American Journal of Biochemistry and Biotechnology 1 (3): 145-152, 2005
ISSN 1553-3668

A New Text Mining Approach for Finding Protein-to-Disease Associations

Hisham Al-Mubaid and Rajit K Singh
University of Houston, Clear Lake, Houston, Texas 77058

Abstract: Discovering significant relationships between biological entities from text documents is an important task for biologists in order to develop biological models for research and discovery, especially with the existing gigantic amounts of biomedical documents and the rate at which they are increasing everyday. We propose a new text mining method to extract associations between biological entities from text documents; and we focus and apply the method in our experiments on discovering proteins-to-diseases associations. The proposed method uses two sets of documents on the topic of interest [a negative set and positive (or relevant) set] and utilizes the concepts of expectation (ex), evidence (ev) and Z-scores in combining positive and negative evidences in determining the significant associations. Moreover, the method offers an efficient way to handle protein names, aliases and abbreviations and to disambiguate them from common abbreviations, gene symbols and such. We evaluated the method in discovering protein-to-disease associations from *Medline* abstracts and the results are very encouraging. We confirmed the correctness of the results, in each experiment, through articles from *Medline*. Our method was able to discover associations between certain proteins and various diseases like *Alzheimer*, *Creutzfeldt-Jakob*, *Crohn Disease*, *Dengue*, *Jaundice*, *Lung cancer* and more. For example, in *Alzheimer* test, the method ran on 83,933 abstracts and discovered that *Alzheimer* has significant association with 6 proteins, among them, Amyloid beta A4 protein precursor, Apolipoprotein E precursor and Presenilin 1 [PMIDs: 8596911, 1465129, 8346443, 12614323, 8766720 and 8878479]. We further tested our method on some already discovered and published relationships between genes and diseases and the method was also successful in supporting those discoveries.

Key words: Biomedical text mining, information extraction, text mining, bioinformatics

INTRODUCTION

The current biomedical repositories of text data and research papers and articles are extremely huge and growing at a very high and unprecedented rate[1-4]. Massive wealth of knowledge is embedded in these texts and waiting to be discovered and extracted. Thus there is a great need for efficient and effective natural language processing (NLP) and text mining techniques to process these texts in order to extract knowledge and discoveries significant for the advancement of science[4-6] and to support the scientific phenomena and discoveries. Specifically, accurate and efficient approaches for discovering relationships between important biological entities, for example proteins-to-diseases associations, from texts are important for biologists to develop biological models for research and discovery.

In this study, we present a new text mining method to extract associations between proteins and diseases from biomedical texts. We utilize information-theoretic concepts of *expectation* (ex), *evidence* (ev) and *Z-score,* which are based on term counts and co-occurrences, to determine significant associations. The method also uses two sets of documents: a set of *positive* (relevant) documents and a set of randomly-selected *negative* documents. Furthermore, the method uses a protein name dictionary and offers an efficient way to handle protein names, aliases and abbreviations and to disambiguate them from common abbreviations, gene symbols and such.

The method was implemented and evaluated extensively. We conducted several tests for discovering protein-disease associations from *Medline*[1] abstracts and the experimental results are very encouraging. Moreover, we confirmed the correctness of the resulting discoveries, in each experiment, through *Medline* articles. The method was able to discover associations between certain proteins and various diseases like *Alzheimer, Creutzfeldt-Jakob, Crohn Disease, Dengue, Huntington, Jaundice, Lung cancer, Spinal Cord Injuries and* more. For example, in *Alzheimer* test, the method ran on *83,933* abstracts and discovered that *Alzheimer* is associated with 6 proteins, among them: *Amyloid beta A4 protein precursor, Apolipoprotein E precursor* and *Presenilin 1*; these results were verified from Medline documents: PMIDs: 8596911, 1465129, 8346443, 12614323, 8766720 and 8878479. Moreover, we examined the method in discovering relationships between various biological terms, like gene-disease and in verifying and supporting discoveries already published in the literature; for example, the method was

Corresponding Author: Hisham Al-Mubaid, University of Houston, Clear Lake, Houston, Texas 77058, Tel: +12812833802

able to discover (support) the relationship between gene *RUNX1* and "*acute myeloid leukemia*"[4]. However, the focus of this work is on applying the method particularly on protein-disease associations as our literature review indicated that this particular type of association has not been well investigated in the past.

Related work: The usage of text mining in the biomedical domain was useful in many applications. A lot of work has been done including *concept term extraction*[7], association *rules discovery*[8,9] and *extracting relationships between various concepts*[4,6,10-12]. Also, several natural Language processing (NLP) techniques were applied to biomedical documents, for example, in information extraction, extracting gene and protein interactions, named entity recognition (NER)[13], protein-gene names disambiguation[3,14] and more.

Although a lot of good research has been conducted for extracting important associations and interactions between various biological entities[2,4-6,10-13,16], discovering protein-disease associations, in particular, has not been investigated well in the literature. For example, Adamic *et al.*[4], has presented a statistical approach for discovering groups of genes related to a given disease. They also provide a way to treat alias symbols and to disambiguate gene symbols from other abbreviations. Their method had identified most breast cancer genes and identified many additional genes that have been tied to breast cancer in the literature. Srinivasan[5] presented open and closed text mining algorithms that are built within the discovery framework established by Swanson and Smalheiser[15,16]. The algorithms represent topics using metadata profiles and generate ranked term lists where the key terms representing novel relationships between topics are ranked high. In another research work[6], a relationships network was constructed between biomedical objects by identifying the object co-occurrences within all available *Medline* records. This method[6], identified all possible implicit relationships starting from concept of interest *A* to the concepts *B* using co-occurrences, then to the concepts *C* also using co-occurrences. This yields a huge number of (*implicit*) relations. They borrow from Fuzzy set theory to model relationships as probabilistic between 0 and 1. Then every two nodes *A* and *C* connected via an "intermediate" node *B* were compared against random network model. This method[6] is different than ours in that, they created a network of relationships between various types of objects from all *Medline* abstracts, where the objects are primary names and synonyms for genes, diseases, phenotypes and chemical/pharmaceutical compounds (*e.g.* they found 3,482,204 relations between objects for a database of 33,539 objects). In our work, we focus on proteins; we collected more than *66080* primary names and synonyms for proteins only (*the gene dictionary contains also another ~32803 distinct gene names*). The method described[17], first extracted gene names from articles' titles and abstracts and then identified the ones relevant to a particular set of keywords, which are assumed to be related with a particular disease and represented the relations as a giant graph. Finally they partitioned the giant graph into smaller communities of related genes. The method identified 682 genes, which were statistically relevant to colon cancer[17].

THE METHODS

Since we focus on discovering protein-to-disease associations, in particular, we will explain our method within this context. We firstly need disease dictionary, protein name dictionary and text dataset:

* The *Disease Dictionary* contains disease names and aliases obtained from *MeSH* database[18] (*http://www.nlm.nih.gov/cgi/request.meshdata).*
* The *Protein Name Dictionary* was created using protein names from three databases: *Swissprot, Tremble and LocusLink*[19-21]. For every protein (or gene), there are typically many synonyms, abbreviations and other symbols used in the literature and listed in these databases. We resolved this issue using a set of rules without losing significant protein/gene names information. We compiled the protein name dictionary as a list wherein all names and symbols of one protein are grouped as one entry. We used a number of rules for creating a protein dictionary. Here are some of these rules:
    * *Protein names having the same primary gene name were considered one protein (e.g. 'major prion protein precursor' and 'prion protein' are considered as same proteins because they have a common primary gene name).*
    * We excluded from the dictionary:
* *Protein names containing single character*[22].
* *Protein names having purely numerical entries*[22]
* *Protein names identical to gene names (e.g. "ZnF20" is an official gene name and also is an alias for "Zinc finger protein 197" protein).*
* *Protein names identical to common English words (e.g. "VAN" is a common English word and also an alias for "Nef-associated factor 1" protein)*
* *Entries consisting only of measures. (e.g. "23 kDa protein")*[22].

Other related research uses similar heuristic rules[22]. We also utilized gene name dictionaries to ensure that each protein name or symbol is mentioned as protein and not referring to a gene as there are cases in which an exact symbol refers to a protein and to a gene[3,14].

Text dataset: The main source of our texts is the *Medline*[0]. The Medline database was created by the US National Library of Medicine (NML)[23]. Medline is considered the main text database in the bioinformatics

domain, because of its free accessibility and huge coverage. Each citation is associated with a set of *MeSH* (Medical Subject Headings) terms[18] that describe the content of the item[23].

The main contributions of this work are first in how we discover the proteins that are related to the input topic of interest (*disease*) by combining positive and negative evidences. And then, more importantly, how we filter, from the discovered relations, those that are statistically significant from the insignificant ones. Our method relies on statistically reliable measures of difference between *expected* and *evidence* of protein and disease counts and co-occurrences in terms of *df* and *tf* (as explained next) between the positive set and the negative set of documents. The details of the method are as follow.

We want to discover, for a given disease name (topic of interest), all the proteins that are significantly associated with that disease. For the input topic of interest (which is a *disease name* in this case) we collect from *Medline* all the abstracts on that topic by querying *Pubmed*[1] using all disease names and abbreviations. The output of this step is the set of abstracts containing one or more instances of the disease. Let us call this set of (*relevant)* abstracts S_1. Thus,

$S_1 = \{A_1, A_2, \ldots, A_n\}$

and A_i is an abstract retrieved from Medline using the topic of interest (*disease name*) as keyword

* Next, we use the protein name dictionary to extract all the proteins mentioned in S_1. We call this set of proteins S_p. Then $S_p = \{P_1, P_2, \ldots, P_m\}$ where P_i is a protein name mentioned in at least one of the abstracts of the set S_1. *[Notice here that each abstracts in S_1 contains a mention of the disease; thus, each protein mentioned in any of these abstracts is considered initially as having a co-occurrence with the disease].*
* We also retrieve from *Medline* another "*control*" set of abstracts; we call it S_2. This set contains abstracts randomly chosen from *Medline* and do not contain any mention of the topic of interest *i.e. Negative Set.* This set (S_2) is used as a *control* set for collecting negative evidences to measure the statistical significance of the discovered relations. We chose the number of random documents ($|S_2|$) to be ~ *40K - 45K* documents (Table 1). We carefully selected this range after we have tried various options, like making the set S_2 double the size of S_1 and so on; and we found this produces the best results and acceptable computability. (In some cases, we repeated the experiment with multiple different random sets if needed). It's worthwhile mentioning at this point and before we delve into the details of the method, that we remove from the set S_1 any abstract talking explicitly about a significant relationship between the disease and any protein. We call such documents *verification documents* and we use them to verify our findings (Table 5).

Method I: Computing ex-ev using DF: At this point, we have the sets S_1 , S_2 and S_p :

* For each protein P_i in S_p (*i.e.,* P_i is mentioned in the abstracts of S_1), we compute *document frequency* (*df*) of P_i in both sets S_1 and S_2 as follows:

Document frequency 1 of protein P_i:
$df_1(P_i)$ = number of S_1 documents in which P_i is mentioned

Document frequency 2 of protein P_i:
$df_2(P_i)$ = number of S_2 documents in which P_i is mentioned

Total document frequency of protein P_i:
$df_t(P_i) = df_1(P_i) + df_2(P_i)$

* Next, we combine the positive and negative evidences of the co-occurrences and frequency counts from S_1 and S_2. For measuring statistical significance of the discovered relations, we want to know for each protein P_i mentioned in S_1 to what "*level of likelihood*" this co-occurrence implies that there is a significant relation between P_i and the underlying disease.

We compute for each protein in the set S_p an *expectation* (*ex*) value and an *evidence* (*ev*) value[3], as follows:

$$ex(P_i) = [df_t(P_i) / |S_1+S_2|] * |S_1| \quad (1)$$

$$ev(P_i) = df_1(P_i) \quad (2)$$

The *expectation* measures how many S_1 abstracts P_i is *normally* expected to appear in; whereas, the *evidence* determines how many S_1 abstracts P_i has actually appeared in. It is obvious now that the larger the difference between *ex* and *ev:* $ev(P_i) - ex(P_i)$ the more the likelihood that P_i and the disease have a significant association.

Table 1: Number of relevant and random documents extracted for each disease in the experiments

Experiment (disease)	Number of relevant docs: $\|S_1\|$	Number of random docs: $\|S_2\|$
Alzheimer	42,077	41,856
Creutzfeldt-Jakob	4,890	41,863
Huntington Disease	7,250	41,863
Crohn Disease	18,642	41,863
Jaundice	20,386	40,237
Dengue	4,780	41,863
Spinal Cord Injuries	16,839	40,292
Lung Cancer	43,933	41,862

Thus, this difference [$ev(P_i)$ - $ex(P_i)$] indicates in how many S_1 abstracts the protein is mentioned minus how many S_1 abstracts in which it is expected to appear.

We need to normalize this difference as the same value of $ev(P_i)$ - $ex(P_i)$ can have different significance in differently distributed proteins. For example, a difference of 10 for a protein that is mentioned in 150 abstracts has less significance than a difference of 10 for a protein mentioned in only 20 abstracts. Hence we normalize the difference by dividing by the $df_t(P_i)$ value of the protein. Then, we define a function *f*:

$$f(P_i) = \frac{ev(P_i) - ex(P_i)}{dft(P_i)} \quad (3)$$

We compute the *f* value for each protein in the set S_p according to (3). Then we sort the proteins according to their *f* values and we use the *Z-score* metric to determine the significant *f* values:

$$Z\text{-}Score(P_i) = [f(P_i) - mean(f)]/SD(f) \quad (4)$$

Where *mean(f)* is the mean of all *f* values of all proteins of S_p and *SD(f)* is the standard deviation of *f* values. Thus, the Z-score measures how many standard deviations each *f* value is greater than the mean *f* value, for all proteins, to indicate statistical significance. The Z-score technique has been used in text mining[13] and is considered a reliable measure of statistical significance.

Method II: Computing ex-ev using tf: So far, we have explained how we compute the significance of the discovered associations by utilizing document frequency (DF). Now, we describe how the significance is computed by utilizing the *term frequency* (TF) statistics for each protein

We then compute for each protein P_i in S_p how many times it occurred in each of S_1 and S_2 as follows*:*

Term frequency 1 of protein P_i:
$tf_1(P_i)$ = number of occurrences (mentions) of P_i in the set S_1

Term frequency 2 of protein P_i:
$tf_2(P_i)$ = number of occurrences (mentions) of P_i in the set S_2

Total term frequency of P_i:
tft(Pi) = tf1(Pi) + tf2(Pi)

Then, we carry out basically the same steps in method I except that we use *tf* instead of *df*. That is, we calculate the *ex* and *ev* values for each protein as follows:

$$ex(P_i) = [tf_t(P_i)/|S_1+S_2|]* |S_1| \quad (5)$$

$$ev(P_i) = tf_1(P_i) \quad (6)$$

And the *f* values are:

$$f(P_i) = [ev(P_i) - ex(P_i)]/tf_t(P_i) \quad (7)$$

Similarly we compute the *Z-score* for each protein P_i in the set S_p using equation (4). In our evaluations, we found high correlation (>90%) between the *Z-scores* computed using methods I and II. Hence the final estimate is by the combination of methods I and II. That is, we consider a protein as having significant association with the disease if it gets *Z*-scores of 1.0 or more in both methods I and II.

As we see, the df_1 and tf_1 values capture the co-occurrence counts of the diseases and proteins in the relevant set of documents (S_1) and hence considered the *positive* evidences, whereas df_2 and tf_2 are the *negative* evidences as they capture the occurrence counts of the proteins in the negative set of documents and counted against the association. There are number of methods in the literature utilizing the co-occurrence counts for discovering significant relations as terms that tend to co-occur more frequently are more likely to have biologically significant relationships[6,24].

RESULTS AND DISCUSSION

The method was evaluated with a number of experiments on various diseases to discover the proteins related to those diseases. We ran experiments on 8 different diseases: *Alzheimer, Creutzfeldt-Jakob, Crohn Disease, Huntington, Jaundice, Lung cancer, Dengue, Spinal Cord Injuries and* for brevity sake, we will discuss only *Alzheimer* and *Huntington* diseases in detail, while a summary of other experiments is included in Table 5.

Alzheimer experiment: We want to find the proteins that are associated with the Alzheimer disease using Medline texts. In this experiment, the method ran on a total of *83,933* Medline abstracts for Alzheimer experiment. First, we downloaded from Medline *42,077* abstracts for this disease (this is the set of relevant abstracts S_1). Then, we retrieved another set of randomly chosen abstracts that does not have any mention of Alzheimer. This is the set S_2 in our method and contains *41,856* abstracts. Then, we extracted from S_1, occurrences for 1163 distinct proteins. This set (1163 proteins) is the set S_p in our method. Of course, each one of these proteins is mentioned with its various abbreviations, synonyms and aliases and this issue was resolved using the disambiguation rules and the protein name dictionary that we created for this purpose. Table 2 contains a sample of 8 of these proteins (for space constrains, we listed only the first 8 proteins from S_p alphabetically). Then, for each protein we computed the f values and Z-scores according to methods I and II. We used a threshold of 1.0 to indicate significant associations as explained earlier. The results are shown in Table 3 using method I and method II: Out of the 1163 proteins associated with Alzheimer, we found only 6 having significant associations (Z-scores >= 1.0

Table 2: Part of the set S_p of all proteins mentioned in the set S_1 of *Alzheimer* abstracts. This part includes only the first 8 proteins (for space constraints). Each line contains the protein name along with its aliases, abbreviations and synonyms used in the abstracts

Alpha-mannosidase II, Alpha-mannosidase II, Golgi alpha-mannosidase II
Amyloid beta A4 protein precursor, Amyloid beta A4 protein, Amyloid protein, Alzheimer's disease amyloid protein, ABPP, PreA4
Apoptosis-inducing protein
Arachidonate 12-lipoxygenase, 12S-type, 12-LOX
Metabotropic glutamate receptor 2 precursor, Metabotropic glutamate receptor 2
Presenilin 1, Presenilin-1, PS 1, PS-1
Neuromodulin, Growth associated protein 43.
Nicastrin

Table 3: The 6 proteins that are associated with *Alzheimer* disease; these proteins have Z-scores >= 1.0 using method I and method II

Protein	Protein freq. in relevant docs (S_1)		Protein freq. in random docs (S_2)		Using *DF* statistics *(Method I)*					Using *TF* statistics *(Method II)*				
	tf_1	df_1	tf_2	df_2	*ex*	*ev*	*ev - ex*	*f()*	*Z-score*	*ex*	*ev*	*ev - ex*	*f()*	*Z-score*
PS-1	1456	842	9	9	426.62	842	415.38	0.49	1.27	734.43	1456	721.57	0.49	1.24
ABPP	472	328	11	11	169.95	328	158.05	0.47	1.22	242.14	472	229.86	0.48	1.21
B-S1P	121	42	3	3	22.56	42	19.44	0.43	1.11	62.16	121	58.84	0.47	1.19
M-AP	407	394	22	22	208.55	394	185.45	0.45	1.17	215.06	407	191.94	0.45	1.14
B-S2P	149	108	8	8	58.15	108	49.85	0.43	1.11	78.71	149	70.29	0.45	1.14
Apo-E	1984	1592	210	200	898.36	1592	693.64	0.39	1.01	1099.89	1984	884.11	0.4	1.01

Table 4: The Z-scores of the proteins that are associated with the *Huntington* disease using method I and method II

Protein	Protein freq. in relevant docs (S_1)		Protein freq. in random docs (S_2)		Using *DF* statistics *(Method I)*					Using *TF* statistics *(Method II)*				
	tf_1	df_1	tf_2	df_2	*ex*	*ev*	*ev - ex*	*f()*	*Z-score*	*ex*	*ev*	*ev - ex*	*f()*	*Z-score*
Huntingtin	1626	590	16	6	87.98	590	502.02	0.84	1.52	242.39	1626	1383.6	0.84	1.51
HIP-1	10	5	1	1	0.89	5	4.11	0.68	1.09	1.62	10	8.38	0.76	1.3
JP-3	19	10	3	1	1.62	10	8.38	0.76	1.3	3.25	19	15.75	0.72	1.2

in both methods I and II). This meant that the remaining 1157 proteins mentioned in S_1 are occurring sporadically (have insignificant associations), which was evidenced from the control set S_2 of random abstracts.

Huntington experiment: In this experiment, the set S_1 consisted of *7,250* abstracts (disease relevant documents) whereas the random set S_2 contained *41,863* abstracts. The set S_p consisted of 403 proteins. We found that out of the 403 proteins mentioned in *Huntington* documents only 3 proteins are significantly associated with *Huntington* disease. The results are in Table 4 using method I and method II.

The protein-disease associations discovered by our method were verified manually, from literature, to see whether these results were published. In *Alzheimer* test, we conducted our experiments on *83,933* abstracts and discovered that *Alzheimer* is associated (with statistical significance) with 6 proteins, among them: *Amyloid beta A4, Presenilin 1, Apolipoprotein E precursor* (more in Table 5). We investigated and researched these results carefully and found that these proteins are actually related with *Alzheimer* according to a number of biomedical papers and for space constraints, we only list the PubMed Ids of these articles. [PMIDs: 8596911, 1465129, 8346443, 12614323, 8766720 and 8878479] Also more details are in Table 5. In the *Huntington* test we verified the discovered associations and found proofs in the following documents: [PMIDs: 10823891, 15064418, 14962977 and 11832235]. Moreover, we found verification articles for all the remaining associations. This implies that the precision of our method is very impressive as all the discovered protein-disease associations were confirmed manually from literature. Recall that the verification documents are not included in the documents mined.

Precision and recall: The *Precision* (P) and *Recall* (R) are two reliable metrics used to measure the performance of such methods like the one presented here[6]. For a given concept of interest (*i.e. disease)* the method produces a number of proteins as associated with that disease. One way to evaluate this is to determine how many of these output proteins are *correctly* and *actually* related to the disease (*precision*) and how many of those proteins actually related to that disease, has our system discovered (*recall*).
That is:

$$P = \frac{\text{number of correct proteins found by the system}}{\text{total number of proteins found by the system}}$$

$$R = \frac{\text{number of correct proteins found by the system}}{\text{total number of proteins actually related to the disease}}$$

The *recall* here cannot be computed since there is no such complete data about *all* proteins associated with a disease. However, we tried to find a simple way to roughly estimate the precision and recall rates of our method. We retrieved three sets of 25 abstracts each related to three different diseases. In each set, we manually extracted all protein mentions and then

Table 5: Summary of results for diseases-protein associations with verification articles

Disease	Associated Proteins	Verification Documents
Alzheimer	Presenilin 1(PS1), Amyloid beta A4 protein precursor, Beta-secretase 1 precursor(Memapsin 1), Microtubule-associated protein, Beta secretase 2 precursor, Apolipoprotein E precursor	Type 3 Alzheimer disease may be the result of impaired proteolytic processing of PS1.(PMID: 8766720). The progressive deposition in the human brain of amyloid filaments composed of the amyloid beta protein is a principal feature of Alzheimer's disease. (PMID: 2960019). Others: [8878479, 11311782, 8596911, 12423367, 1465129, 7891887, 8673924, 8346443, 10671320 , 10501182, 9626772, 12614323]
Huntington	Huntingtin, Huntingtin interacting protein 1, Junctophilin 3	Huntington's disease (HD) is an autosomal dominant condition, resulting from a mutation in huntingtin.(PMID: 11765125). Others: [10823891, 15064418, 14962977, 11832235, 15468075, 14557581, 15876586]
Creutzfeldt-Jakob	Major prion protein precursor, Peroxiredoxin 6(1-Cys Prx), Gamma enolase(Neuron-specific enolase)	Abnormal accumulations of prion protein (PrP) can be detected in the spleen, lymph nodes and tonsils of patients with variant Creutzfeldt-Jakob disease(11476840). Others: [3917302, 8035877, 10987652, 12970341, 12210213, 3309455, 10081943, 2663293].
Crohn Disease	Thiopurine S-methyltransferase, Trefoil factor 2 precursor, Trefoil factor 3 precursor(hP1.B)	Trefoil peptides are widely distributed in the intestine in human inflammatory bowel disease and are of considerable potential functional importance(8283019). Others: [10833476, 8419234, 8368306, 8346203]
Jaundice	Plasma kallikrein precursor, UDP-glucuronosyltransferase 1-3 precursor, Secretory component, Glucuronosyltransferase, Canalicular multispecific organic anion transporter 1, Beta-2-microglobulin precursor	Tubular dysfunction, manifested by increased urinary excretion of B2MG(Beta-2-microglobulin) occurs in patients with hepatorenal syndrome and deep jaundice(3884478). Others: [1084679, 9738861, 12502904, 7439618, 6198239]
Dengue	Heterogeneous nuclear ribonucleoprotein K, HERV-E envelope glycoprotein(Envelope protein), Envelope glycoprotein, CD209 antigen-like protein 1, RNA helicase, CD209 antigen	The heterogeneous nuclear ribonucleoprotein K (hnRNP K) interacts with dengue virus core protein(11747608). Others: [15579065, 1339466, 1342710, 9256277, 10964773]
Spinal Cord Injuries	Semaphorin 3A precursor(semaphorin 3A), Reticulon 4 receptor precursor(nogo-66 receptor), Heparin-binding growth factor 1 precursor(Acidic fibroblast growth factor), Tenascin-R precursor, Adenosine A1 receptor, Brevican core protein precursor, Microtubule-associated protein 2, Bone morphogenetic protein 7 precursor(Osteogenic protein-1)	The recent discovery of the Nogo family of myelin inhibitors and the Nogo-66 receptor opens up a very promising avenue for the development of therapeutic agents for treating spinal cord injury(15317586). Others: [14727128, 12764110, 15247588, 9418975, 8594213, 15525355, 12895450, 9326288, 15573078, 10338277]
Lung Cancer	Mucin short variant SV10 (episialin) Methylthioadenosine phosphorylase	Data suggest that MeSAdo phosphorylase deficiency is frequently found in non-small cell lung cancers(8382555). Others: [9677444, 8971171]

carefully reviewed the abstracts to infer and induce the proteins that are actually/correctly related to the disease as can be inferred by a careful reader who is looking for proteins-disease relationships particularly. Then we ran our system on each one of these sets separately to compare the system's results against our manual finding. In the first case, the system produced a total of 17 proteins and correctly identified 16 proteins out of 18 proteins that we manually found. While in the second case, the system recalled correctly 13 out of 15 proteins related to the disease and manually the proteins found were 24. And in the third case, the system recalled correctly 22 out of 24 proteins related to the disease and manually the proteins found were 35. These results are as follows:

Set 1: P = 16/17 = 0.94, R = 16/18 = 0.89
Set 2: P = 13/15 = 0.87, R = 13/24 = 0.54
Set 3: P = 22/24 = 0.92, R = 24/35 = 0.69

On average, our method achieved a precision rate of 0.91 and a 0.71 recall rate.

Supporting known relationships: To further evaluate our method, we tested the method on some already known and published relationships between genes and diseases; and conducted three such tests as follows.

* We ran the first experiment on an already published association[4] which states that "RUNX1" gene has a strong connection with "acute myeloid leukemia". Our method correctly identified this association with Z-score values >= 1.
* The method described[25], predicted the involvement of "synapsin I" in "long-term potentiation (LTP)" which had been demonstrated[26] and also with "calcium calmodulin kinase type II" which had been established[27]. Our method successfully extracted relevance between

"LTP" and "synapsin I" with Z-Scores >= 1 and between "LTP" and "calcium calmodulin kinase type II" also with Z-scores >= 1.

* Finally we ran an experiment on a published association[28] between "Parkinson's disease" and various genes and our method extracted (with high Z-Score values), the relevance of genes "PARK1", "PARK2" and "PARK7" with "Parkinson's disease".

CONCLUSION

We presented a new approach for identifying significant associations between diseases and proteins. Finding such protein to disease relationships is not an easy process and not much research has been done on this task. The novelty in this approach is two fold; first in discovering important associations, it depends not only on relevant documents on the topic of interest but also on another set of negative (*randomly chosen)* documents. The latter set is used as a *control* set to help in determining the statistical significance of the discovered associations. Second is that it depends on a new way of measuring the significance of an association between two biological terms.

In the future endeavor of this research we want to apply the method in discovering more relations between various biological entities like gene-to-disease associations and gene-to-drugs relations. We also would like to investigate applying weights to different types of term co-occurrences, for example, co-occurrence within the title, within certain window size, or within the abstract. Furthermore, we plan in the continuation of this research to investigate new methods to determine the type of the protein-disease associations.

REFERENCES

1. Medline: accessed using Entrez PubMed Interface: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
2. Hirschman, L., J.C. Park, J. Tsujii, L. Wong and C.H. Wu. 2002. Accomplishments and challenges in literature data mining for biology. Bioinformatics, Vol. 18.
3. Ginter, F., J. Boberg, J. J¨arvinen and T. Salakoski, 2004. New Techniques for Disambiguation in Natural Language and Their Application to Biological Text. JMLR, 5.
4. Adamic, L.A., D. Wilkinson, B.A. Huberman and E. Adar, 2002. A literature based method for identifying gene-disease connections. IEEE Computer Soc. Bioinformatics Conf.
5. Srinivasan, P., 2004. Text mining: generating hypotheses from MEDLINE. J. Am. Soc. Information Sci. and Technol., 55: 396-413.
6. Wren, J.D., R. Bekeredjian, J.A. Stewart, R.V. Shohet and H.R. Garner, 2004. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics, 20: 3.
7. Uramoto, N., H. Matsuzawa, T. Nagano, A. Murami and H. Takeuchi, 2004. A text-mining system for knowledge discovery from biomedical documents.
8. Hristovski, D., J. Stare, B. Peterlin and S. Dzeroski, 2001. Supporting discovery in medicine by association rule mining in Medline and UMLS. Proc. MedInfo Conf., London, England, Sep. 2-5, 10: 1344-1348.
9. Creighton, C. and S. Hanash, 2003. Mining gene expression databases for association rules. Bioinformatics, 19-1: 79-86.
10. Palakal, M., M. Stephens, S. Mukhopadhyay, R. Raje and S. Rhodes, 2002. A Multi-level Text Mining Method to Extract Biological Relationships. Proc. IEEE Computer Soc. Bioinformatics (CSB) Conf., pp: 97-108.
11. Weeber, M. *et al.*, 2003. Generating hypothesis by discovering implicit Associations in the literature: a case report of a search for new potential therapeutic uses of thalidomide. J. Am. Med. Inform. Assoc., 10: 252-259.
12. Srinivasan, P. and B. Libbus, 2004. Mining MEDLINE for Implicit Links between Dietary Substances and Diseases. ISMB 2004 and in Bioinformatics (Supplement).
13. Andrade, M. and A. Valencia, 1998. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. Bioinformatics, Vol.14.
14. Hatzivassiloglou, V., P.A. Dubou´e and A. Rzhetsky, 2001. Disambiguating proteins, genes and RNA in text: A machine learning approach. Bioinformatics, vol. 17.
15. Swanson, D.R., 1986. Fish oil, Raynaud's syndrome and undiscovered public knowledge. Perspectives in Biol. and Med., 30.
16. Smalheiser, N.R. and D.R. Swanson, 1998. Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. Comp. Meth. Prog. Biomed, Vol. 57.
17. Wilkinson, D.M. and B.A. Huberman, 2004. A method for finding communities of related genes. Proc. Natl. Acad. Sci. U.S.A., 101 Suppl. 1: 5241-5248.
18. Lowe, H.J. and G.O. Barnett, 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. JAMA, 271.
19. Swissport home: http://www.ebi.ac.uk/swissprot/
20. Tremble: http://www.ebi.ac.uk/trembl/
21. LocusLink at NCBI:http://www.ncbi.nlm.nih.gov/LocusLink/

22. Sergei Egorov, Anton Yuryev and Nikolai Daraselia, 2004. A simple and Practical Dictionary-bases Approach for Identification of proteins in Medline Abstracts. J. Am. Med. Informatics Assoc., 11: 174-178.
23. U.S. National Library of Medicine. http://www.nlm.nih.gov
24. Jenssen, T.K., A. Laegreid, J. Komorowski and E. Hovig, 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet., Vol. 28.
25. Hao Chen and Burt M. Sharp. Content-rich biological network constructed by mining PubMed abstracts
26. Fukunaga, K., D. Muller and E. Miyamoto, 1995. Increased phosphorylation of Ca2+/calmodulin-dependent protein kinase II and its endogenous substrates in the induction of long-term potentiation. J. Biol. Chem., 270: 6119-24. doi: 10.1074/jbc.270.11.6119.
27. Malinow, R., H. Schulman and R.W. Tsien, 1989. Inhibition of postsynaptic PKC or CaMKII blocks induction but not expression of LTP. Science, 245: 862-866.
28. Pankratz, N. and T. Foroud, 2004. Genetics of Parkinson Disease. NeuroRx, April 1, 1: 235-242.