Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Andrew Hoblitzell

Entitled

Biomedical Literature Mining with Transitive Closure and Maximum Network Flow

For the degree of Master of Science

Is approved by the final examining committee:

Snehasis Mukhopadhyay	05/07/2010
Chair	
Yuni Xia	05/07/2010
Shiafoen Fang	05/07/2010

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Snehasis Mukhopadhyay

Approved by: Shiaofen Fang 05/07/2010
Head of the Graduate Program Date

Graduate School Form 20
(Revised 9/10)

PURDUE UNIVERSITY
GRADUATE SCHOOL

Research Integrity and Copyright Disclaimer

Title of Thesis/Dissertation:
Biomedical Literature Mining with Transitive Closure and Maximum Network Flow

For the degree of Master of Science

I certify that in the preparation of this thesis, I have observed the provisions of *Purdue University Executive Memorandum No. C-22,* September 6, 1991, *Policy on Integrity in Research.**

Further, I certify that this work is free of plagiarism and all materials appearing in this thesis/dissertation have been properly quoted and attributed.

I certify that all copyrighted material incorporated into this thesis/dissertation is in compliance with the United States' copyright law and that I have received written permission from the copyright owners for my use of their work, which is beyond the scope of the law. I agree to indemnify and save harmless Purdue University from any and all claims that may be asserted or that may arise from any copyright violation.

Andrew P. Hoblitzell
Printed Name and Signature of Candidate

05/07/2010
Date (month/day/year)

*Located at http://www.purdue.edu/policies/pages/teach_res_outreach/c_22.html

BIOMEDICAL LITERATURE MINING WITH TRANSITIVE CLOSURE AND MAXIMUM NETWORK FLOW

A Thesis

Submitted to the Faculty

of

Purdue University

by

Andrew P. Hoblitzell

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2011

Purdue University

Indianapolis, Indiana

To my family and BJ

ACKNOWLEDGMENTS

I want to thank family whose encouragement have been an essential ingredient in the development of this thesis. Their continuous questions regarding the status of my thesis and my graduate studies have always kept me focused and on track.

I would also like to thank the entire staff of the Department of Computer and Information Science at IUPUI for their assistance during my graduate study. I would like to thank the IUPUI and Department of Computer and Information Science for the University Fellowship and the Teaching Assistantship support which they offered to me during my graduate study. I would also like to thank Anita Park and Mark Jaeger from the Purdue University Graduate School Office, who helped make my thesis stylistically complete.

I would especially like to thank my professor and research advisor, Dr. Snehasis Mukhopadhyay for giving me an opportunity to work under his supervision. His support and advice throughout the course of my undergraduate and graduate study and research work have helped me to successfully complete this thesis. His constant encouragement and leadership made it possible for me to explore and learn about new things within Computer Science.

TABLE OF CONTENTS

Page

ABBREVIATIONS

TMS - Text Mining System

ABSTRACT

Hoblitzell, Andrew P. M.S., Purdue University, May, 2011. Biomedical Literature Mining with Transitive Closure and Maximum Network Flow. Major Professors: Snehasis Mukhopadhyay.

The biological literature is a huge and constantly increasing source of information which the biologist may consult for information about their field, but the vast amount of data can sometimes become overwhelming. Medline, which makes a great amount of biological journal data available online, makes the development of automated text mining systems and hence “data-driven discovery” possible. This thesis examines current work in the field of text mining and biological literature, and then aims to mine documents pertaining to bone biology. The documents are retrieved from PubMed, and then direct associations between the terms are computers. Potentially novel transitive associations among biological objects are then discovered using the transitive closure algorithm and the maximum flow algorithm. The thesis discusses in detail the extraction of biological objects from the collected documents and the co-occurrence based text mining algorithm, the transitive closure algorithm, and the maximum network flow which were then run to extract the potentially novel biological associations. Generated hypotheses (novel associations) were assigned with significance scores for further validation by a bone biologist expert. Extension of the work in to hypergraphs for enhanced meaning and accuracy is also examined in the thesis.

CHAPTER 1. INTRODUCTION

Bone diseases affect tens of millions of people and include bone cysts, osteoarthritis, fibrous dysplasia, and osteoporosis among others. With osteoporosis, the density of bone mineral is reduced, the proteins of the bone are altered, and the microarchitecture of the bone is disrupted. (Holroyd et al., 2008)

Osteoporosis affects an estimated 75 million people in Europe, USA and Japan, with 10 million people suffering from osteoporosis in the United States alone. Osteoporosis may significantly affect life expectancy and quality of life and is a component of the frailty syndrome. Teriparatide (parathyroid hormone, PTH), approved by the Food and Drug Administration (FDA) on 26 November 2002, is used in the treatment of some forms of osteoporosis and is the only FDA-approved drug that replaces bone lost to osteoporosis. (Saag et al., 2007)

The extraction and visualization of relationships between biological entities appearing in biological databases offers a chance to keep biologists up to date on the research and also possibly uncover new relationships among biological entities.

1.1. Motivation for the Problem of Biological Text Mining

Bioinformatics, the application of information technology and computer science to the field of molecular biology, has seen a great amount of development since the term was first coined in 1979. (Hogeweg et al., 1979) The field is varied and includes databases, algorithms, computational and statistical

techniques, and theory to solve formal and practical problems arising from the massive amounts of data.

The field has been of particular interest for informaticists and biologists to develop automatic methods to extract embedded knowledge from literature data. This particular problem of relationship extraction has been studied by numerous researchers in the field. (Oyama et al., 2002; Marcotte et al., 2001; Ono et al., 2001; Humphreys et al., 2000; Thomas et al., 2000)

1.1.1. Text Mining Applications

The biological literature is a huge and constantly increasing source of information which the biologist may consult for information about their field, but the vast amount of data can sometimes become overwhelming. It is thus important for science and new technologies to help discover new relationships and increase the efficiency of biological information workers.

Medline, which makes a great amount of biological journal data available online, makes the development of automated text mining systems and hence “data-driven discovery” possible. A method known as text mining, which will draw on elements of natural language processing and artificial intelligence, allows for the extraction of knowledge contained in the literature, and holds promising developments.

1.1.2. Artificial Intelligence

Artificial intelligence, which was coined in the middle 1950s by John McCarthy, is typically defined as “the study and design of intelligent agents” where intelligent agents try to maximize their reward within a well defined environment. (Poole et al., 1998) Machine learning and unsupervised learning without class labels are very typical within many artificial intelligence research

problems. Artificial intelligence has found broader user in the field of computer science in a wide variety of problems, including graph searching.

1.1.3. Natural Language Processing

Natural language processing (NLP), which many times is used to translate information from computer databases into readable human language, also finds an application in converting samples of human language into more formal representations which many times include parse trees, first-order logic structures, or other data structures.

Statistical natural-language processing uses stochastic, probabilistic and statistical methods to resolve difficulties. One such statistical method is Markov models, where it is assumed that there are purely random processes and that future states can be inferred from the current state. NLP has overlap with the computational linguistics, and is very closely related to the field of artificial intelligence within the subject of computer science.

1.2. Statement of Goals of the Thesis

The specific objectives of this thesis are:

1) To design a scalable and fault-tolerant text mining system which inexpensively mines from publicly available biomedical literature
2) To empirically evaluate the above TMS with regards to accuracy and meaning

1.3. Contributions of the Thesis

The main contributions of this thesis are:

- (i) In terms of the computational methodologies, for the first time, a maximal network flow based algorithm is presented to determine, in a theoretically sound manner, a confidence score for the derived transitive associations.

- (ii) In terms of the application domain, a specific pathway in bone biology consisting of a number of important proteins is subjected to the text mining approach.
- (iii) In terms of the experimental results, this paper reports for the first time (to the authors' knowledge) that a significant higher agreement with an expert's knowledge can be obtained with transitive mining than that with only direct associations. Further, both direct as well as the transitive associations were in much better agreement with the expert's knowledge than a random association matrix. These results demonstrate the usefulness of such text mining methodologies in general, and the transitive mining methods in particular.

1.4. Organization

This thesis is organized into six chapters. An Introduction, along with the problem definition and motivation, overall approach and contributions is provided in this chapter. Chapter 2 discusses the background and the related work on this thesis. Chapter 3 describes the design and implementation details of the Text Mining System and discusses the design decision and optimizations that lead to making this system more efficient. Chapter 4 provides the results of the Experiments related to accuracy and performance of the text mining system.

Chapter 5 provides the possible future extensions for the work in to hypergraphs. Finally, Chapter 6 concludes the research and identifies areas for potential future work.

CHAPTER 2. RELATED WORK

Building a text mining system that can integrate vast amounts of data is a difficult problem. Using that data to make novel predictions for bone biologists, and then condensing this meaning in to succinct, understandable, and meaningful visualizations is an even more difficult problem which has been studied by many others academically. A summarization of ongoing related work in text mining and biological literature is presented in this section.

2.1. Related Work

There are many text mining and bioinformatics examples in the literature, with each approach having its own advantages and disadvantages. This section presents some examples of these other approaches. It lays out a starting point for some of the work used in this thesis.

There are numerous varieties of bioinformatics tools available to extract knowledge through literature. One such tool is the Online Mendelian Inheritance in Man (OMIM) database and its associated morbid map MedMiner may be used to query Genecards using terms related to physiologic pathways. PubGene uses a similar design to allow the user to query genes using the HUGO approved gene symbols in the database. Other approaches for exploring the literature are explored in the following section.

2.1.1. Complementary Literatures: A Stimulus to Scientific Discovery

In their 1997 paper “Complementary Literatures: A Stimulus to Scientific Discovery”, Swanson et al. introduce informatics techniques to process the

output of Medline (Medline Plus) searches. Swanson et al. begin with a list of viruses that have weapons potential development and present findings meant to act as a guide to the virus literature to support further studies of defensive measures.

The initial Medline searches presented identified two kinds of virus literatures, those concerning genetic aspects of virulence and those concerning the transmission of viral diseases. The paper's method downloaded the Medline records for the two virus literatures and extracted all virus terms common to both.

The authors took the fact that the resulting virus list included an earlier independently published list of viruses as proof of the high degree of statistical significance of the test, thus supporting an inference that the new viruses on the list share certain important characteristics with viruses of known biological warfare interest.

2.1.2. Automatic Term Identification and Classification in Biology Texts

In the 1999 paper “Automatic Term Identification and Classification in Biology Texts”, Collier et al. discuss the rapid growth of literature databases and the difficulty that they pose to academics wishing to efficiently access relevant data. Collier et al. examine information extraction methods for the identification and classification of terms which appear in biological abstracts from the online database MEDLINE.

The paper makes use of a decision tree for classification and term candidate identification, and uses a variant of shallow parsing for identification. The paper conducted experiments against a corpus of 100 expert tagged abstracts. Collier et al.'s results indicate that while identifying term boundaries is non-trivial, a high success rate can eventually be obtained in term classification.

2.1.3. Predicting Emerging Technologies with the Aid of Text-Based Data Mining

In the 2001 paper “Predicting Emerging Technologies with the Aid of Text-Based Data Mining: A Micro Approach”, N. R. Smalheiser outlines how text mining can connect complementary pieces of information across domains of scientific literature. Smalheiser's paper again attempted to predict genetic engineering technologies that may impact on viral warfare in the future.

The paper's analysis was carried out using a combination of conventional Medline searches. Smalheiser's findings strongly indicated genetic packaging technologies as plausible candidates for study that had not previously been examined. The method of the paper was to define two fields that are hypothesized to contain complementary information, to identify common factors that bridge the two disciplines, and to progressively shape the query once initial findings were obtained. Thus the process was somewhat manual and involved a great amount of feedback from domain experts.

2.1.4. Literature Mining in Molecular Biology

“Literature mining in molecular biology”, a 2002 paper by Bruijn and Martin, examines a variety of literature mining in Medline abstracts or full text articles. It divides the process in to text categorization, named entity tagging, fact extraction, and collection-wide analysis.

Text categorization is defined to divide a collection of documents into disjoint subsets. The goal of the named entity tagging is to identify what entities or objects the article mentions. Fact extraction aims to grasp the interactions or relationships between those entities. Finally, collection wide analysis opens the door to knowledge discovery, where combined facts form the basis of a novel insight.

The paper finds that the scalability of algorithms becomes a more urgent issue as the size of the data grows, but that literature mining systems will move closer towards the human reader.

2.1.5. Accomplishments and Challenges in Literature Data Mining for Biology

"Accomplishments and challenges in literature data mining for biology", a paper by Hirschman et al., reviewed recent results in literature data mining for biology through 2002. Hirschman et al. trace literature data mining from its recognition of protein interactions to its solutions to a range of problems such as improving homology search and identifying cellular location, and note that the field has progressed from simple term recognition to the actual extraction of much more complex interactions between degrees of entities.

The paper examines successful work from the natural language processing perspective and notes that templates may now be used to increase sensitivity. The paper also examines progress in biomedical applications, specifically in organizing a challenge evaluation, the extraction of biological pathways, and automated database curation and ontology development.

2.1.6. Hybrid approach to Protein Name Identification in Biomedical Texts

In the 2005 paper "A hybrid approach to protein name identification in biomedical texts", Mostafa et al. examined a hybrid approach to identifying protein names in biomedical texts.

The paper's method uses heuristics for protein detection and uses a probabilistic model for completing protein names, while an expert protein name dictionary is complementarily consulted. Mostafa et al. automatically create a large-scale corpus annotated with protein names to train their probabilistic model.

The paper's experiments yielded results that the automatically constructed corpus is equally useful in training as compared with manually annotated corpora.

2.1.7. TransMiner

Transminer is a system developed by Narayanasamy et al. in 2004 for finding transitive associations among various biological objects using text-mining from PubMed research articles. Transminer is based on the principles of co-occurrence and transitivity for extracting novel associations.

The extracted transitive associations are given a significance score which is calculated based on the well-known tf*idf method. This method of assigning significance score is most effective and was adopted in this research. The paper by Cheng et al. in 2009 applies such transitive text mining methods to find genetic associations between breast cancer and osteoporosis diseases.

Similar work included the paper by Vaka and Mukhopadhyay on finding novel associations among biological objects and the paper by Jayadevaprakash et al. on generating association graphs of non-co-occurring text objects using text-mining. Jayadevaprakash paper specifically discusses extracting transitive associations with and without using metadata. The paper used an automated vocabulary discovery algorithm for extracting various biological objects from the text and then performed mining using the generated objects. The paper used a tf*idf method to assign significance scores to the extracted transitive associations.

Another paper by Mukhopadhyay et al. on generation of hypergraphs representing multi-way association among various biological objects presented two methods. The paper gave exhaustive and apriori methods and found same results with later method taking less computational time. Associations thus

extracted are represented in a cognitive-rich hypergraph environment in order to better assist biological researchers.

2.1.8. Summary

All the above systems have other applications or drawbacks which this thesis tries to address. This thesis presents a new method which makes use of the maximum network flow method, which is not believed to have been applied to this problem before.

2.2. New Work Presented

This work attempts to add on to work which has already been done on text mining and biological literature in some of the following ways:

(i) In terms of the computational methodologies, for the first time, a maximal network flow based algorithm is presented to determine, in a theoretically sound manner, a confidence score for the derived transitive associations.

(ii) In terms of the application domain, a specific pathway in bone biology consisting of a number of important proteins is subjected to the text mining approach.

(iii) In terms of the experimental results, this paper reports for the first time (to the authors' knowledge) that a significant higher agreement with an expert's knowledge can be obtained with transitive mining than that with only direct associations. Further, both direct as well as the transitive associations were in much better agreement with the expert's knowledge than a random association matrix. These results demonstrate the usefulness of such text mining methodologies in general, and the transitive mining methods in particular.

(iii) Further, there is the design of a generic architecture for a service-oriented mining environment that will be scalable, fault tolerant, and extendible. The realization of this architecture is in a real world application

scenario by creating a system using data publicly available through the PubMed system. An empirical validation of the premise that an efficient TMS can be achieved by using existing data using new algorithmic approaches is validated. Finally, an establishment of a working TMS with transitive predictions for future biological study resulted.

CHAPTER 3. TRANSITIVE CLOSURE AND MAXIMUM NETWORK FLOW

This chapter presents the design of the TMS (Text Mining System) employing a simple Java environment. It starts with the overall design and assumptions of the TMS in Section 3.1. The implementation of the TMS uses Java. Section 3.2 provides information about pair relationships and their relevance to the TMS. Section 3.3 explores how the transitive closure and maximum flow algorithms may successfully applied on the pair relationships to generate potentially meaningful results. Section 3.4 discusses some of the implementation details of the TMS system.

3.1. Document Representation

To extract entity relationships from the biological literature, this paper examines flat relationships, which simply state there exists a relationship between two biological entities.

A Thesaurus-based text analysis approach is used to discover the existence of relationships. The approach relies on multiple Thesauri, representing domain knowledge which can be constructed using existing organizational sources. In this case, the information has been derived by consulting experts in the domain of interest, who are users of the system as well.

The document representation step next converts the downloaded text documents into data structures which are able to be processed without the loss of any meaningful information. The process uses a thesaurus, an array T of atomic tokens (or terms) identified by a unique numeric identifier. The thesaurus

is useful for normalizing the terms versus the frequency of their occurrence and for replacing an uncontrolled vocabulary set with a controlled set. (Rothblatt et al., 1994)

The tf*idf (the term frequency multiplied with inverse document frequency) algorithm (Rothblatt et al., 1994) is applied to achieve a refined discrimination at the term representation level. The inverse document frequency (idf) component acts as a weighting factor by taking into account inter-document term distribution, over the complete collection given by:

$W_{ik}=T_{ik} \times \log(N/n_k)$

where Tik represents the number of occurrences of term Tk in document i, Ik=log(N/nk) provides the inverse document frequency of term Tik in the base of documents, N is the number of documents in the base of documents, and nk is the number of documents in the base that contains the given term Tk. To deal with the fact that the number of documents in the stream may be too small for the idf component to be meaningful, a table is maintained containing total frequencies of all terms in the base as a whole. The purpose of this step is to represent each document as a weight vector whose elements give a proportional frequency of occurrence of each term within the given document.

3.2. Pair Relationships

The goal is to discover entity pairs from the collection of retrieved text documents such that the entities in each pair are related to each other somehow. While two entities being related to each other depends on a somewhat subjective notion of “being related”, we have investigated entity-pair discovery from a collection of Medline abstracts using the Vector-Space tf*idf method and a thesaurus consisting of entity terms.

Each document d*i* is converted to an M dimensional vector where W where W_{ik} denotes the weights of the k^{th} gene term in the document and M

indicates the number of total terms in the thesaurus. W_{ik} will increase with the term frequency (T_{ik}) and decrease with the total number of documents in the collection (n_k).

After the vector representations of all of the documents have been computed, the associations between entities k and l are computed using the following equation:

$association[k][l]=\sum(W_{ik}*W_{lk})$, k=1…m, l=1…m, i=1…N

For any pair of entity terms co-occurring in even a single document, the association[k][l] will always be greater than zero. 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} entity terms. A decision can be made about the existence of a strong relationship between entities using a user-defined threshold on the elements of the Association matrix.

3.3. Application of Transitive Closure and Maximum Flow

A very useful extrapolation of these results can be achieved through “transitive text mining”.

The basic premise of transitive text mining is that if there are direct associations between objects A and B, as well as direct associations between objects B and C, then an association between A and C may be hypothesized even if the latter has not been explicitly seen in the literature. Such transitive associations may be efficiently determined by combing the transitive closure of the direct association matrix.

Further, to determine a confidence measure for such transitive (indirect/implicit) associations, we propose for the first time in this paper, the application of a well-known graph theoretic problem and algorithm, i.e. the maximal flow algorithm. In this, the direct association strengths are viewed as capacities of the corresponding edges, and the confidence measures of all pairs of transitive associations are computed as the maximal flow between the direct pairs.

This is based on what we term the “separation of evidence principle”, where evidence (i.e., a part of the capacities) once used along a transitive path may not be used again along another transitive path in defining the confidence measure of a transitive association. To our knowledge, this is the first application of the maximal flow algorithm in biomedical text mining.

The transitive closure of a binary relation R on a set X is the smallest transitive relation on X that contains R. A relation R on a set S is transitive if, for all x,y,z in S, whenever x R y and y R z then x R z. A relationship which is already transitive will have the same relationship as its transitive closure, while a relationship which is not transitive will have a different relationship as its transitive closure. The union of two transitive relations will not necessarily be transitive, so the transitive closure would have to be taken again to ensure transitivity. (Lidl and Pilz, 1998:337)

Input

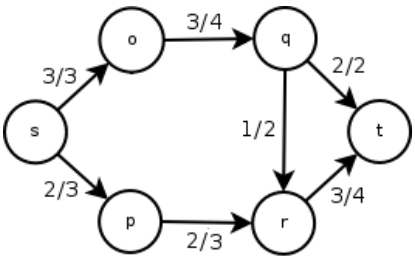
Output

The Floyd-Warshall algorithm may be used to find the transitive closure:

```
1 procedure FloydWarshall ()
2    for k := 1 to n
3       for i := 1 to n
4          for j := 1 to n
5             cost[i][j] |= (cost[i][k] & cost[k][j]);
```

The algorithm is an example of dynamic programming, and was discovered by Robert Floyd in 1962 and by Bernard Roy in 1959 and again by Stephen Warshall in 1962. In Warshall's original formulation of the algorithm, the graph is unweighted and represented by a Boolean adjacency matrix. Floyd-Warshall assumes that there are no negative cycles. (Warshall, 1962)

The maximum flow problem, seen as a special case of the circulation problem, finds a maximum flow through a single-source, single-sink flow network:



The problem is based on the premise that if every edge in a flow network has capacity, then there exists a maximal flow through the network. The problem may be solved using the Ford-Fulkerson algorithm. (Cormen et al., 2001)

The Ford-Fulkerson algorithm, published in 1956, computes the maximum flow in a flow network. The algorithm works such that as long as there is a path from the source to the sink with unused capacity on all edges in the path, flow is sent along any one of the paths. A path with such available capacity is called an augmenting path. The algorithm runs until there a maximum flow is found: (Ford et al., 1956)

```
1 procedure FordFulkerson ()
2    f(u,v)=0 for all edges (u,v)
3    While there is a path p from s to t in Gf, such that cf(u,v) > 0 for all edges:
4          Find cf(p) = min{cf(u,v)}
5             For each edge (u,v)
6                f(u,v)=f(u,v)+cf(p)
7                f(v,u)=f(v,u)-cf(p)
```

The Edmonds-Karp algorithm is an implementation of the Ford-Fulkerson method. The algorithm was published by Yefim Dinic in 1970, and again independently by Jack Edmonds and Richard Karp in 1972. The algorithm defines the augmenting path of the Ford-Fulkerson algorithm such that the path found must be the shortest path which has available capacity. This is found by a breadth-first search, allowing edges to have unit length: (Edmonds et al., 1972)

```
1  procedure EdmondsKarp ()
2  while true
2     m, P := BreadthFirstSearch(C, E, s, t)
3     if m = 0
```

```
4       break
5    f := f + m
7    v := t
8    while v ≠ s
9       u := P[v]
10      F[u,v] := F[u,v] + m
11      F[v,u] := F[v,u] - m
12      v := u
13 return (f, F)
```

The Edmonds-Karp algorithm is applied for each transitive association (a,b), where a is viewed as the source and b is viewed as the sink.

CHAPTER 4. TEXT MINING FOR BONE BIOLOGY

Chapter 3 provided the design and implementation details of the TMS. This chapter discusses all the Experiments that were carried out to validate the TMS and analyze its performance, scalability and other features. The following chapter presents the results of those experiments.

4.1. Motivation

Bone diseases such as osteoporosis, which is characterized by reduced bone mass and debilitating fractures and which affects millions of people in the United States, have limited and expensive treatments available to patients. Bone biologists may be overwhelmed by the amount of literature constantly being generated, thus the identification and extraction of existing and novel relationships among biological entities or terms appearing in the biological literature is an ongoing problem. The problem has become more and more pressing with the development of large online publicly-available databases of biological literature.

Extraction and visualization of relationships between biological entities appearing in these databases offers the opportunity of keeping researchers up-to-date in their research domain. This may be achieved through helping them visualize possible biological pathways and by generating likely new hypotheses concerning novel interactions through methods such as transitive closure network flow.

All generated predictions can be verified against already existing data, and possible new relationships can be verified against experiment. This paper presents a method for the extraction and visualization of potentially meaningful relationships.

4.2. Metrics

To test our search strategy we chose to explore potential novel relationships between NMP4/CIZ (nuclear matrix protein 4/cas interacting zinc finger protein; hereafter referred to as Nmp4 for clarity) and proteins that may interact with this signalling pathway. Briefly, Nmp4 is a nuclear matrix architectural transcription factor that represses genes that support the osteoblast phenotype (Childress et al., 2010).

Clinically, Nmp4 has been linked to osteoporosis susceptibility (Jin et al., 2009; Garcia-Giralt et al., 2005), indicating that changes in the function of this gene have real consequences in the human population.

We chose the following proteins or terms to probe the existence of unrecognized biological relationships with Nmp4: beta-catenin, zyxin, p130Cas, PTH (parathyroid hormone), PTHR1 (parathyroid hormone/parathyroid hormone-related peptide reactor 1), ECM (extracellular matrix), receptor for advanced glycation end products, HMGB1 (high mobility group box I protein), HMG-motif (high-mobility group-motif), architectural transcription factor, R-smad (receptor regulated Sma- and Mad-related protein), Smad4, CF (cystic fibrosis), actin, and alpha actinin. The rationale for these choices is explained elsewhere in detail (Childress et al., 2010).

A summary of the terms used is presented in the following legend:

ID	Term
1	beta-catenin
2	Zyxin
3	p130Cas
4	PTH
5	PTHR1
6	ECM
7	receptor for advanced glycation endproducts
8	HMGB1
9	HMG-MOTIF
10	architectural transcription factor
11	R-Smad
12	Smad4
13	Nmp4
14	actin
15	alpha-actinin

4.3. Direct Association Results

Using the terms given above and the document representation matrix laid out in the methodology section, the following direct association matrix was generated:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	125.8														
3	45000.4	627.7													
4	1444.3	27.7	24559.4												
5	7.4	0.0	387.9	10047.0											
6	449.8	96.5	6775.3	1044.6	0.0										
7	0.0	0.0	0.0	2.9	0.0	0.0									
8	9.3	5.4	1699.1	129.8	0.0	0.0	406.5								
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0							
10	917.3	5.2	3280.8	436.8	38.7	329.5	0.0	73.2	0.0						
11	130.7	0.0	60.5	0.0	0.0	71.4	0.0	0.0	0.0	0.0					
12	2003.4	0.0	3267.3	262.5	0.0	368.4	0.0	0.0	0.0	303.9	3379.3				
13	64.3	160.5	648.2	823.2	0.0	25.3	0.0	50.1	0.0	48.0	0.0	0.0			
14	3496.7	1540.2	14216.3	1759.1	21.0	1717.1	11.6	196.1	0.0	1074.5	33.7	221.4	105.0		
15	211.4	667.9	683.6	46.3	0.0	155.2	0.0	7.2	0.0	37.9	0.0	1.8	0.0	8666.9	

4.4. Transitive Closure and Maximum Network Flow Results

The Floyd-Warshall algorithm was then run over the data to determine the transitive closure of the direct association matrix. After this step, the Ford-Fulkerson algorithm was run with the Edmonds-Karp algorithm to determine the maximum network flow over the data.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	3257.07														
3	53860.84	3257.07													
4	30991.73	3257.07	30991.73												
5	10502.03	3257.07	10502.03	10502.03											
6	11032.97	3257.07	11032.97	11032.97	10502.03										
7	421.03	421.03	421.03	421.03	421.03	421.03									
8	2184.98	2184.98	2184.98	2184.98	2184.98	2184.98	421.03								
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00							
10	6545.74	3257.07	6545.74	6545.74	6545.74	6545.74	421.03	2184.98	0.00						
11	3675.73	3257.07	3675.73	3675.73	3675.73	3675.73	421.03	2184.98	0.00	3675.73					
12	6725.09	3257.07	6725.09	6725.09	6725.09	6725.09	421.03	2184.98	0.00	6545.74	3675.73				
13	1924.62	1924.62	1924.62	1924.62	1924.62	1924.62	421.03	1924.62	0.00	1924.62	1924.62	1924.62			
14	25044.91	3257.07	25044.91	25044.91	10502.03	11032.97	421.03	2184.98	0.00	6545.74	3675.73	6725.09	1924.62		
15	10478.13	3257.07	10478.13	10478.13	10478.13	10478.13	421.03	2184.98	0.00	6545.74	3675.73	6725.09	1924.62	10478.13	

4.5. Analysis of Results

The Direct Association Matrix was normalized against the maximum score divided by 1000 to give scores ranging from 0 to 1000. A thresh holding value of 152.1 was then obtained and used for examining and analyzing the data.

The MNF matrix, seen in was normalized by dividing the maximum score by 1000 to give scores ranging from 0 to 1000. A thresh holding value of 7000.2 was obtained from inspection of the scores. The values were then threshholded as follows:

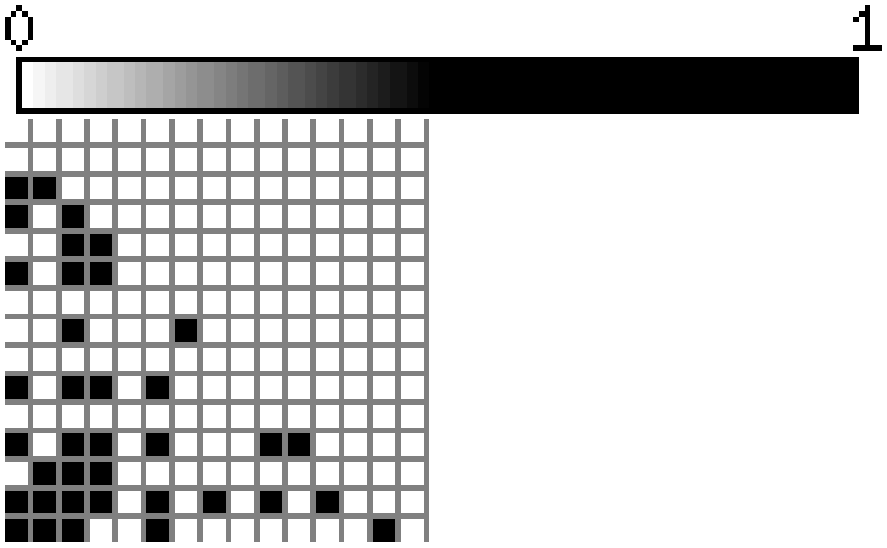
- [0, 7000.2): 0 (minimum association)
- [7000.2, 53860.84]: 1 (significant association)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	0														
3	1	0													
4	1	0	1												
5	1	0	1	1											
6	1	0	1	1	1										
7	0	0	0	0	0	0									
8	0	0	0	0	0	0	0								
9	0	0	0	0	0	0	0	0							
10	0	0	0	0	0	0	0	0	0						
11	0	0	0	0	0	0	0	0	0	0					
12	0	0	0	0	0	0	0	0	0	0	0				
13	0	0	0	0	0	0	0	0	0	0	0	0			
14	1	0	1	1	1	1	0	0	0	0	0	0	0		
15	1	0	1	1	1	1	0	0	0	0	0	0	0	1	

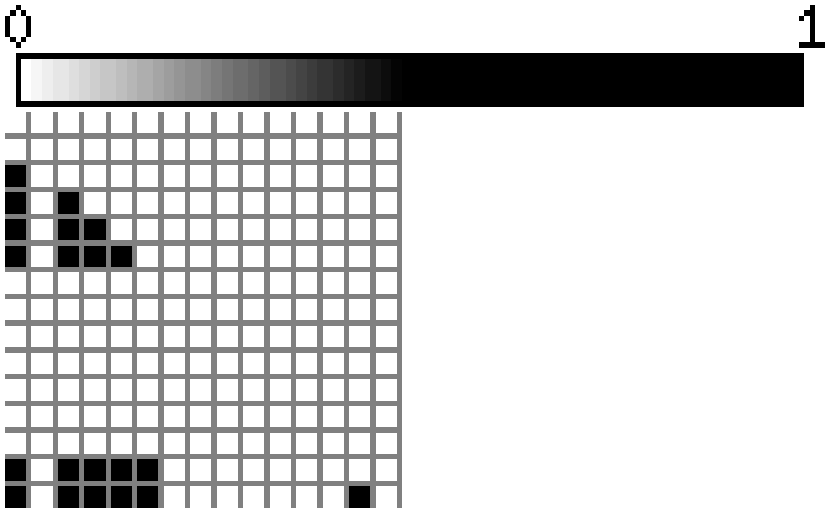
The normalized predicted scores were then able to be compared with the expert provided rankings which are shown below.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1															
2	0														
3	0	0													
4	1	0	0												
5	1	0	0	1											
6	1	1	0	1	1										
7	0	0	0	0	0	0									
8	0	0	0	0	0	0	1								
9	0	0	0	0	0	0	1	1							
10	0	0	0	0	0	0	0	1	1						
11	1	0	0	1	0	0	0	0	1						
12	1	0	0	1	0	0	0	0	1	0	1				
13	0	0	1	0	0	0	0	0	0	1	0	0			
14	0	0	0	1	1	1	0	0	0	0	0	0	0		
15	0	0	0	0	0	1	0	0	0	0	0	0	0	1	

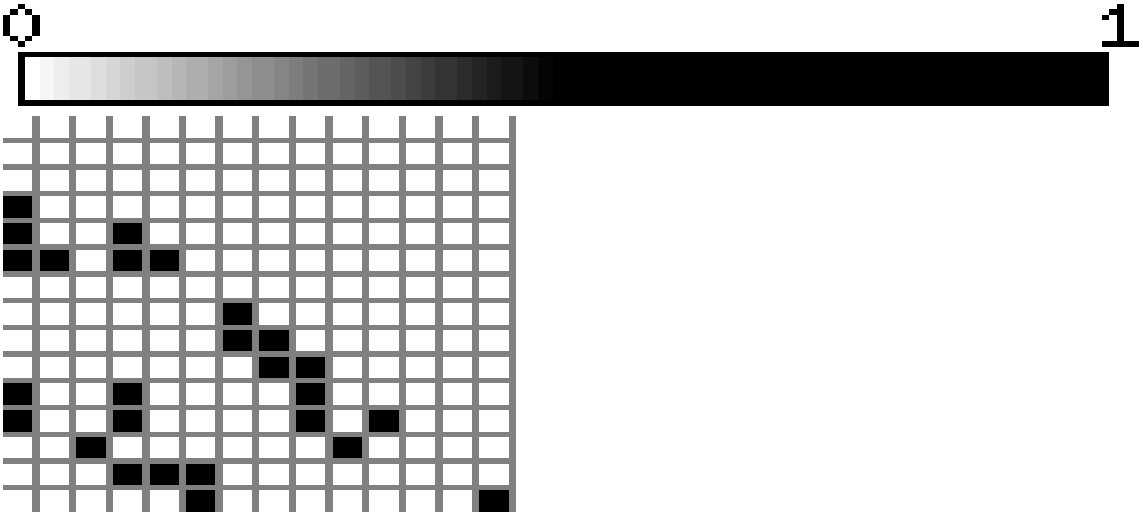
The tables were able to generate the following heat maps:



Direct Association Heat Map



MNF Heat Map



Expert Heat Map

The results from were then compared against expert provided scores. The average error was then computed as follows:

$\sum|Expert(l,k)-Predicted(l,k)|/N_r$

where Expert(l,k) is the expert provided score of a relationship between entities l and k, Predicted(l,k) is the predicted score of a given relationship between entities l and k, l is one entity, k is another entity, and N_r is the total number of relations. The resulting average error of the maximum network flow method was found to be 0.24, a significant improvement over the corresponding direct association error rate of 0.35 and a random average error rate of 0.58.

It may be seen that the application of the maximum flow algorithm to this problem offers a significant improvement over direct associations or random rankings in matching the expert provided rankings.

CHAPTER 5. EXTENSION TO HYPERGRAPHS

5.1. Introduction

A hypergraph is a generalization of a GRAPH, where EDGES can connect any number of VERTICES. (Dauber, 1969) Formally, a hypergraph *H* is a pair *H* = (*X*, *E*) where *X* is a set of elements, called *nodes* or *vertices*, and *E* is a set of non-empty subsets of *X* called *hyperedges*. (Berg et al., 1972)

Numerous problems have been studied on hypergraphs including transitive closure, transitive reduction, flow and cut problems, and minimum weight traversal problems. The “maximum hyperflow problem” is said to consist of finding a suitable hyperflow in hypergraph H which maximizes the amount of hyperflow entering the sink node. (Austiello et al., 2001) This problem has been likened to the capacitated minimum cost flow problem on directed hypergraphs. (Cambini et al., 2007)

Directed hypergraphs are a powerful tool in modeling and solving several relevant problems in many application areas. (Gallo et al., 1999) Most algorithms which solve the maximum flow problem on hypergraphs first necessitate the transformation of these hypergraphs into directed graphs, while some algorithms solve the maximum flow problem directly on hypergraphs. (Pistorius et al., 2003)

5.2. Motivation

The relevance hypergraphs to the current research is extracting associations which would involve more than two objects or nodes in the network. Hyperedges would be of length three i.e., edges involving three vertices. In this

case, edges would refer to associations and vertices would refer to objects of interest. Ternary associations (associations that involve three objects) would then be extracted using this hypergraph based approach.

The multi-way associations can be determined by co-occurrence based mining from textual literature. The association strength between three objects *i*, *j*, and *k* can be calculated by extending the previous association formula to the following:

$$Association[i][j][k] = \sum_{l=1}^{N} W_l[i] * W_l[j] * W_l[k]$$

where $W_e[i]$, $W_l[j]$, and $W_l[k]$ are the weights of objects *i*, *j*, and *k* in document *l* in the vector-space tf*idf numerical representation of the document, and *N* is the number of documents in the collection to be analyzed.

Association strengths between more than three objects can be evaluated by taking the product of their tf*idf weights in the same document, summed over all documents. Such a method is called an exhaustive extraction method because it works over all possible values.

The number of possible hyper-edges in the exhaustive extraction method grows quickly and in fact grows exponentially with the number of objects, since, the number of subsets of a set A of cardinality n (i.e., the cardinality of the power set of A) is 2^n. This is in contrast to the binary graph, encoding binary relationship, where the number of possible associations is simply n^2.

Because of the exponential growth, any exhaustive attempt to check for all hyper-edges would run into extremely high computational complexity, especially because in many cases the total number of objects would be large. Two possible ways to mitigate this are to limit the number of objects that can be related in hyper-edges or to take in to account the fact that by the very nature of co-

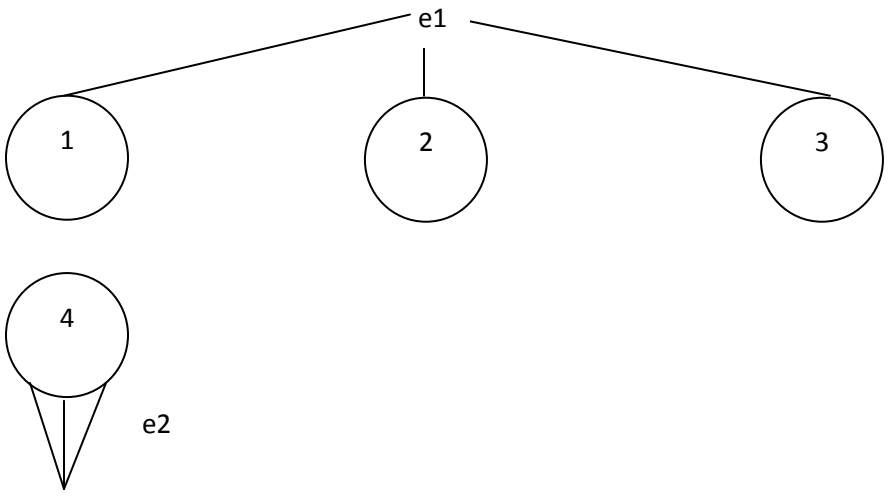
occurrence based associations, is that all q-subsets of the corresponding m objects (1<q<m) must be related. Using these assumptions would be expected to result in great improvements in efficiency.

5.3. Case Study 1

5.3.1. Diagram

A hypergraph is a graph in which generalized edges (called hyperedges) may connect more than two nodes. The network maximum flow problem is to find a flow through a single-source and single-sink flow network that is maximum. The maximum flow problem can be seen as a special case of more complex network flow problems, such as the circulation problem.

The goal in this case study is to find the maximum hyperflow through the illustrated network:



Edge	Capacity
e1	1.0
e2	2.0

5.3.2. Input

Node 1	Node 2	Node 3	Input
1	1	1	0.0
1	1	2	0.0
1	1	3	0.0
1	1	4	0.0
1	2	1	0.0
1	2	2	0.0
1	2	3	1.0
1	2	4	0.0
1	3	1	0.0
1	3	2	1.0
1	3	3	0.0
1	3	4	0.0
1	4	1	0.0
1	4	2	0.0
1	4	3	0.0
1	4	4	0.0
2	1	1	0.0
2	1	2	0.0
2	1	3	1.0
2	1	4	0.0
2	2	1	0.0
2	2	2	0.0
2	2	3	0.0
2	2	4	0.0
2	3	1	1.0
2	3	2	0.0
2	3	3	0.0
2	3	4	0.0
2	4	1	0.0

Node 1	Node 2	Node 3	Input
2	4	2	0.0
2	4	3	0.0
2	4	4	0.0
3	1	1	0.0
3	1	2	1.0
3	1	3	0.0
3	1	4	0.0
3	2	1	1.0
3	2	2	0.0
3	2	3	0.0
3	2	4	0.0
3	3	1	0.0
3	3	2	0.0
3	3	3	0.0
3	3	4	0.0
3	4	1	0.0
3	4	2	0.0
3	4	3	0.0
3	4	4	0.0
4	1	1	0.0
4	1	2	0.0
4	1	3	0.0
4	1	4	0.0
4	2	1	0.0
4	2	2	0.0
4	2	3	0.0
4	2	4	0.0
4	3	1	0.0
4	3	2	0.0
4	3	3	0.0

(Continued)

Node 1	Node 2	Node 3	Input
4	3	4	0.0
4	4	1	0.0
4	4	2	0.0
4	4	3	0.0

(Continued)

5.3.3. Output

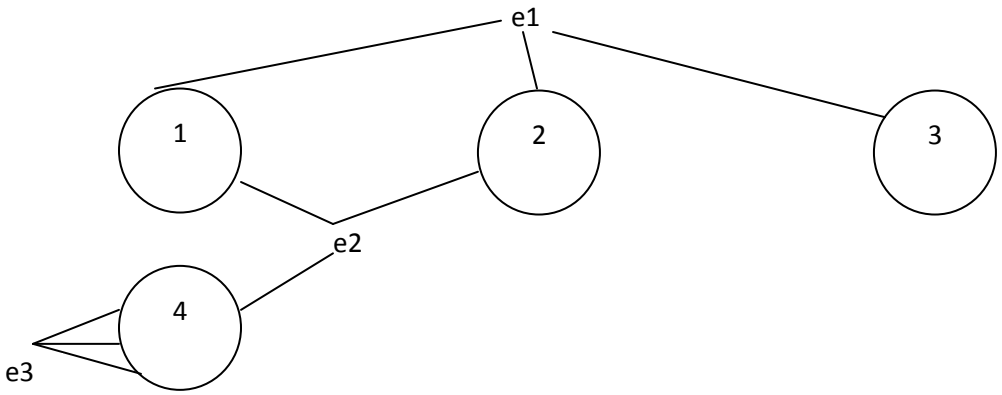
Source	Sink	Output
1	1	0.0
1	2	1.0
1	3	1.0
1	4	0.0
2	1	1.0
2	2	0.0
2	3	0.0
2	4	0.0
3	1	1.0
3	2	0.0
3	3	0.0
3	4	0.0
4	1	0.0
4	2	0.0
4	3	0.0
4	4	2.0

5.4. Case Study 2

5.4.1. Diagram

A hypergraph is a graph in which generalized edges (called hyperedges) may connect more than two nodes. The network maximum flow problem is to find a flow through a single-source and single-sink flow network that is maximum. The maximum flow problem can be seen as a special case of more complex network flow problems, such as the circulation problem.

The goal in this case study is to find the maximum hyperflow through the illustrated network:



Edge	Capacity
e1	0.5
e2	2.0
e3	4.0

5.4.2. Input

Node 1	Node 2	Node 3	Input
1	1	1	0.0
1	1	2	0.0
1	1	3	0.0
1	1	4	0.0
1	2	1	0.0

Node 1	Node 2	Node 3	Input
1	2	2	0.0
1	2	3	0.5
1	2	4	2.0
1	3	1	0.0
1	3	2	0.5
1	3	3	0.0
1	3	4	0.0
1	4	1	0.0
1	4	2	2.0
1	4	3	0.0
1	4	4	0.0
2	1	1	0.0
2	1	2	0.0
2	1	3	0.5
2	1	4	2.0
2	2	1	0.0
2	2	2	0.0
2	2	3	0.0
2	2	4	0.0
2	3	1	0.5
2	3	2	0.0
2	3	3	0.0
2	3	4	0.0
2	4	1	2.0
2	4	2	0.0
2	4	3	0.0
2	4	4	0.0
3	1	1	0.0

(Continued)

Node 1	Node 2	Node 3	Input
3	1	2	0.5
3	1	3	0.0
3	1	4	0.0
3	2	1	0.5
3	2	2	0.0
3	2	3	0.0
3	2	4	0.0
3	3	1	0.0
3	3	2	0.0
3	3	3	0.0
3	3	4	0.0
3	4	1	0.0
3	4	2	0.0
3	4	3	0.0
3	4	4	0.0
4	1	1	0.0
4	1	2	2.0
4	1	3	0.0
4	1	4	0.0
4	2	1	2.0
4	2	2	0.0
4	2	3	0.0
4	2	4	0.0
4	3	1	0.0
4	3	2	0.0
4	3	3	0.0
4	3	4	0.0
4	4	1	0.0

(Continued)

Node 1	Node 2	Node 3	Input
4	4	2	0.0
4	4	3	0.0
4	4	4	4.0

(Continued)

5.4.3. Output

Source	Sink	Output
1	1	0.5
1	2	2.5
1	3	0.5
1	4	0.5
2	1	2.5
2	2	0.5
2	3	0.5
2	4	0.5
3	1	0.5
3	2	0.5
3	3	0
3	4	0.5
4	1	0.5
4	2	0.5
4	3	0.5
4	4	4

CHAPTER 6. CONCLUSION AND FUTURE WORK

Bioinformatics, the application of information technology and computer science to the field of molecular biology, has seen a great amount of development since the term was first coined in 1979. (Hogeweg et al., 1979) The field has been of particular interest for informaticists and biologists to develop automatic methods to extract embedded knowledge from literature data.

The biological literature is a huge and constantly increasing source of information which the biologist may consult for information about their field, but the vast amount of data can sometimes become overwhelming. Medline, which makes a great amount of biological journal data available online, makes the development of automated text mining systems and hence “data-driven discovery” possible. There are numerous varieties of existing bioinformatics tools available to extract knowledge through literature.

6.1. Conclusions of the Research

The aims in this paper were to present a method which uses a maximal network flow based algorithm to determine a confidence score for the derived transitive associations. A specific pathway in bone biology consisting of a number of important proteins is subjected to the text mining approach. We show that a significant higher agreement with an expert’s knowledge can be obtained with transitive mining than that with only direct associations. Both direct as well as the transitive associations were in much better agreement with the expert’s knowledge than a random association matrix. These results demonstrated the

usefulness of such text mining methodologies in general, and the transitive mining methods in particular.

Further, there is the design of a generic architecture for a service-oriented mining environment that will be scalable, fault tolerant, and extendible. The realization of this architecture is in a real world application scenario by creating a system using data publicly available through the PubMed system. An empirical validation of the premise that an efficient TMS can be achieved by using existing data using new algorithmic approaches is validated. Finally, an establishment of a working TMS with transitive predictions for future biological study resulted.

6.2. Future Work

Future work on this problem would be very likely to include an extended set of vocabulary terms and extended work on the development of visualizations which are more meaningful to a bone biologist information expert. This would be achieved by extending the analysis in to hyperedges and hypergraphs, and through further collaboration with other computer scientists.

6.2.1. Causal Model Development

A systematic procedure for constructing causality models from text mining knowledge could also be developed.

Such a method would focus on developing Bayesian networks, a model that encodes relationships among variables of interest. A Bayesian Network would offer many benefits:

- Bayesian networks can be used to model causal relationships, and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention.
- Bayesian networks have both causal and probabilistic semantics, it would be an ideal representation for combining prior knowledge and data.

- Bayesian networks are very useful in modeling uncertainty in a domain.

Two different approaches may be used to construct Bayesian networks, a data-based approach or a knowledge-based approach. The data-based approaches use conditional independence semantics of Bayes nets to induce models from the prior data. The knowledge-based approach use causal knowledge of domain experts in constructing Bayesian networks. The knowledge-based approach is especially useful in situations where domain knowledge is crucial and availability of data is scarce.

Because text mining can derive causal knowledge, probability encoding techniques may be used to assess the numerical parameters of the resulting Bayes nets. The process of constructing Bayesian network from text then consists of three steps:

1) Derivation of the causal maps
2) Modification of the causal maps to construct Bayesian Causal Maps
3) Derivation of the Parameters of the Bayesian Causal Maps

6.2.2. Biomedical Knowledge Visualization

The development of a visualization environment would eventually aim to assist the biology domain experts to understand the integrated data sets and information and assist them in combining their domain expertise in the knowledge discovery and the hypothesis generation process.

To achieve these goals, the visualization environment would be designed to provide the visual summaries of the problem domain, and to facilitate users' reasoning process because users' decision making is driven by the visual cues and supported by interactions that could explore and manipulate the data sets. The knowledge visualization environment would eventually address the demand of knowledge discovery in the context of a multi-level graph of biological entities.

For example, in investigating Osteoporosis using association values from literature mining, the multi-level graph of biological entities is a group of inter-connected graphs, each of which have nodes belonging to a certain functional category. The graph could be acquired from diverse source to visualize a complex graph to enable knowledge discovery.

6.3. Summary

Helping bone biologists visualize possible biological pathways and generate likely new hypotheses concerning novel interactions through methods such as transitive closure and maximal network flow offer a new method to help find a cost-effective treatment to bone diseases such as osteoporosis.

The thesaurus based method presented in this paper obtains a significant improvement over random guessing. Future work on this problem would be very likely to include an extended set of vocabulary terms and extended work on the development of visualizations which are more meaningful to a bone biologist information expert. Work on extending the examination in to associations between multiple proteins or terms could also be conducted in an effort to further improve the accuracy and obtain more meaningful results. This would be achieved by extending the analysis in to hyperedges and hypergraphs.

Finally, a systematic procedure for constructing causality models from text mining using Bayesian networks could also be included. The development of a visualization environment would eventually aim to assist the biology domain experts in integrating the transitive text mining results and causal models to enhance knowledge discovery.

REFERENCES

REFERENCES

[1] Giorgio Ausiello, Paolo Giulio Franciosa, Daniele Frigioni, Directed Hypergraphs: Problems, Algorithmic Results, and a Novel Decremental Approach, Proceedings of the 7th Italian Conference on Theoretical Computer Science, p. 312-327, October 04-06, 2001.

[2] Claude Berge, and Dijen Ray-Chaudhuri, “Hypergraph Seminar”, Ohio State University 1972, Lecture Notes in Mathematics 411 Springer-Verlag.

[3] B. de Brujin and J. Martin, “Literature Mining in Molecular Biology,” Proc. EFMI Workshop Natural Language, p. 1-5, 2002.

[4] R. Cambini, G. Gallo, and M. G. Scutella. Flows on hypergraphs, Mathematical Programming B, 78 (1997) 195-217.

[5] B. Cheng, H. Vaka, and S. Mukhopadhyay. “Gene-Gene Association Study Between Breast Cancer and Osteoporosis Using Transminer Text Mining System”, p. 411-414, The proceedings of the 2009 IEEE International Conference on Bioinformatics & Biomedicine (BIBM), Washington D.C, 2009.

[6] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein (2001). “26. Maximum Flow”. Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill. p. 643-668. ISBN 0-262-03293-7.

[7] E. A. Dinic (1970). “Algorithm for solution of a problem of maximum flow in a network with power estimation”. Soviet Math. Doklady (Doklady) Vol 11: 1277-1280.

[8] E. Dauber, in Graph theory, ed. F. Harary, Addison Wesley, (1969) p. 172.

[9] Jack Edmonds and Richard M. Karp (1972). “Theoretical improvements in algorithmic efficiency for network flow problems”. Journal of the ACM 19 (2): 248-264. doi:10.1145/321694.321699.

[10] Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures,” in Proceedings of the Pacific Symposium on Biocomputing, 505-516.

[11] Ford, L. R., Fulkerson, D. R. (1956). “Maximal flow through a network”. Canadian Journal of Mathematics 8: 399-404.

[12] Giorgio Gallo, Maria G. Scutella, Directed Hypergraphs as a Modelling Paradigm, University of Pisa, 1999.

[13] L. Hirschman, J. Park, J. Tsujii, L. Wong, and C. Wu. Accomplishments and challenges in literature data mining for biology. Bioinformatics, 18:1553-1561, 2002.

[14] Hogeweg, P. (1978). Simulating the growth of cellular forms. Simulation 31, 90-96; Hogeweg, P. and Hesper, B. (1978) Interactive instruction on population interactions. Comput Biol Med 8: 319-27.

[15] “Human Genome Project Completion: Frequently Asked Questions”. genome.gov. http://www.genome.gov/11006943. Retrieved 2010-03-11.

[16] Humphreys, K., Demetrios, G., and Gaizauskas, R. (2000). Two Applications of Information.

[17] Jayadevaprakash N., Mukhopadhyay S., Palakal M. Generating Association Graphs of Non-CooccurringText Objects using Transitive Methods. Proceedings of the 2005 ACM symposium on Applied computing.

[18] Lidl, R. and Pilz, G. 1998, Applied abstract algebra, 2nd edition, Undergraduate Texts in Mathematics, Springer, ISBN 0-387-98290-6.

[19] Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001). “Mining Literature for Protein-Protein Interactions.” Bioinformatics, 17: 359-363.

[20] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US); [updated 2005 Aug 12; cited 2005 August 11]. Available from: http://medlineplus.gov/.

[21] Mukhopadhyay, S.; Palakal, M.; Maddu, K., “Multi-way Association Extraction from Biological Text Documents Using Hyper-Graphs,” Bioinformatics and Biomedicine, 2008. BIBM '08. p.257-262, 3-5 Nov. 2008.

[22] Narayanasamy, V., Mukhopadhyay, S., Palakal, M., and Potter, D. (2004). TransMiner: Mining Transitive Associations among Biological Objects from Text. Journal of Biomedical Sciences, 11(6): 864-873.

[23] Nmp4/CIZ: Road block at the intersection of PTH and load. Paul Childress, Alexander G. Robling, Joseph P. Bidwell. Bone 19 October 2009 (Article in Press doi: 10.1016/j.bone.2009.09.014).

[24] Nobata, C., Collier, N., & Tsujii, J. (1999). Automatic term identification and classification in biology texts. In Proceedings of the 5th natural language processing pacific rim symposium (p. 369-374).

[25] N. R. Smalheiser, Predicting emerging technologies with the aid of text-based data mining: the micro approach, Technovation, Volume 21, Issue 10, October 2001, Pages 689-693, ISSN 0166-4972, doi: 10.1016/S0166-4972(01)00048-7. (http://www.sciencedirect.com/science/article/B6V8B-43RJ3G5-6/2/eef94ef0cbe80d3130ae92c0ccba6637).

[26] Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), 2010 March 11. World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim.

[27] Ono T., Hishigaki H., Tanigami A., and Takagi T. (2001). “Automatic Extraction of Information on Protein-Protein Interactions from the Biological Literature,” Bioinformatics, 17(2):155-161.

[28] Oyama T., Kitano K., Satou K., and Ito T. (2002). “Extraction of Knowledge on Protein-Protein Interaction by Association Rule Discovery,” Bioinformatics, 18(5):705-714.

[29] J. Pistorius and M. Minoux. An improved direct labeling method for the max-flow min-cut computation in large hypergraphs and applicatsions. International Transactions in Operational Research, 10(1):1-11, 2003.

[30] Poole, David; Mackworth, Alan; Goebel, Randy (1998), Computational Intelligence: A Logical Approach, Oxford University Press, p. 1, http://www.cs.ubc.ca/spider/poole/ci.html.

[31] Rothblatt, J., Novick, P., Stevens, T. (1994). Guidebook to the Secretory Pathway. Oxford University Press Inc., New York.

[32] Saag KG, Shane E, Boonen S, et al. (November 2007). “Teriparatide or alendronate in glucocorticoid-induced osteoporosis”. The New England journal of medicine 357 (20): 2028-39. doi:10.1056/NEJMoa071408. PMID 18003959. http://content.nejm.org/cgi/pmidlookup?view=short&pmid=18003959&promo=ONFLNS19.

[33] Swanson, D.R. and Smalheiser, N.R. (1997). “An Interactive System for Finding Complementary Literatures: A Stimulus to Scientific Discovery,” Artificial Intelligence 91: 183-203.

[34] Thomas, J., Milward, D., Ouzounis, C., Pulman, S., and Carroll M. (2000). “Automatic Extraction of Protein Interactions from Scientific Abstracts,” in Proceedings of the Pacific Symposium on Biocomputing, 541-551.

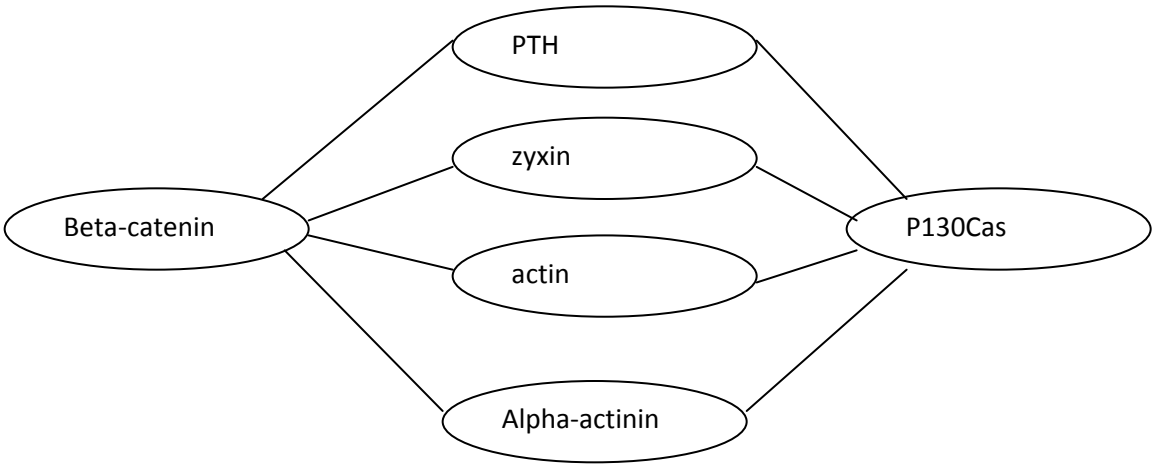
[35] Tor-Kristian Jenssen, Astrid Lægreid, Jan Komorowski & Eivind Hovig (May 2001). “A literature network of human genes for high-throughput analysis of gene expression”. Nature Genetics 28 (1): 21-28. doi:10.1038/ng0501-21. PMID 11326270.

[36] Warshall, Stephen (January 1962). “A theorem on Boolean matrices”. Journal of the ACM 9 (1): 11-12. doi:10.1145/321105.321107.

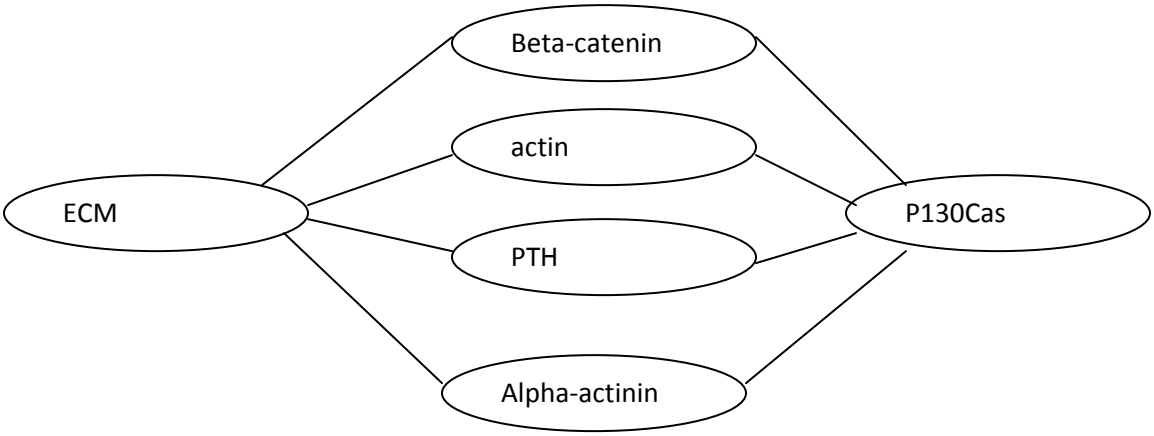
[37] Vaka H.G.G, Mukhopadyay S. Knowledge Extraction and Extrapolation Using Ancient and Modern Biomedical Literature. Accepted for 2008 IEEE BioCom Workshop 2009, conjunction with AINA 2009.

APPENDIX

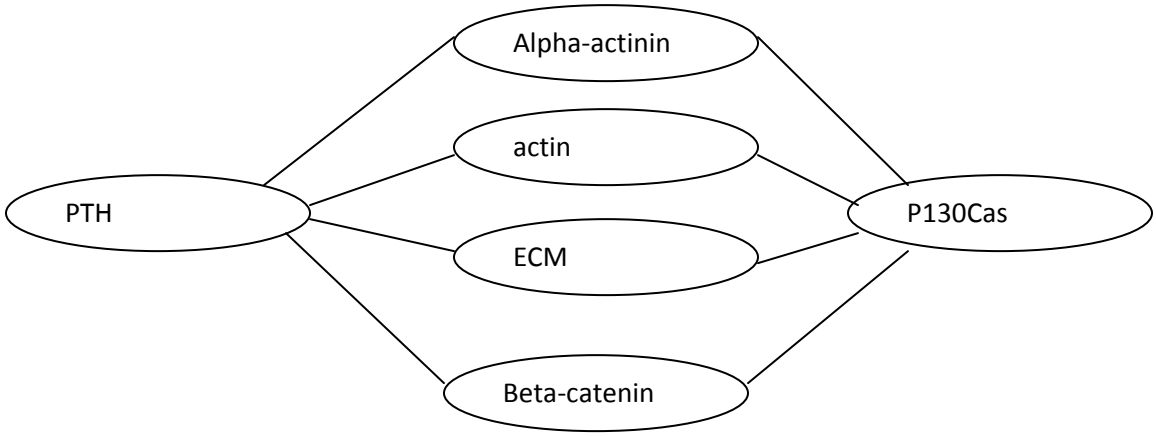
APPENDIX: SELECTED TRANSITIVITIES FOR FURTHER STUDY



- Associations computed by direct occurrences in text log-normalized over total occurrences in literature
- Aid for biologist in finding potentially meaningful transitive relationships which require verification
- Potentially new beta-catenin-zyxin-p130Cas path, beta-catenin-actin-p130Cas path, and alpha-actinin-zyxin-p130Cas paths



- Associations computed by direct occurrences in text log-normalized over total occurrences in literature
- Aid for biologist in finding potentially meaningful transitive relationships which require verification
- Potentially new ECM-beta-catenin-p130Cas, ECM-actin-p130Cas, and ECM-alpha-actinin-p130Cas paths



- Associations computed by direct occurrences in text log-normalized over total occurrences in literature
- Aid for biologist in finding potentially meaningful transitive relationships which require verification
- Potentially new PTH-alpha-actinin-p130Cas, PTH-actin-p130Cas, PTH-ECM-p130Cas, and PTH-beta-catenin-p130Cas paths.