Genomics & Informatics Vol. 2(3) 131-133, September 2004

MediScore: MEDLINE-based Interactive Scoring of Gene and Disease Associations

Hye-Young Cho, Bermseok Oh, Jong-Keuk Lee, Kuchan Kimm and InSong Koh*

Division of Epidemiology and Bioinformatics, National Genome Research Institute, National Institute of Health, 5 Nokbun-Dong, Eunpyung-Gu, Seoul 122-701, Korea

Abstract

MediScore is an information retrieval system, which helps to search for the set of genes associated with a specific disease or the set of diseases associated with a specific gene. Despite recent improvement of natural language processing (NLP) and other text mining approaches to search for disease associated genes, many false positive results come out due to diversity of exceptional cases as well as ambiguities in gene names. In order to overcome the weak points of current text mining approaches, MediScore introduces statistical normalization based on binomial to normal distribution approximation which corrects inaccurate scores caused by common words not representing genes and interactive rescoring by the user to remove the false positive results. Interactive rescoring includes individual alias scoring for each gene to remove false gene synonyms, referring MEDLINE abstracts, and cross referencing between OMIM and other related information.

Availability : Mediscore is freely available from http://www.ngri.re.kr/mediscore.

Supplementary Information : http://www.ngri.re.kr/mediscore/manual

Keywords: interactive scoring, MEDLINE, text mining

Introduction

Research in biomedical science in the past decades has generated a large volume of data stored in databases such as MEDLINE, mentioning the relationships between genes and diseases. When someone wants to study genes related with a specific disease, there are two popular ways one may take. One is to search the related literatures manually through MEDLINE keyword search. The huge amounts of literatures, however, make it impractical for one to find out the literatures of interest in this way. Another is to hunt up the list of genes from various disease-related databases for a specific disease, but the availability of data in these databases is limited. More improved way is to list up the relative associations based on biomedical literatures using some computational tools, which crudely mimic manual search. Currently there seems to be two directions for this approach. First is to use so called natural language processing (NLP) which intends to extract meanings from sentence structures. Due to huge diversity of exceptions and the structural ambiguities of biomedical literature text, however, NLP still has a certain limitation and has not been used practically yet (Friedman *et al.*, 2001). Second is to apply text mining and various algorithms to score frequency of gene and disease co-occurrence in the literature database (Marcotte *et al.*, 2001; Chaussabel and Sher, 2002; Perez-Iratxeta *et al.*, 2002; Hu *et al.*, 2003). This method is helpful, but still produces significant number of false positives due to ambiguities of gene symbols.

In this note, we present a complementary way of correcting the text mining results to reduce false prediction rates. We developed an information retrieval system, MediScore (***MEDLINE***-based ***I***nteractive ***Scor***ing of gene and disease associations), which lists up the possible sets of genes associated with a user-selected disease or sets of diseases associated with a user- selected gene with much more improved accuracy.

Methods

MediScore was implemented as follows: First, we indexed all aliases of genes and those of diseases. And then their results were formatted as a relational database. We counted the number of co-occurrence of each gene against all diseases. Finally we applied a statistical normalization to the counts of co-occurrences in order to get the normalized association scores.

MediScore Database

Each gene description, *i.e.* official/preferred symbol and its full description came from the LocusLink. The list of 60 disease terms were selected from MeSHs (Medical

*Corresponding author: E-mail insong@ngri.re.kr, Tel +82-2-380-1416, Fax +82-2-354-1063
Accepted 15 August 2004

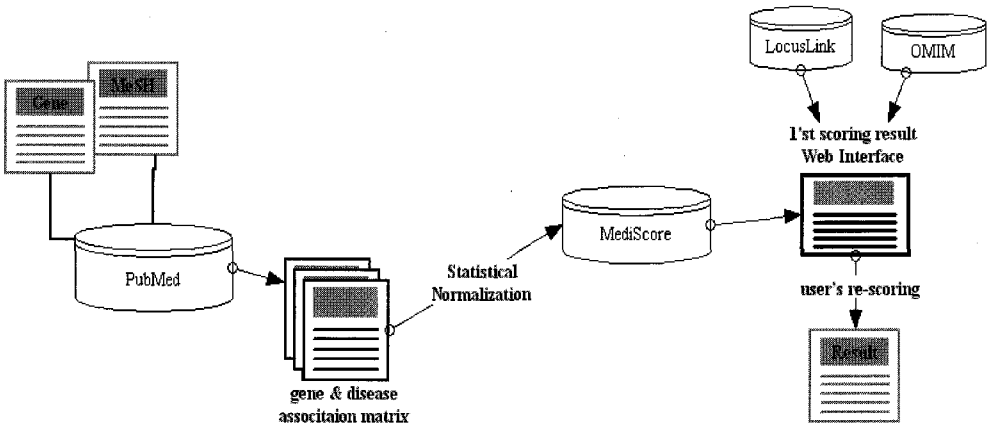

Fig. 1. The system flow of MediScore

Subject Headings) based on the priorities of our research interests and limitation of available hardware capability. All the information above was formatted as a relational database, and the OMIM, PubMed and some additional public gene information were referenced as hyperlinks.

Text Mining Algorithms

First, we indexed all occurrences of a set of synonyms for all genes from MEDLINE abstracts. We made each alias of a gene to have a number of PubMed IDs. Therefore, when a certain gene is selected, all of its aliases and their PubMed IDs are also integrated. In this manner, all the related PubMed IDs are indexed for a certain gene. In case of disease term, after being indexed all disease terms individually, including MeSH, they are also integrated. These results are constructed as matrices of the gene and disease associations per each decade. This approach assumes that most co-citations describe a positive association (Hu *et al.*, 2003). And then all the scores are normalized by the statistical manner as described below.

Statistical Basis

Let N_A be the total number of MEDLINE abstracts and N_G be the number of abstracts mentioning a specific gene one or more times, N_D be the number of those mentioning a specific disease likewise, and N_S=$N_{G\&D}$ be the number of abstracts in which gene and disease names co-occur. Then the probability of co-occurrence of a specific disease and a gene in one abstract $p_{G\&D}$ is

$$p_{G\&D} = \frac{N_S}{N_A}$$

Since the event of co-occurrence only belongs to one of the two cases, happened or not happened, its probability distribution takes binomial distribution. The binomial distribution denoted by *Bin(n,p)* is specified by the number of observations *n* and the probability of occurrence *p*. The expected value and variance for the binomial distribution *Bin(n,p)* are *E(x)=np* and *Var(x)=np(1-p)*. Since the binomial distribution approximates normal distribution for large enough *n* (Central Limit Theorem), the mean and variance of the *Bin(n,p)* can be approximated as $\mu=np$ and $\sigma^2=np(1-p)$. As we are interested in highly non-random events, we apply standardization of normal distribution, and then list the high *Z* score cases meaning non-random events.

$$Z=\frac{X-\mu}{\sigma} \sim N(0,1).$$

In order to get the normalized association score *A*, the expected value *E(X)* which represents the portion of random co-occurrences in the total number of co-occurrences N_S, is subtracted from N_S and then its result is to be divided by the standard deviation *sd(X)*. Applying these, we get the normalized association score *A*.

$$A=\frac{X-E(X)}{sd(X)}=\frac{N_S-np}{\sqrt{np(1-p)}}=\frac{N_S-N_S\cdot p_{G\&D}}{\sqrt{N_S\cdot p_{G\&D}(1-p_{G\&D})}}.$$

The larger value of *A* means the less random co-occurrence, thus indicating strong association of gene and disease. Through this statistical normalization, we could filter out random co-occurrences with a non-specific high *Ns* score caused by the general words used for gene symbols such as MASS(2200), CELL(1057), UP(7378), MICE(4280), IMAGE(64589) and FAT(2195) - LocusLink ID in parentheses.

Interactive Re-scoring

After a user gets the list of candidate genes and their

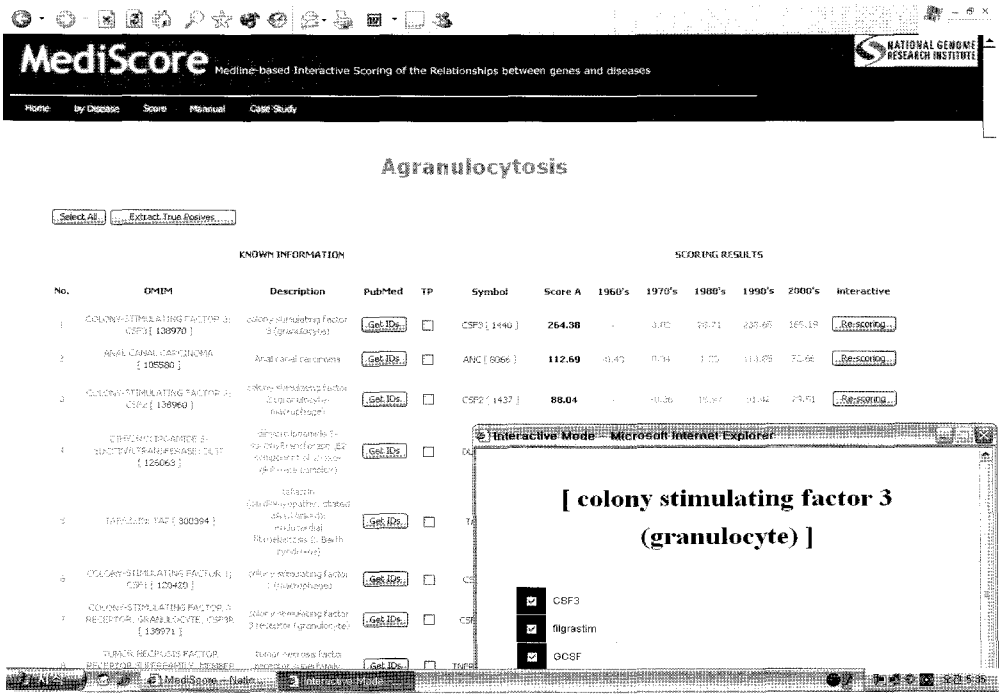

Fig. 2. Snapshots of MediScore on web interface.

scores that the MediScore system has recommended, the user does the interactive re-scoring. In this step the user verifies whether the level of associations or ranks of the candidate genes are true or false with the help of various information offered by the system. First, the user may recheck original gene meanings represented not by gene symbols but by full descriptions of genes. In many cases this simple procedure is enough to filter out some false positives. Second, the user may remove false positive results through individual alias scoring for each gene to remove false gene synonyms. For example, gene symbol ER stands for two different genes *i.e.* estrogen receptor and epiregulin. This kind of gene symbol confusion may be rechecked by scoring with individual gene symbol. Abrupt change in association score helps the user to determine true or false gene symbols. Third, MediScore also suggests definition of a disease associated with a gene based on the OMIM database. Additionally, the user may actually take a look at some MEDLINE abstracts to check real usage of gene synonyms in the context, which is also supported by the system through hyperlinks. Through this series of verification procedures, the user may eventually get a highly accurate list of genes associated with a specific disease.

Acknowledgement

This work was supported in part by Health Foundation, Korea and in part by the intramural fund of National Institute of Health, Korea.

References

Chaussabel, D. and Sher A. (2002). Mining microarray expression data by literature profiling. *Genome Biol.* 3(10):RESEARCH0055.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetscky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17, Suppl. 1, S74-S82.

Hu, Y., Hines, L.M., Weng, H., Zuo, D., Rivera, M., Richardson, A., and LaBaer, J. (2003). Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.* 2(4), 405-412.

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

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