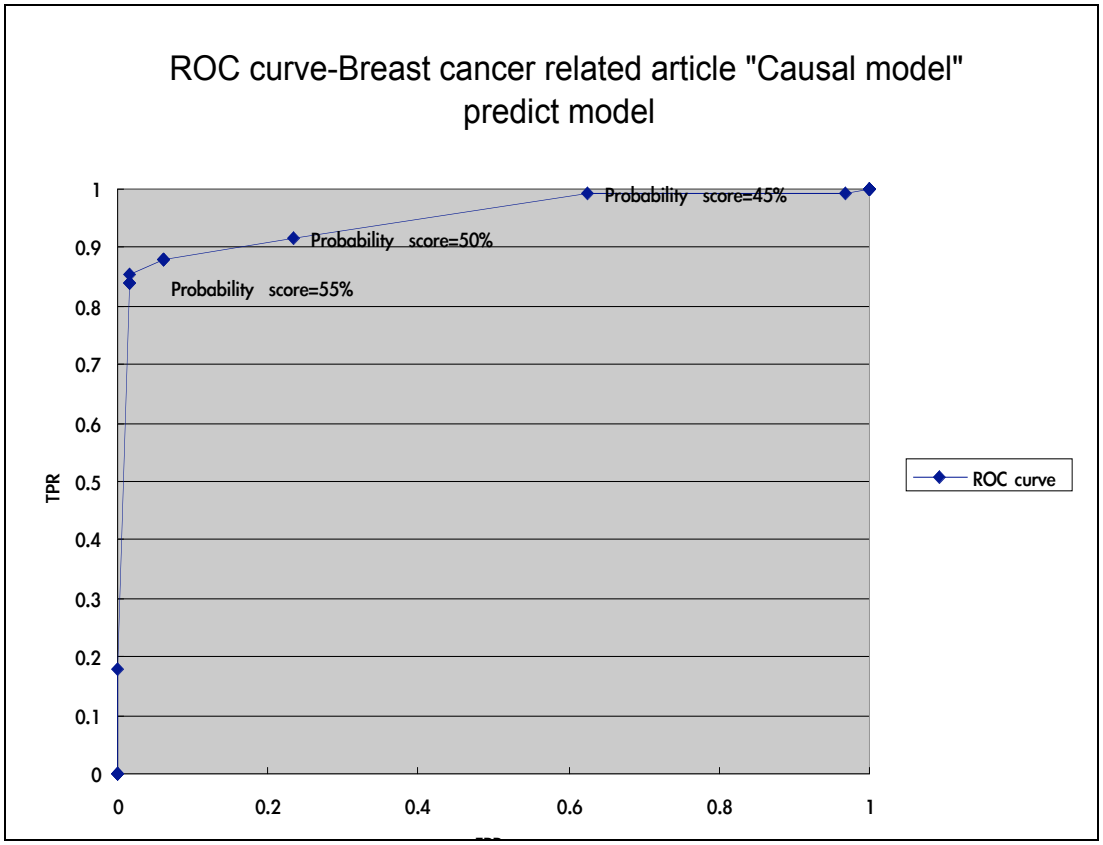

Figure 40 ROC curve of Breast related articles using "Causal model" predict model

In figure 41 and 42, the 463 test set are re-plotted in "Structural Learning model" and "Causal model"; which are belongs to asthma disease model. The two probabilistic models of asthma related articles are used to calculate the probabilistic score of the articles. In these two graphs, the asthma related probabilistic score of the 463 test set are re-plotted to show the TPR and FPR of the prediction of asthma articles.

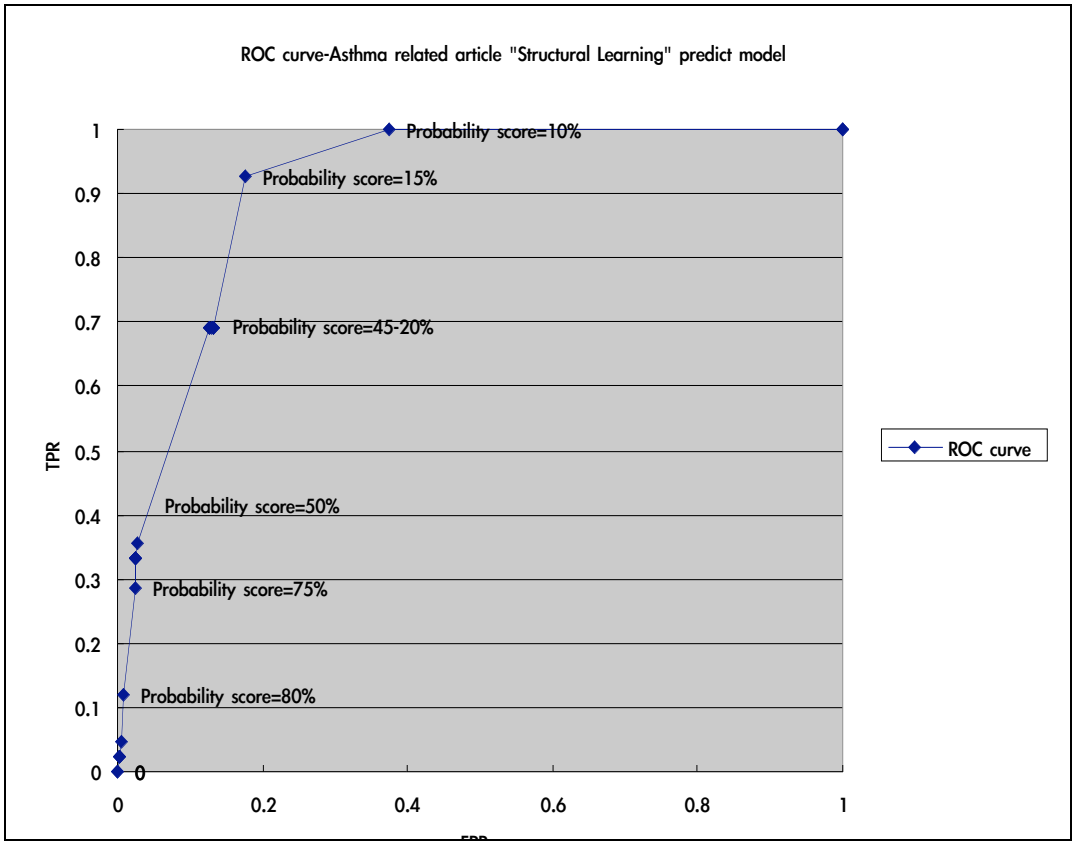

Figure 41 ROC curve of "Asthma related articles" using "Structural Learning" predict model

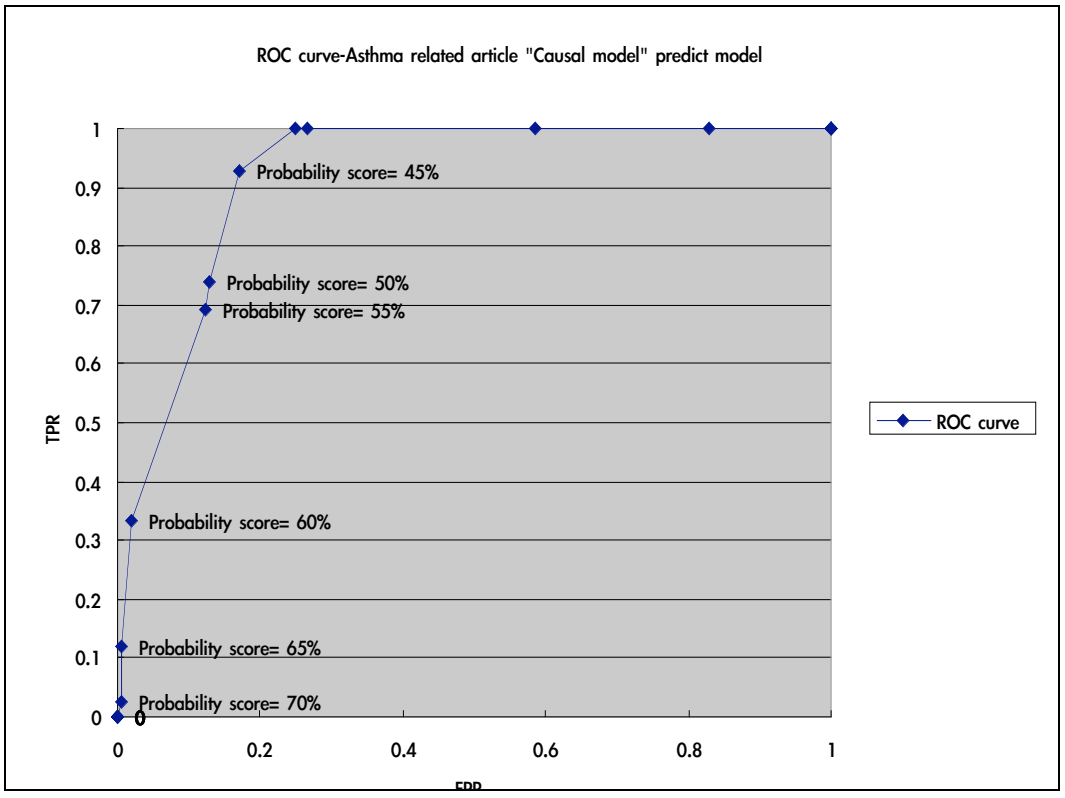

Figure 42 ROC curve of "Asthma related articles" using "Causal model" predict model

Chapter 6

Discussion

In this research, we focus on gene-disease relationship research, and try to build up a causal network to represent the knowledge from biomedical knowledge. From result, we successfully build a probabilistic model to illustrate the genes of "Asthma" disease and "Breast Cancer" disease. In ROC curve analysis, the choosing genes are strong enough to discriminate the two diseases articles.

6.1 Using gene probability to predict the disease state

As expecting, using frequency analysis to capture the discriminating genes from disease articles to build up the probabilistic model can represent the knowledge from biomedical literature database. For examples, the BRCA1 genes is strongly related breast cancer, as a common sense; when we setting BRCA1 gene state "YES 100%", we can observe the Breast cancer state "Yes" turning to 97% and the Asthma state "Yes" turning to 1.3%.

Further more, we can expend this using two genes to inference our model. For example, when we setting "SCYA5" and "CSF2" two genes state "YES 100%", we can observe the Asthma state "Yes" turning to 98.6% and the Breast state "Yes" turning to 4.4%.

6.2 Comparing two causal networks

6.2.1 The precision of “Structural Learning Model” and “Causal model”

In breast cancer and asthma two diseases, the two models can successfully tell from asthma and breast cancer articles. From ROC curve, the Structural Learning model in Breast cancer is better than Causal model, but in Asthma, the Causal model is slightly better.

6.2.2 The comparison of probability between two models

From the probability table, we can observe that the variation of probability in different disease categories is bigger in “Structural Learning Model”. For example, the two peaks in histogram of Breast “Structural Learning Model” are 100% (Breast related articles) and 55% (Breast cancer not-related articles) and the two peaks in histogram of Breast “Causal model” are 80% (Breast related articles) and 45%. In Asthma category, we can also observer the same result. It is mean that if we need the better probability variation, the “Structural Learning Model” offers a better choice.

6.2.3 The comparison of system performance in two models

All the programs and software used in this research is developed by JAVA program language and is running at Microsoft XP platform which using AMD Athlon 800 CPU, 384MB RAM and RAID 1 mode 4 * 45G hard drives (Table 8).

Table 8 The comparison of system performance in two models

	Structural Learning model (Asthma model)	Causal model (Asthma model)
File size	23.4MB	474MB
Open time	10 sec	3 min30 sec
Used RAM	176 MB (Before load)→205MB	176 MB (Before load)→ 245MB
Turn to run mode	1sec	8sec
1 node inference	<1sec	2sec
Rest to initiate state	<1sec	6sec

6.2.4 The comparison of Asthma and Breast cancer related articles in "Structural Learning model"

In directed acyclic graph, the most important thing for system performance is the number of parent nodes to each node, which is shown as figure 43. We also compared the parent node and child node number of each node in Asthma and Breast cancer Structural Learning model (Table 9).

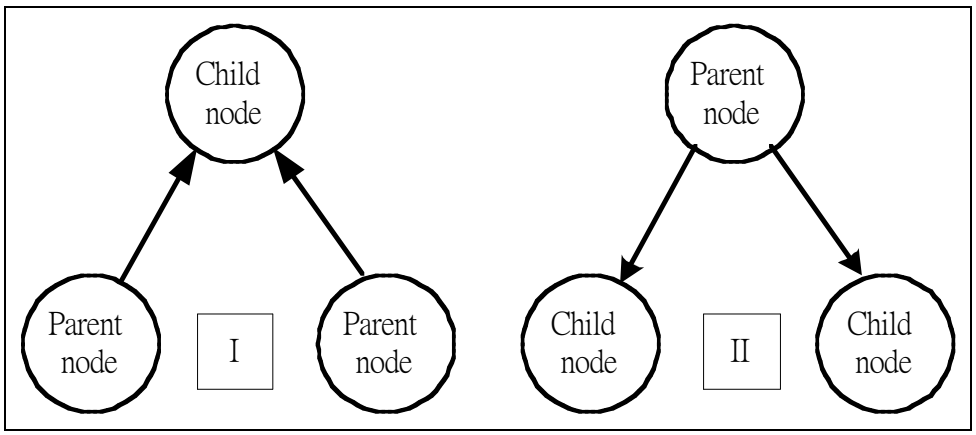

Figure 43 The parent node and child node in directed acyclic graph

Table 9 The comparison of Asthma and Breast cancer related articles in "Structural Learning model"

	Structural Learning model of Breast cancer			*Structural Learning model of Asthma*		
	Parent node	*Child node*	*Total*	*Parent node*	*Child node*	*Total*
ALOX5AP	0	1	1	0	1	1
BRCA1	9	2	11	11(MAX)	0	11
BRCA2	8	1	9	5	3	8
Disease node	*Breast ca* 16(MAX)	0	16	*Asthma* 7	3	10
CCR3	1	1	2	1	1	2
CCR4	0	2	2	0	1	1
CD80	1	1	2	1	0	1
CD86	0	2	2	0	1	1
CD8A	2	2	4	0	7	7
CSE1L	2	3	5	1	5	6
CSF2	5	4	9	4	4	8
EGF	0	9	9	9	1	10

EGFR	0	8	8	7	1	8
ERBB2	1	5	6	0	5	5
IL13	0	2	2	0	1	1
IL4	2	1	3	2	1	3
IL9	0	2	2	0	2	2
MS4A2	0	1	1	0	1	1
SCYA5	0	1	1	0	3	3
SERPINB5	2	1	3	0	3	3
TNF	5	2	7	2	5	7
File size	23.3MB			1.3MB		

From observing the parent node number, the files size is correlated with the 2^n of parent node. In "Causal model" the maximum number of parent nodes is 20, the number in "Breast cancer Structural Learning model" is 16, and in "Asthma Structural Learning model" is 11. The result is compatible with the file size is correlated with the 2^n of parent node (Table 10).

Table 10 The comparison of maximum parent node and file size in three models

	Maximum parent node	File size
Causal model	20	474 MB
Breast cancer Structural Learning model	16	23.3 MB
Asthma Structural Learning model	11	1.3 MB

6.3 Limitations

6.3.1 The size of training set in three disease categories

The number of articles in three different disease categories is quite different. The AD related articles are much smaller than two others. There are only 79 training set of AD, so it is not enough to extract the strong discriminating genes to tell from AD and asthma.

6.3.2 The Pathological pathway in three disease categories

Another reason is that, the AD and Asthma both belong to MsSH "Hypersensitivity, Immediate" category and it is obvious that this two diseases share some the same pathological pathway. Because the size of articles, the IR technology can not tell from them from gene name, we can expect that if there are enough articles in the future, we can build up the suitable to differentiate these two diseases.

In the other hand, the Breast cancer has quite different pathological pathway from two others. Even though, there are some genes appearing in three diseases. For example, "CSF2, alias GM-CSF" and" TNF tumor necrosis factor" two genes frequently appear in all three diseases categories. So some computational tools are needed to calculate the gene likelihood in each disease category.

6.4 Comparisons with other biomedical information retrieval tools

In this research, we use two different kinds of models to describe the gene and disease relationship by probabilistic model. From literature review, there is no other tool combing information retrieval technology and probabilistic model to illustrate the gene and disease relationship.

For example, if we input disease names when using "GeneCards", the web site responses the selecting genes by keyword searching. The result can not provide the

statistic information to give us the additional information. So the method using in this research is combining with IR technology and probabilistic model.

Chapter 7

Conclusion and Suggestion

7.1 The advantage of gene-disease causal network

In the functional genomic era, the biomedical researchers interest not only sequence but also mapping the sequence to human being. So how to explain from genes to diseases or how to link genes to disease by probabilistic model provides a new method to approach functional genomic research.

7.2 Contribution

In nearly future, the price of microarray is getting down quickly and this technology will have a great impact in clinical research. To bridge the gap between microarry and clinical evidence, we need this kind of research to help clinical researchers and biomedical researchers work together.

Especially, the biomedical literature database becomes the more and more important resource in modern science. How to help researchers using those information without pain and more efficiently is the bioinformatics research interesting point, and accelerate genome research.

7.3 Limitations

In this research, all the information and the probabilistic model derive from related articles. So when encountering little information or some less talking about topic, usually the IR technology would drop the valuable information. Furthermore, in biomedical related topic article, they are a lot of domain specific information, for

example gene-gene interaction, gene A inhibit gene B, gene B induce gene C or gene act at location A and some many other. So using IR without Natural Language Processing (NLP), we can not capture some detail information and usually capture too much noise. So this gives us a hint that IR technology is suitable to observe some stable and global information, but is not enough to dig further information.

7.4 Future Work

Nowadays, we do care not only one gene-disease relationship but also multi-gene-disease relationship, therefore there are a lot of effort can be done to improve the scale of this research; for example, expanding the number of disease categories and genes. There are several aspects can be done as following.

7.4.1 Expanding the number of disease categories

In this research, we use only three diseases to evaluate our hypothesis. But in high-throughput microarry, we have to face the whole genome wide human diseases. So it is necessary that expanding the disease category to exam the relationship between diseases. For examples, cancer is the most popular disease in our human life; the clinical researchers are interesting in detecting cancer diseases in one microarry chip. Therefore, how to illustrate the gene-disease probabilistic model from all cancer wide disease is the most important work, which has to be solved. In the future, we hope that all the disease can be included in the same probabilistic model.

7.4.2 Expending the number of genes

Using “Structural Learning Model” to represent the gene disease network is much simpler and lower complexity for computer calculating than “Causal model”. So if we want to expend the number of genes, the “Structural Learning Model” is more

suitable for computation and reduction of the complexity. At the meantime, the PC algorithm is adopted in this research to find the causal relationship, and some refined algorithm; which is derived from original PC algorithm (Peter Spirtes, Clark Glymour) is needed for searching the causal relationship of biomedical literature database.

Reference List

1. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-11, 2000.

2. Andrade MA, Valencia A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics 14: 600-7, 1998.

3. Baldi P, Long AD. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics 17: 509-19, 2001.

4. Blaschke C, Andrade MA, Ouzounis C, Valencia A. Automatic extraction of biological information from scientific text: protein-protein interactions. Proc Int Conf Intell Syst Mol Biol 60-7, 1999.

5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-71, 2001.

6. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol 77-86, 1999.

7. Eckman BA, Kosky AS, Laroco LA Jr. Extending traditional query-based integration approaches for functional characterization of post-genomic data. Bioinformatics 17: 587-601, 2001.

8. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17 Suppl 1: S74-82, 2001.

9. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol 7: 601-20, 2000.

10. Hishiki T, Collier N, Nobata C, Okazaki-Ohta T, Ogata N, Sekimizu T, Steiner R, Park HS, Tsujii J. Developing NLP Tools for Genome Informatics: An Information Extraction Perspective. Genome Inform Ser Workshop Genome Inform 9: 81-90, 1998.

11. Ichikawa K. A-Cell: graphical user interface for the construction of biochemical reaction models. Bioinformatics 17: 483-4, 2001.

12. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28: 21-8, 2001.

13. Kostoff RN, DeMarco RA. Extracting information from the literature by text mining. Anal Chem 73: 370A-378A, 2001.

14. Leung S, Mellish C, Robertson D. Basic Gene Grammars and DNA-ChartParser for language processing of Escherichia coli promoter DNA sequences. Bioinformatics 17: 226-36, 2001.

15. Marcotte EM, Xenarios I, Eisenberg D. Mining literature for protein-protein interactions. Bioinformatics 17: 359-63, 2001.

16. Ohta Y, Yamamoto Y, Okazaki T, Uchiyama I, Takagi T. Automatic construction of knowledge base from biological papers. Proc Int Conf Intell Syst Mol Biol 5: 218-25, 1997.

17. Proux D, Rechenmann F, Julliard L, Pillet V V, Jacq B. Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. Genome Inform Ser Workshop Genome Inform 9: 72-80, 1998.

18. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 14: 656-64, 1998.

19. Rindflesch TC, Tanabe L, Weinstein JN, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput 517-28, 2000.

20. Rzhetsky A, Koike T, Kalachikov S, Gomez SM, Krauthammer M, Kaplan SH, Kra P, Russo JJ, Friedman C. A knowledge model for analysis and simulation of regulatory networks. Bioinformatics 16: 1120-8, 2000.

21. Selley JN, Swift J, Attwood TK. EASY--an Expert Analysis SYstem for interpreting database search outputs. Bioinformatics 17: 105-6, 2001.

22. Spirtes, P., Glymour, C., and Scheines, R., 1993. *Causation, Prediction, and Search*, New York:

Springer-Verlag

23. Usuzaka Si, Sim KL, Tanaka M, Matsuno H, Miyano S. A Machine Learning Approach to Reducing the Work of Experts in Article Selection from Database: A Case Study for Regulatory Relations of S. cerevisiae Genes in MEDLINE. Genome Inform Ser Workshop Genome Inform 9: 91-101, 1998.

24. Wojcik J, Schachter V. Protein-protein interaction map inference using interacting domain profile pairs. Bioinformatics 17 Suppl 1: S296-305, 2001.