DETECTING GENE RELATIONS FROM MEDLINE ABSTRACTS

M. STEPHENS, M. PALAKAL, S. MUKHOPADHYAY, R. RAJE
Department of Computer & Information Science
Indiana University Purdue University Indianapolis
Indianapolis, Indiana 46202, USA
mpalakal@cs.iupui.edu

J. MOSTAFA
Information Science & Informatics
Indiana University
Bloomington, Indiana 47405, USA
jm@cs.indiana.edu

Research in bioinformatics in the past decade has generated a large volume of textual biological data stored in databases such as MEDLINE. It takes a copious amount of effort and time, even for expert users, to manually extract useful information embedded in such a large volume of retrieved data and automated intelligent text analysis tools are increasingly becoming essential. In this article, we present a simple analysis and knowledge discovery method that can identify related genes as well as their shared functionality (if any) based on a collection of relevant retrieved relevant MEDLINE documents. The relative computational simplicity of the proposed method makes it possible to process and analyze large volumes of data in a short time. Hence, it significantly contributes to and enhances a user's ability to discover such embedded information. Two case studies are presented that indicate the usefulness of the proposed method.

1. Introduction

In bioinformatics there is a need to extract biological information from a plethora of literature. For example, how proteins relate to each other is important information that is used in the development of molecular pathways. Several techniques have been implemented in recent years to try and accomplish this task. BioNLP uses natural language processing (NLP) and pattern matching and has been shown to be an effective tool when combined with BioKleiski, a search engine, and BioJake, an interface for building metabolic pathways[1]. Another technique used to find protein-protein interaction involves NLP to process and tag all parts of text. The technique then infers gene interactions based on common verbs used to describe these interactions[2]. This implementation is also being extended to produce SGML documents so information extraction can be done for a multitude of tasks[3]. A different method uses finite-state lexical tools and a Hidden Markov Model (HMM) for speech tagging, checking against a collection of finite-state error recovery modules[4]. In some cases the NLP techniques have had accuracy of over 90%.

Other techniques are based more on statistics than NLP and extract information from keywords in text. One such technique uses a group of related documents against a set of random documents to extract domain specific information. This information could include gene function and interactions[5]. Some systems have extended this idea by first finding the most frequently seen keywords and checking them against a dictionary of genes using patterns created by surface clues to come up with the relationship[6].

Many of these techniques are computationally intensive and their applications to on-line analysis of a large set of retrieved documents will require significant waiting time on the part of users. Furthermore, very few of them attempt to assist in the text analysis with domain knowledge, available from authoritative organizations, experts, or users themselves. In this paper we present a Thesaurus-based text analysis approach and tool to discover the existence and the functional nature of relationships between genes relating to a problem domain of interest. The approach relies on multiple Thesauri, representing domain knowledge as gene names and terms describing gene functions. These Thesauri can be constructed using existing organizational sources (e.g., NCBI and EBI), by consulting experts in the domain of interest, or by the users themselves. Thesauri can also be constructed using automated vocabulary discovery techniques being developed by the Information Extraction (IE) or Information Retrieval (IR) communities. In its simplest form, a Thesaurus consists of a linear list of terms and associated concepts. However, the proposed association and function discovery techniques can extend to more complex Thesaurus structures incorporating synonyms or hierarchical relationships. Once Thesauri describing gene names and gene functions are established, the user initiates a retrieval process from MEDLINE and the analysis algorithms are applied on the large retrieved set to identify relationships and functions of genes. The analysis involves Thesaurus-based content representation of the retrieved documents, identification of associations (relationships) and finally detecting gene functionality from the represented retrieved document set. These primary steps are described in detail in the following sections along with some experimental results. The experiments, conducted using two different Thesauri, show the methods to be highly successful in identifying related gene pairs and moderately successful in identifying the functional nature of such relationships.

2. Text Document Representation

The document representation step converts text documents into structures that can be efficiently processed without the loss of vital content. At the core of this process is a thesaurus, an array T of atomic tokens (e.g., a single term) each identified by a unique numeric identifier culled from authoritative sources or automatically

discovered. A thesaurus is an extremely valuable component in term-normalization tasks and for replacing an uncontrolled vocabulary set with a controlled set[7]. Beyond the use of the thesaurus, the *tf.idf* (the term frequency multiplied with inverse document frequency) algorithm[7] is applied as an additional measure for achieving more accurate and refined discrimination at the term representation level. In this formula, the *idf* component acts as a weighting factor by taking into account inter-document term distribution, over the complete collection given by:

$$W_{ik} = T_{ik} \times \log(N / n_k) \qquad 2\text{-}1$$

Where T_{ik} is the number of occurrences of term T_k in document i, $I_k = \log(N/n_k)$ is the inverse document frequency of term T_k in the document base, N is the total number of documents in the document base, and n_k is the number of documents in the base that contain the given term T_k.

As document representation is conducted on a continuous stream, the number of documents present in the stream may be too few for the *idf* component to be usefully applied. To deal with this, a table is maintained containing total frequencies of all thesaurus terms in a sufficiently representative collection of documents as a base (randomly sampled documents from the source used as the training set). It is worth pointing out that such a table can be pre-constructed off-line before any on-line analysis of retrieved documents is attempted. The purpose of the document representation step is to convert each document to a weight vector whose dimension is the same as the number of terms in the thesaurus and whose elements are given by equation (2-1).

3. Gene-pair Relationship

In this section, we describe how the document vectors can be used to identify Gene-pair relationships. The goal is to discover pairs of genes from a collection of retrieved text documents such that the genes in each pair are related to one other in some manner. This is similar, in spirit, to the problem of association rule discovery, extensively studied in the database mining literature. However, there are differences between gene-association discovery and association rule discovery in databases:

(i) Association rule discovery is frequently based on transaction records, stored in specific formats; whereas the gene relationships are discovered from natural language text.

(ii) Commonly, database association rule discoveries are based on frequencies of individual items as well as the joint frequencies of pairs. In the context of a text document, these parameters are insufficient.

The relative "importance" of each gene, as well as the strength of their joint occurrences, play important roles. The vector space model attempts to compute the importance of terms (or, combinations) on the basis of term frequencies within a document and within an entire document collection. This influence of the document collection on the relative importance of terms is a distinguishing feature from just frequency based association rule discovery.

It is clear that whether two genes are to be automatically discovered to be related depends on the somewhat subjective notion of "being related". We have investigated Gene pair discovery from a collection of MEDLINE abstracts using the Vector-Space *tf***idf* method and a thesaurus consisting of Gene terms. Each Gene term, in turn, contains several synonymous keywords that are gene names. Each document d_i is converted to a *M* dimensional vector W_i where $W_i[k]$ denotes the weight of the k^{th} gene term in the document and M indicates the number of terms in a Thesarus. $W_i[k]$ is computed by the equation 3-1, which is a reformulation of equation (2-1) described earlier:

$$W_i[k] = T_i[k] * \log(N / n[k]) \qquad 3\text{-}1$$

Where $T_i[k]$ is the frequency of the k^{th} gene term in document d_i, *N* is the total number of documents in the collection, and *n*[k] is the number of documents out of *N* that contain the k^{th} gene term.

It is clear that $W_i[k]$ increases with term frequency $T_i[k]$. However, it decreases with *n*[k], i.e., if a gene term occurs in increasingly larger number of documents in the collection, it is treated as a common term and its weight is decreased.

Once the vector representation of all documents are computed, the association between two gene terms *k* and *l* is computed as follows:

$$association[k][l] = \sum_{i=1}^{N} W_i[k] * W_i[l] \quad \text{k = 1...m, } l = 1...\text{m} \qquad 3\text{-}2$$

For any pair of gene terms co-occurring in even a single document, the *association*[k][l] will be non-zero and positive. However, the relative values of *association*[k][l] will indicate the product of the importance of the k^{th} and l^{th} term in each document, summed over all documents. This computed association value is used as a measure of the degree of relationship between the k^{th} and l^{th} gene terms. A decision can be made about the existence of a strong relationship between genes using a user-defined threshold on the elements of the Association matrix.

4. Functional Nature of Relationships Between Gene-Pairs

Once a "relationship" has been found between genes, the next step is to find out what that relationship is. The approach taken here requires an additional thesaurus containing terms relating to possible relationships between genes that a user may be interested in. This thesaurus is then applied to sentences, which contain co-occurring gene names. If a word in the sentence containing co-occurrences of genes, matches a relationship in the thesaurus, it is counted as a score of one. The highest score over all sentences for a given relationship is then taken to be the relationship between the two genes or proteins. A score of as little as one could be significant because a relationship may be only mentioned in one abstract. A higher score, however, would be more likely to indicate that relationship because they are often reiterated in multiple abstracts. The following is an equation summarizing the relationship:

$$score[k][l][m] = \sum_{i=1}^{N} p_i \,;\, (p_i = 1 : Gene_k, Gene_l, \mathrm{Re}\,lation_m \text{ all occur in sentence } i) \qquad 4\text{-}1$$

where, N is the number of sentences in the retrieved document collection, p_i is a score equal to 1 or 0 depending on whether or not all terms are present, and $Gene_k$ refers to the gene in the gene thesaurus with index *k*, and $relation_m$ refers to the term in the relationship thesaurus with index *m*. The functional nature of the relationship is chosen as arg_m score[k][l][m]. It is worth pointing out that the function array is computed only for the gene pairs where relations are found using the criteria described in section 3.

The idea is to narrow down the search to a few relationships which the user can check. If a functional relationship cannot be found the user can still check against articles where the terms co-occurred to see if a function might have been missing from the function thesaurus containing the relationships. Overall, this will help the user to quickly develop potential pathways and speed up the process of finding genetic interactions.

5. Experimental Results

Two experiments show how this technique performs in accuracy and as a tool for discovering a legitimate pathway based on retrieved data. The list of potential relations used for both examples, determined manually using a Molecular Biology text book[9], is shown in figure 5-1. The first experiment uses the list shown in figure 5-2.

"activates, activator"	"inhibits, inhibitor"	"phosphorylates"
"binds, binding, complexes"	"catalyst, catalyses"	"hydrolysis, hydrolyzes"
"cleaves"	"adhesion"	"donates"
"regulates"	"induces"	"creates"
"becomes"	"transports"	"exports"
"releases"	"suppresses, suppressors"	

Figure 5-1: The Thesaurus of Relationships

This list includes genes and proteins not taken from any particular pathway but is associated with cell structure and muscle cells. The other example includes a known list of genes and relationships for the pathway of ER transport to the Golgi[8],

"actinin"	"actn2"	"ank1, ankyrin"	"atf4"
"ca3"	"CD36"	"cd54"	"COI"
"cox1"	"CSE1"	"cst3"	"desmin"
"FKBP51"	"FKBP54"	"FUS, TLS"	"GAPDH"
"hmsh2"	"hrv"	"hsp90"	"importin"
"lim"	"mcm4"	"myoglobin"	"nebulin"
"nfatc"	"myosin"	"nop-30"	"NPI-1"
"p55"	"titin"	"ubiquinone"	"filamin"

Figure 5-2: Thesaurus of Genes (Unknown Pathway)

"SEC13"	"SEC12"	"SEC16"	"SAR1"
"SEC23"	"YPT1"	"Rab1"	"SEC21"
"BET1"	"ARF1"	"SEC7"	"SLY1"

Figure 5-3 : Thesaurus of Genes (Known Pathway)

is shown in figure 5-3. Some protein names represent more than one protein or gene. For example, myosin would include all instances of myosin I and myosin II. The training documents are created by taking an equal number of abstracts from the MEDLINE database for each gene. For the known pathway 1835 abstracts were used and for the unknown pathway 5072 abstracts were used. The analysis was then performed on each set of documents.

The results are shown in table 5-1 for the unknown pathway and table 5-2 for the known pathway. In these tables, the **Gene-Pair** shows the pairs of genes found to have **Association Strength** calculated using equation 3-2 that has a value greater

than zero. Higher association strength represents a stronger relationship between the genes. The **Predicted Relationship**, described in section 4, and the **Relation Score** , calculated using equation 4-1, is also shown in the table.

Table 5-1: Results for Identified Gene Pairs and Functional Relationship: Unknown Pathway

Gene-Pair	**Association Strength**	**Predicted Relationship * = correct**	**Relation Score**	**Known Relationship**
actinin-ank1	40.62	Unidentified	0	Both are actin binding proteins
actinin-desmin	1771.35	Unidentified	0	Make up cell cytoskeleton
actinin-filamin	738.42	binds *****	2	Both are actin binding proteins
actinin-myosin	2270.31	binds *	1	Actinin binds myosin
actinin-titin	5240.56	binds *	7	Actinin binds titin
actinin-nebulin	2481.77	binds*	1	Both are sarcomeric constituents
ank1-desmin	228.39	Unidentified	0	Both cytoskeletal proteins
ank1-myosin	606.31	binds *	3	Ankrin is involved in binding phosphorylated myosin
cse1-importin	2209.70	Unidentified	0	Cse1 recycles importin back to the cytoplasm
desmin-filamin	160.43	Unidentified	0	Both cytoskeletal proteins
desmin-TLS	25.99	Unidentified	0	Both immunohistachemical and molecular markers
desmin-nebulin	3857.15	Unidentified	0	Both cytoskeletal proteins
desmin-myosin	1819.45	Adhesion	2	Both cytoskeletal proteins
desmin-titin	4060.39	Unidentified	0	Both cytoskeletal proteins
filamin-nebulin	201.76	Unidentified	0	Both cytoskeletal proteins
filamin-titin	125.69	Unidentified	0	Both cytoskeletal proteins
fkbp54-hsp90	104.07	Unidentified	0	Both part of an Avian progesterone receptor complex
nebulin-myosin	2991.14	binds *	4	Nebulin binds to myosin
nebulin-titin	10079.16	binds	1	Both involved in the alignment of thick and thin filaments
myosin-titin	3689.34	binds	9	MyBP-C binds to both proteins

If all potential relationships have a relation score of zero, then the predicted relationship is marked as "Undefined". This shows that there were no sentences involving the gene pair along with a term in the relationship thesaurus. The **Known Relationship** were done by hand using MEDLINE abstracts along with [8, 9]. Due to limited space, only one known relationship for each gene-pair is shown and the thesaurus for the known pathway uses only a partial list of genes

Table 5-2: Results for Identified Gene Pairs and Functional Relationship: Known Pathway

Gene-Pair	Association Strength	Relationship * = correct	Relation Score	Known Relationship
sec13-sec16	97.03	Unidentified	0	Sec13 exhibits synthetic lethality in combination with sec16
sec13-sec23	458.92	Unidentified	0	Sec13 exhibits synthetic lethality in combination with sec23
sec13-sec21	157.47	Unidentified	0	Both involved in vesicle transport
sec12-sec16	274.05	suppressors	4	When both genes are mutated the cell does not grow.
sec12-sar1	72.64	Unidentified	0	Elevated expression of sar1 suppresses sec12
sec23-ypt1	13.45	Unidentified	0	Biochemical interaction
sec23-sec21	126.04	Unidentified	0	Both involved in vesicle transport
ypt1-rab1	90.23	Unidentified	0	Are members of the Ras super family of small GTP binding proteins
ypt1-bet1	248.53	Unidentified	0	Overproduction of bet1 suppresses the phenotype of mutant ypt1
ypt1-sec7	42.59	Unidentified	0	Both involved in vesicle transport
ypt1-sly1	265.08	Unidentified	0	Genetic interaction
sec21-bet1	114.15	Unidentified	0	Genetic interaction
sec21-arf1	659.96	Binds*	1	Genetic interaction
sec21-bos1	113.02	Unidentified	0	Both involved in vesicle transport
bet1-arf1	100.49	Unidentified	0	Genetic interaction
bet1-sly1	198.54	Unidentified	0	Both involved in vesicle transport

involved in the ER transport to Golgi pathway. The inverse relationship to this is shown in figures 5-4 and 5-5 as a graph where genes with greater association strength are closer together. A discussion will follow regarding the results.

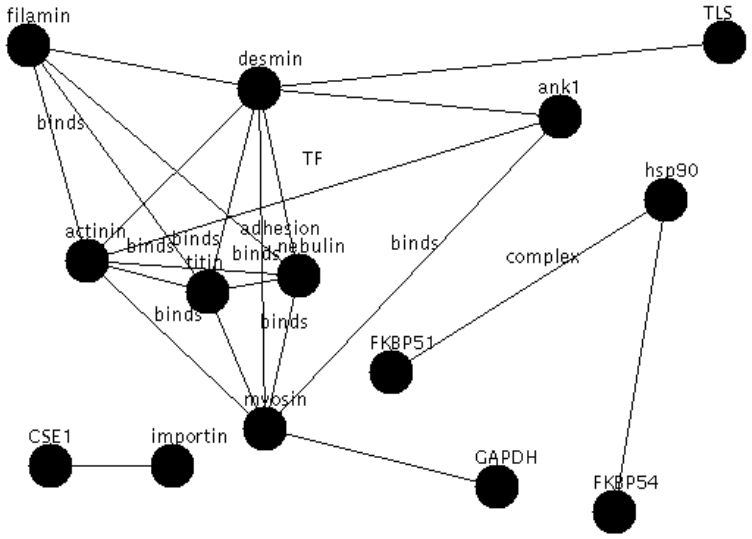

Figure 5- 4: Graph showing relationships between genes in Unknown Pathway(s). The higher the Association strength the closer the genes appear on the graph. In this way the related genes are clustered together and can be picked out.

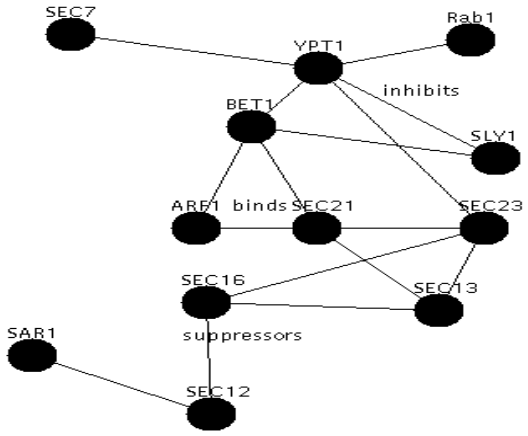

Figure 5- 5: Graph showing relationships between genes in Known Pathway. The higher the Association strength the closer the genes appear on the graph. In this way the related genes are clustered together and can be picked out.

6. Discussions of Experimental Results

The relationship discovery aspect of the proposed method was excellent. Almost all gene pairs had a relationship of some kind. This was verified by looking at the actual abstracts on the basis of which associations were computed. Using [8], the sensitivity of the known pathway was calculated at 61% and the specificity at 89%. The relationships missed are due to the absence of abstracts describing the relationship. Using [9] and the abstracts the sensitivity for the unknown pathway was 100%. The sensitivity could be lower depending on the source. The specificity of the unknown pathway depends strongly on how one interprets the data. The strong central cluster includes proteins involved in construction of the cytoskeleton. Weakly associated TLS and GAPDH do not belong putting the specificity around 78%. The cluster containing CSE1 and importin are involved in the process of recycling importin and the other cluster contains proteins involved in making a steroid receptor complex both having specificity of 100%.

Finding the actual nature of the relationship between proteins had a specificity of 67% in the unknown pathway and specificity of 50% in the known pathway. Several problems with this technique contributed to the low specificity. First of all, it was seen that very few abstracts actually contained both gene pairs and the functional relationships in the same sentence. In the case of the unknown pathway, there were only 49 sentences out of 5072 abstracts containing both a gene pair and a function. Either having more abstracts or the complete articles could correct this. All relationships mentioned in the articles were found for the gene pairs if the relationship was mentioned in the thesaurus.

Furthermore, the relationship between the two genes may not be direct. For example: "Utropin is a large multidomain protein that belongs to a superfamily of actin **binding** proteins, which includes dystrophin, alpha-**actinin**, beta-spectrin, fimbrim, **filamin**, and plectin". The bold words are the function and two genes in one sentence. One can see that actinin does not bind filamin but the relationship is portrayed because they share a common function. It is therefore necessary to find a way to separate relationships that are based on association and relationships based on actual interaction between proteins.

Lastly, the relationship between two co-ocurring genes may not exist at all. This was in very few cases but nevertheless a problem. For example: "The maturation of the sarcomeres was characterized by a short delay in the establishment

of the pattern for M-line epitopes of **titin** with respect to Z-disk epitopes and the incorporation of the M-lin component myosin, which proceeded that of **myosin binding** protein-C". The relationship found here is that titin binds myosin. The correct relationship here is that protein-C binds myosin. One way to correct this is to just have the user look at the associated article and delete the wrong associations. This also gives the users an advantage because they may pick up a relationship they did not consider before. So for the purpose of knowledge discovery this problem could actually be an advantage.

Perhaps a better way to approach the problem of matching functions to proteins is a distance relationship where one no longer looks at the sentences but rather the whole document keeping track of an index for each keyword. Once all keywords are indexed, a relationship could be found taking the average index of two genes, finding the closest relationship term associated with the two genes, and scoring the relationship based on distance. In this way, functions with strong associations between genes will score higher and can stand out among genes that score higher with other gene pairs. This would enable a user to select the desired specificity while allowing the flexibility to find new relationships.

The ability to cluster proteins together into potential pathways shows that the gene pairing technique is a potentially powerful tool when the genes are known. What happens when the genes are not known in advance? For this, additional functionality would need to be added as it is impractical to use all known genes and functions. A better way of doing this may involve first automatically finding all genes and functions mentioned in a set of training articles. From there, the two thesauri could then be used to find relationships. This would allow the algorithm to be effective for any set of articles as well as a known set of gene names.

7. Future Directions and Enhancements:

The Association and functional relationship discovery algorithm described in this paper are based on the information contained in the retrieved documents from MEDLINE. As pointed out in section 6, the abstracts extracted from MEDLINE in some cases lacked specific information concerning gene functions. One way to remedy this problem is to access and analyze full documents, rather than only abstracts. The retrieved collection can also be augmented by accessing other text-based collections, such as the On-Line Mendelian Inheritance in Man (OMIM) collection, in addition to MEDLINE. The sequence databases (e.g., GeneBank) also contain functional information about genes, which can be utilized. Finally, predicting functions of genes from sequences using computational models (e.g. Hidden Markov Models or Neural Networks) is an important and on-going international effort. Accessing sequence databases with the gene names to retrieve

their sequence data and then applying the prediction models on that sequence data would help leverage finding new relationships by accurate computational models. The intersection of the sets of functions (either accessed from sequence databases or predicted form sequences) of associated genes can be a useful pointer for identifying the nature of the relationship.

Another future direction of great usefulness is to integrate association discovery tools with profile-based information filtering (IF) engines. Such biological IF systems retrieve documents on the basis of stable long-term automatically learned profiles of user interests, rather than specific user queries. Such an integrated filtering and analysis system will help the user to keep up-to-date with evolving document and information collections.

Acknowledgement

This project is supported in part by a grant from the NSF Digital Libraries Initiative Phase 2 grant IIS-9817572 and Elli Lilly & Company.

References

1. N. See-Kiong and M.Wong, "Toward Routine Automatic Pathway Discovery from On-line Scientific Text Abstracts" *Genome Informatics* 10: 104-112 (1999).
2. T. Sekimizu, H. Park, and J. Tsujii, "Identifying the interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts" *Genome Informatics* 9: 62-71 (1998).
3. D. Proux, F. Rechenmann, L. Julliard, V. Pillet, and B. Jacq, "Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction" *Genome Informatics* 9: 72-80 (1998).
4. T. Hishiki, N. Collier, C. Nobata, T. Okazaki-Ohta, N. Ogata, H. Park, C. Nobata, T. Sekimizu, and J. Tsujii, "Developing NLP Tools for Genome Informatics: An Information Extraction Perspective" *Genome Informatics* 9: 81-90 (1998).
5. M. Andrade and A. Valencia, "Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families" *Bioinformatics*, 14:600-6007 (1998)
6. O. no, A. anigami, A. is iga i, O. a agi, "Automatic extraction of Information on Protein-Protein Interaction from Scientific Literature" *Genome Informatics* 10: 296-297 (1999)
7. G. Salton, Automatic Text Processing. Addison-Wesley (1989)
8. Rothblatt J., Novick P., Stevens T. Guidebook to the Secretory Pathway. Oxford University Press Inc., New York (1994)

9. Lodish H., Berk A., Matsudaira P., Baltimore D., Zipursky S., Darnell J. Molecular Cell Biology. Third Edition. Scientific Books, Inc. New York (1995)
10. Wilbur WJ, Yang Y. "An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts" Comput Biol Med. 1996 May; 26(3):209-22.