Discovering the gene and disease relationship from biomedical literature database by applying the information retrieval technology

Ming-Chin Lin, Yu-Chuan Li, Chien-Yeh Hsu, I-Ren Chiang

Graduate Institute of Medical Informatics, Taipei Medical University, Taiwan

Summary

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 brings 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 clinical 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's almost impossible for clinicians to digest the knowledge. So it's 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.

Keyword:

gene, disease, information, retrieval, biomedical database

Introduction and background

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. 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.

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. It's 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.

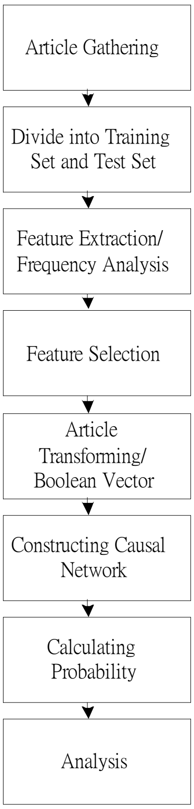
Besides, until right now, there are approximate more twenty thousand genes are discovered, and it's difficult for clinicians to recognize all the symbols in it.

Further more, there are too many genes

describing in biomedical literature, it's very difficult for clinicians to describing which gene is more important and which is less important. Therefore, for clinicians, it's 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.

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 1.



Article Gathering
Divide into Training Set and Test Set
Feature Extraction/ Frequency Analysis
Feature Selection
Article Transforming/ Boolean Vector
Constructing Causal Network
Calculating Probability
Analysis

Figure 1 The research flow

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 pas

2. 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's 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.

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. 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. Form 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,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}

4. 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..

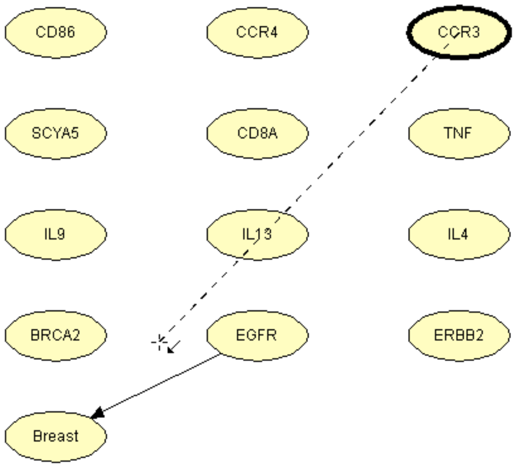


Figure 2 "Building causal model"

5. 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.

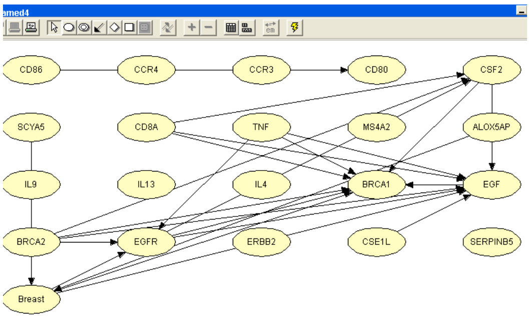
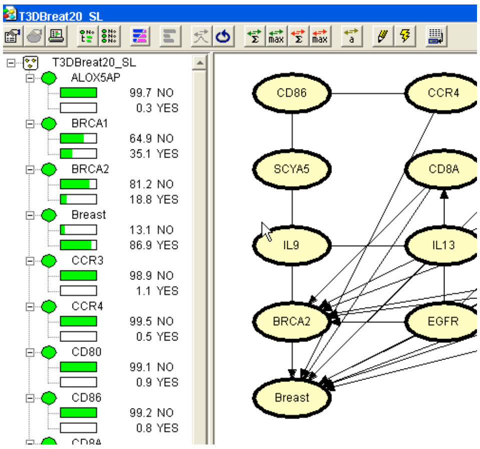


Figure 3. Building "Structural Learning model"

6. 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 score



T3DBreat20_SL
ALOX5AP
99.7 NO
0.3 YES
BRCA1
64.9 NO
35.1 YES
BRCA2
81.2 NO
18.8 YES
Breast
13.1 NO
86.9 YES
CCR3
98.9 NO
1.1 YES
CCR4
99.5 NO
0.5 YES
CD80
99.1 NO
0.9 YES
CD86
99.2 NO
0.8 YES
CD8A

Figure 4. Calculating the conditional probability

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

Result and Analysis

After pre-processing, we divide the article set into two parts, training set and test set (table1).

Table 1. 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

In this project, two model "Structural Learning model" and "Causal model" are build to represent the gene-disease relationship (Figure 5, 6)

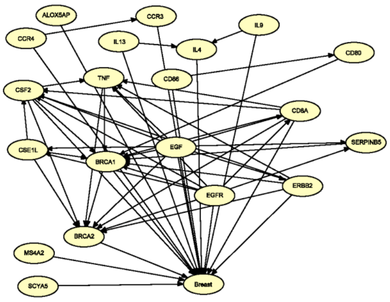


Figure 5 "Structural Learning model"

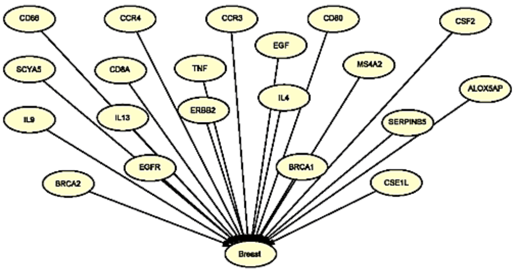


Figure 6 "Causal model"

Finally, we use test set to evaluate the precision of two models. And the, the ROC curve is illustrated to compare the two models (Figure 7).

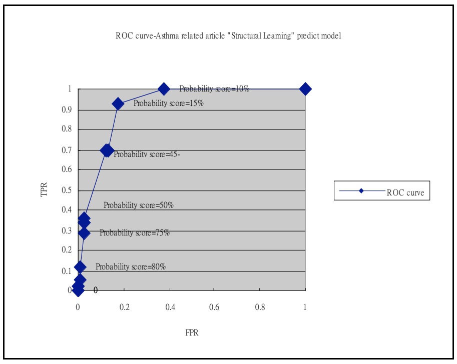


Figure 7. ROC curve of "Asthma related articles" using "Structural Learning" predict model

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

1. Comparing two causal networks

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.

2. The comparison of probability score between two models

From the probability score table, we can observe that the variation of probability score 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's mean that if we need the better probability variation, the "Structural Learning Model" offers a better choice

Reference

1. Alizadeh, A. A.; Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A.; Boldrick, J. C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J. I.; Yang, L.; Marti, G. E.; Moore, T.; Hudson, J. Jr; Lu, L.; Lewis, D. B.; Tibshirani, R.; Sherlock, G.; Chan, W. C.; Greiner, T. C.; Weisenburger, D. D.; Armitage, J. O.; Warnke, R.; Staudt, L. M., and et, a. l. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000 Feb 3; 403(6769):503-11.

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

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

4. Eckman, B. A.; Kosky, A. S., and Laroco, L. A. Jr. Extending traditional query-based integration approaches for functional characterization of post-genomic data. Bioinformatics. 2001 Jul; 17(7):587-601.

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

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

7. Jenssen, T. K.; Laegreid, A.; Komorowski, J., and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001 May; 28(1):21-8.

8. Marcotte, E. M.; Xenarios, I., and Eisenberg, D. Mining literature for protein-protein interactions. Bioinformatics. 2001 Apr; 17(4):359-63.

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

15. Spirtes, P., Glymour, C., and Scheines, R., 1993. *Causation, Prediction, and Search*, New York: Springer-Verlag