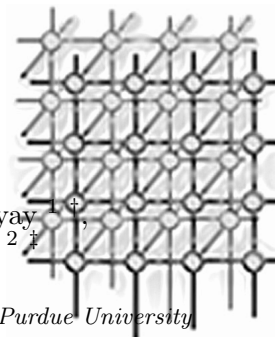


# Multi-Level Text Mining for Bone Biology



Omkar Tilak<sup>1 †</sup>, Andrew Hoblitzell<sup>1 †</sup>, Snehasis Mukhopadhyay<sup>1 †</sup>,  
Qian You<sup>1 †</sup>, Shiaofen Fang<sup>1 †</sup>, Yuni Xia<sup>1 †</sup>, Joseph Bidwell<sup>2 ‡</sup>

<sup>1</sup> *Department of Computer and Information Science, Indiana University Purdue University Indianapolis, Indianapolis, IN, 46202*

<sup>2</sup> *Department of Anatomy and Cell Biology, Indiana University School of Medicine, Indianapolis, IN, 46202*

---

## SUMMARY

Osteoporosis is characterized by reduced bone mass and debilitating fractures and is likely to reach epidemic proportions. Due to vigorous research taking place in the fields related to osteoporosis, bone biologists are overwhelmed by the amount of literature being generated on a regular basis. These problems can be alleviated by inferring and extracting novel relationships among biological entities appearing in the biological literature. With the development of large online publicly available databases of biological literature, such an approach becomes even more appealing. The novel relationships between biological pathways thus discovered constitute new hypotheses which can be verified using experiments. This paper presents a novel method called multi-level text mining for the extraction of potentially meaningful biological relationships. Multi level mining uses transitive maximum flow graph analysis coupled with set combination operations of union and intersection. Set operators are applied along and across the paths of a transitive flow graph to combine the data. In the first level of the multi-level mining process, protein domain names are used. Novel relationships between domains are extracted by the transitive text mining analysis. In the second level, these newly discovered relationships are used to extract relevant protein names. Set operators are used in various combinations to obtain different sets of results.

KEY WORDS: Bone Biology, Text Mining, Transitive Closure, Network Flow, Biological Literature, Artificial Intelligence

## 1. INTRODUCTION

---

<sup>†</sup>E-mail: {otilak, ahoblitz, smukhopa, qiyou, sfang, yxia}@cs.iupui.edu

<sup>‡</sup>E-mail: jbidwell@iupui.edu



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 micro architecture of the bone is disrupted. 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 [1].

*Teriparatide* (parathyroid hormone, *PTH*), approved by the Food and Drug Administration (FDA), is used in the treatment of some forms of osteoporosis and is the only FDA-approved drug that replaces bone lost to this disease [2]. However, it is the least cost-effective therapy therefore considerable research has been devoted to improving the efficacy of this drug [3].

Due to vigorous research taking place in the fields related to osteoporosis, bone biologists are overwhelmed by the amount of literature being generated on a regular basis. These problems can be alleviated by inferring and extracting novel relationships among biological entities appearing in the biological literature. With the development of large online publicly available databases of biological literature, such an approach becomes even more appealing. The novel relationships between biological pathways thus discovered constitute new hypotheses which can be verified using experiments. This paper presents a text mining based method for the extraction of potentially meaningful biological relationships using transitive maximum flow graph analysis. 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 has the potential to uncover new relationships among biological entities.

Text mining is the process of searching, collating, inferring and deriving useful knowledge from textual data. Text mining has recently been applied in the domain of the biomedical research for discovering relationships between biological entities including proteins, drugs, and metabolic pathways. In this paper, we describe a novel multi-level, transitive text mining strategy for the extraction of relationships between two objects (protein names, protein domain names etc.) using graph theoretic algorithms. In particular, we propose a multi-level, transitive text mining approach to extract and predict relationships among the functional domains of *Nmp4*, *Rage*, and their known associated proteins within the context of *PTH*- and load-responsive bone cell signaling pathways that enhance bone formation [3]. Multi-level text mining combines transitive text mining with the set operations (namely union and intersection). The first level of transitive text mining generates direct and transitive association graph between various objects. A comparison of these graphs is done to obtain objects of interest and these objects are considered in the second level of mining. The set operations are applied on the paths in the transitive flow graph. Set operations are applied to the segments along the path and then across various paths in the transitive graph. Various combinations of these operators yield different and interesting results (see Section 4).

The contributions of the current paper with respect to the state-of-the-art in this field can be summarized as follows:

1. A maximal network flow based algorithm is used to determine, in a theoretically sound manner, a confidence score for the derived transitive associations.



2. Various pathways in bone biology are subjected to the text mining approach. The results obtained from individual pathways are combined to make novel predictions.
3. In terms of the experimental results, a significant agreement with an expert's knowledge was obtained with transitive mining than that with only direct associations. This demonstrates the usefulness of such text mining methodologies in general, and the transitive mining methods in particular.

This paper is organized as follows. Section 2 gives a brief review of the text mining approaches used in the biological domain. Section 3 describes the proposed multiple-level transitive text mining technique using maximum flow analysis coupled with set combination operations along multiple pathways. Section 4 describes the experimental results obtained from the application of these algorithms. Section 5 concludes the paper.

## 2. RELATED WORK

Text mining algorithm has been used to discover novel uses for *Curcuma longa* (turmeric), a dietary substance [4]. Several disease such as retinal diseases, Crohn's disease and disorders related to the spinal cord were identified. The text mining process is initiated with a single topic  $T$  of any type, such as a disease, a pharmacological substance or a gene. Starting with a topic and navigating through intermediate topics, the goal is to reach terminal topics that shed new light on  $T$ . A topic's profile is a representation of the topic, which is built from a set of documents that are relevant to the topic. A topic profile is a set of terms (single words and phrases) extracted from these documents. A numerical weight is associated with each term using the  $TF \times IDF$  algorithm. This weight indicates the relative importance of the term in representing the topic. Then top terms are selected to form next set of topics and so on.

The utility of physical protein interactions for determining gene-disease associations is studied in [5] by examining the performance of seven recently developed computational methods. It was found that random-walk approaches individually outperform clustering and neighborhood approaches. The paper shows that combining these methods into a consensus method yields Pareto optimal performance. A quantification of the effect of a diffuse topological distribution of disease-related proteins on the prediction quality is presented. This allows the identification of diseases especially amenable to network-based predictions and others for which additional information sources are absolutely required.

The study in [6] aims to identify and rank genes involved in osseous augmentation to obtain groups with more numerous predicted associations called the leader gene clusters. An iterative search was performed for which only genes involved in a specific process were identified. The iterative search comprises of a consecutive expansion-filtering loop. For each gene, predicted associations with all other involved genes were obtained from the Web-available database (STRING database) and the weighted number of links (WNL), given by the sum of only high-confidence predicted associations (results with a score  $\geq 0.9$ ), allowing gene ranking. Genes belonging to higher clustering classes were then identified.

An inference network approach is used in [7] to predict implicit gene-disease associations. Genes and diseases are represented as nodes and are connected via two types of intermediate



nodes: gene functions and phenotypes. To estimate the probabilities involved in the model, two learning schemes are compared; one baseline using co-annotations of keywords and the other taking advantage of free text. Additionally, domain ontologies are used to complement data sparseness and examine the impact of full text documents. The validity of the proposed framework is demonstrated on the benchmark data set created from real-world data.

In [8], an approach for automated pathway synthesis is proposed that acquires facts from hand-curated knowledge bases. To comprehend the incompleteness of the knowledge bases, the facts are automatically extracted from Medline abstracts. Logical reasoning is applied to the acquired facts based on the biological knowledge about pathways. The reasoning is represented by encoding the logic representation in the form of pre- and post-conditions of pharmacokinetic properties that describe the course of drug disposition in the body, which includes drug absorption, distribution, metabolism and excretion. By representing such biological knowledge, the reasoning component is capable of assigning ordering to the acquired facts and interactions that is necessary for pathway synthesis. An existing pharmacokinetic pathway available in PharmGKB is reconstructed using this approach. The results show that the approach is capable of synthesizing these pathways and uncovering information that is not available in the manually annotated pathways.

A curated source (PharmGKB and DrugBank) and an automatic text-mining source (Pharmspresso) is used in [9] for extraction of drug-gene relationships. A corpus of full-text articles is first tokenized into sentences. Pharmspresso then marks up the sentences by identifying terms associated with genes and drugs. A drug-gene network is then created by drawing edges between genes and drugs that co-occur at the sentence level. One classifier is trained using each of the two types of knowledge sources and then validated against a gold standard set of drug-gene relationships to allow comparison of the two sources.

In [10], authors report the results of text mining for a bone biology pathway including SMAD genes. A text mining tool is used to analyze the PubMed literature database and integrates the available genomic information to provide a detailed mapping of the genes and their interrelationships within a particular network such as osteoporosis. To filter the most significant findings, a ranking system is devised to rate our predicted novel genes. The results obtained from the text mining program show that existing genomic data within the PubMed database can effectively be used to predict novel genes for osteoporosis research that have not previously been reported in the literature.

Transminer [11] is a system developed 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  $TF \times IDF$  method. This method of assigning significance score is used in this research.

There are several PPI databases and text mining tools that have been developed in the bioinformatics community [12]. The *STRING* project is a representative of such tools that attempts to predict PPI based on multiple sources and heuristic techniques. However, the difference and novelty of our approach to these database/tools in general are delineated below.

1. To our knowledge, none of these tools use transitivity as the basis for predicting relationships, nor do they use graph algorithms (such as Maximum flow) to estimate the



strength of indirect evidences. In our view, transitivity is one of the primary mechanisms human researchers use in postulating new hypothesis.

2. Our techniques are not limited to proteins. The nodes in the association graph can be any biological entity of interest (such as drugs, diseases etc.).
3. Most of these tools pre-compute the binary associations. Our approach computes "on-the-fly" associations for a set of relevant objects. So the results can be easily adapted to changing literature and/or granularity of the specific objects of interest.
4. To our knowledge, none of these tools work with protein domains.

### 3. MULTI-LEVEL TRANSITIVE TEXT MINING

#### 3.1. Level 1 - Extract Direct and Transitive Relationships Between A Pair of Terms

To extract entity relationships from the biological literature, we use a Thesaurus-based text mining approach. In our case, a thesaurus representing domain knowledge was constructed by consulting experts in the bone biology, who are users of our system as well. Each text document is converted in a format that is amenable to mathematical processing while still keeping the original meaning intact. The thesaurus is an array  $T$  of atomic terms identified by a unique numeric identifier. In our case, the thesaurus has *protein domain names* that are relevant to the bone biology. We use  $TF \times IDF$  (Term Frequency multiplied with Inverse Document Frequency) algorithm [13] to convert the document in a mathematically accessible format.  $TF \times IDF$  algorithm calculates the weight of each thesaurus term in every document as follows:  $W_{ik} = T_{ik} \times \log(N/n_k)$  where  $T_{ik}$  represents the number of occurrences of term  $T_k$  in document  $i$ ,  $I_k = \log(N/n_k)$  provides the inverse document frequency of term  $T_{ik}$  in the base of  $N$  number of documents and  $n_k$  is the number of documents in the base that contains the given term  $T_k$ .

The association strength between two terms in the Thesaurus is then calculated as  $association[k][l] = \sum_{i=1}^N W_{ik} \times W_{il}, k = 1, 2, \dots, m; l = 1, 2, \dots, m$ , where  $m$  is the total number of terms in the Thesaurus. Thus the 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 terms. A suitable threshold is then used to convert these arbitrary association values into a binary association format. Thus if  $association[k][l] \geq \text{threshold}$   $association[k][l] = 1$ ; else  $association[k][l] = 0$

The  $association[k][l]$  values thus obtained provide a *direct* association between two terms in the Thesaurus. We extrapolate and exploit the direct association values to calculate *transitive* association between different terms. If there is a direct association between terms  $A$  and  $B$  as well as a direct association between terms  $B$  and  $C$ , then a *transitive* association between  $A$  and  $C$  may be hypothesized even if the latter has not been explicitly seen in the literature. Such transitive associations can be calculated by using the transitive closure of the direct association matrix. To calculate the association strength for such transitive relationships, we propose the application of the maximal flow algorithm [14]. The direct association strengths are viewed



as capacities of the corresponding edges, and the association strength of all pairs of transitive associations are computed as the maximal flow between the direct pairs. This process uses the separation of evidence” principle, where evidence (i.e., a part of the direct association strength) once used along a transitive path are not used again along another transitive path in defining the confidence measure of a transitive association. We used Edmonds-Karp [15] algorithm to implement the maximum flow analysis. The Edmonds-Karp 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 augmenting path found is such that it is the shortest path which has available capacity. This path is found by a breadth-first search. The algorithm runs until maximum flow is found. To our knowledge, this is the first application of the maximal flow algorithm in biomedical text mining.

### 3.2. Use Set Operations to Extract Further Relevant Information

In addition to Thesaurus, level 2 of multi-level mining incorporates another list of terms called as a Search List. Thesaurus is used to find the relevant set of documents and these documents are then further explored to find the terms appearing in the Search List. Level 2 of the multi-level text mining begins with a pair of terms (from Thesaurus) who have transitive association but no direct association. These pairs are interesting because they represent a hypothesis obtained from our analysis during level 1. No direct association between a pair of terms means that there is no information in the literature which mentions these two domains together in the same document. In other words, nobody has explored any relationship between them. But a presence of transitive association between this same pair indicates that our analysis has spotted this pair as a potential pair which has not been considered yet and which might be useful for further analysis.

To further analyze each domain pair, we create a domain pair association subgraph. This graph contains all the paths that connect the two domains under consideration. Figure (1) shows a hypothetical association subgraph between two protein domains  $X$  and  $Y$ . Each node in the graph represents a domain name. All the domain names that appear in the subgraph belong to the original Thesaurus containing the term list. As the graph demonstrates, there are multiple ”paths” that ”connect” domains  $X$  and  $Y$ . Each path can have different length and nodes across different paths can be distinct. Each segment of the complete path represents a direct association between two domains (say  $I_1$  and  $I_2$ ). For each such domain pair, all the abstracts that show direct association between the domain pair  $I_1$  and  $I_2$  were collected. This set was then analyzed to obtain a list of protein names mentioned in this set of abstracts. Similar analysis was carried out for every segment of all the paths present in the subgraph. Lets indicate the protein list corresponding to the direct association between domains pair  $I_1$  and  $I_2$  as  $P_{I_2}^{I_1}$ . Also, we use  $P_1^{Path}$  to indicate the protein list generated by processing the direct associations between domains on  $Path1$  and  $P^{SGraph}$  to indicate the protein list generated for the entire subgraph. We use  $\cup$  to indicate union of sets with duplicates eliminated and  $\cap$  indicates the intersection of sets. We also calculate the frequency of occurrence of every protein based on all the sets of abstracts corresponding to every path segment of the entire subgraph. We assume that the frequency count indicates the ”goodness” of a protein in making

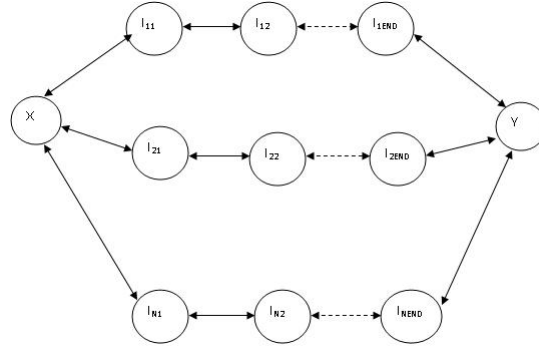


Figure 1. Transitive Flow Graph

the hypothesis. This frequency count is used to rank the proteins in decreasing order of their frequency thus putting "better" proteins at the top. By combining the protein lists generated along each individual segment of the path, we generate a list for the entire path. Then by combining the protein names list generated for each path, we generate the list corresponding to the entire subgraph. The list combinations are done using the union and intersection set operators described earlier.

We explored following combinations of set operations for various domain pairs.

**UNION along a path and UNION across all paths (UAUA):** For each path in the subgraph, we first calculate the *union* of protein lists *along* the segments of the path to produce the path protein list. Then we calculate the *union* of all the path protein lists to get the protein list for the entire subgraph. The subgraph list was ordered in the decreasing order based on the frequency count of the individual proteins.

$$P_i^{Path} = P_{I_{i1}}^X \cup P_{I_{i2}}^{I_{i1}} \cup P_{I_{i3}}^{I_{i2}} \cup \dots \cup P_Y^{I_{iEnd}}$$

$$P^{SGraph} = \bigcup_{i=1}^N P_i^{Path}$$

**UNION along a path and INTERSECTION across all paths (UAIA):** For each path in the subgraph, we first calculate the *union* of protein lists *along* the segments of the path to produce the path protein list. Then we calculate the *intersection* of all the path protein lists to get the protein list for the entire subgraph. The subgraph list was ordered in the decreasing order based on the frequency count of the individual proteins.

$$P_i^{Path} = P_{I_{i1}}^X \cup P_{I_{i2}}^{I_{i1}} \cup P_{I_{i3}}^{I_{i2}} \cup \dots \cup P_Y^{I_{iEnd}}$$

$$P^{SGraph} = \bigcap_{i=1}^N P_i^{Path}$$



**INTERSECTION along a path and UNION across all paths (IAUA):** For each path in the subgraph, we first calculate the *intersection* of protein lists *along* the segments of the path to produce the path protein list. Then we calculate the *union* of all the path protein lists to get the protein list for the entire subgraph. The subgraph list was ordered in the decreasing order based on the frequency count of the individual proteins.

$$P_i^{Path} = P_{I_{i1}}^X \cap P_{I_{i2}}^{I_{i1}} \cap P_{I_{i3}}^{I_{i2}} \cap \dots \cap P_Y^{I_{iEnd}}$$

$$P^{SGraph} = \bigcup_{i=1}^N P_i^{Path}$$

## 4. EXPERIMENTAL RESULTS

### 4.1. Using Only Level-1

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*) and proteins that may interact with this signaling pathway. Briefly, *Nmp4* is a nuclear matrix architectural transcription factor that represses genes that support the osteoblast phenotype [3]. Clinically, *Nmp4* has been linked to osteoporosis susceptibility, indicating that changes in the function of this gene have real consequences in the human population [16]. We chose the following proteins or terms to probe the existence of unrecognized biological relationships with *Nmp4*: *beta catenin*, *zyxin*, *p130Cas*, *PTH* (parathyroid hormone), *PTHrP* (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 [3].

Using the protein names (terms) given above and the methodology described in 3.1, following direct and transitive association heat maps were generated:

The direct association heat map was normalized against the maximum score divided by 1000 to give scores ranging from 0 to 1000. A threshold value of 152.1 was then used for examining and analyzing the data.

The transitive association heat map was normalized by dividing the maximum score by 1000 to give scores ranging from 0 to 1000. A threshold value of 7000.2 was obtained from inspection of the scores.

The transitive association heat map was then compared with the expert provided Heat Map.

$$\sum \frac{|Expert(l,k) - Predicted(i,k)|}{N_f}$$

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*, and

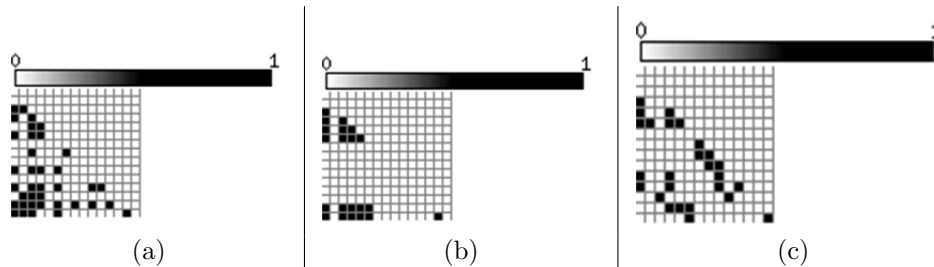


Figure 2. Proteins Association Heat Maps: (a) Direct (b) Transitive (c) Expert

$N_f$  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  $0.58$ . This indicates 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.

#### 4.2. Level-1 Followed by Level-2

The Thesaurus consists of Protein Domain Names and the Search List consists of all the proteins found in Humans. In particular, Thesaurus consists of following protein domains: *SH3 domain*, *Proline-rich*, *SH2-binding*, *PTB-binding*, *Serine-rich*, *Src-binding*, *HLH motif*, *LD domain*, *LIM domain*, *FERM domain*, *FAT domain*, *PR1 domain*, *PR2 domain*, *SH3-binding*, *Cys2His2 zinc finger*, *polyglutamine/alanine repeat*, *serine/threonine-rich*, *Armadillo repeat*, *MH1*, *MH2*, *SH2 domain*, *Bilobed kinase domain*.

Level-1 analysis was carried out on the terms in the Thesaurus to obtain direct and transitive association heat maps to discover novel relationships between protein domains. There are many such "novel hypothesis" domain pairs. To begin Level-2 analysis, we used the domain pairs which have most "sticky" domains, i.e. those domains found in adaptor proteins like *p130Cas*, a central molecule in our mechanosome network [3]. Adaptor proteins typically contain numerous domains and interact with a wide variety of proteins. Therefore we chose domain pairs *SH3-Domain - SH2-binding* (which are on *p130Cas*) and *SH3-Domain - MH2* for further analysis.

*UNION along a path and UNION across all paths (UAUA)*: For *SH3-Domain - SH2-binding*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency count are as follows: *SHP-1*, *FAK*, *proline-rich tyrosine kinase 2*, *SHP-2*, *Akt*, *Nck*, *PI3K*, *c-Abl*, *CD3*, *Cas*, *SLP-76*, *Nef*, *IRS-1*, *PYK2*, *SHIP*.

For *SH3-Binding - MH2*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency count are as follows: *PRP*, *FAK*, *proline-rich tyrosine kinase 2*, *p38*, *tRNA*, *ovalbumin*, *Nck*, *c-Abl*, *Nef*, *PYK2*, *Cas*, *activin*, *Akt*, *p62*, *IL-2*.

*UNION along a path and INTERSECTION across all paths (UAIA)*: For *SH3-Domain - SH2-binding*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency

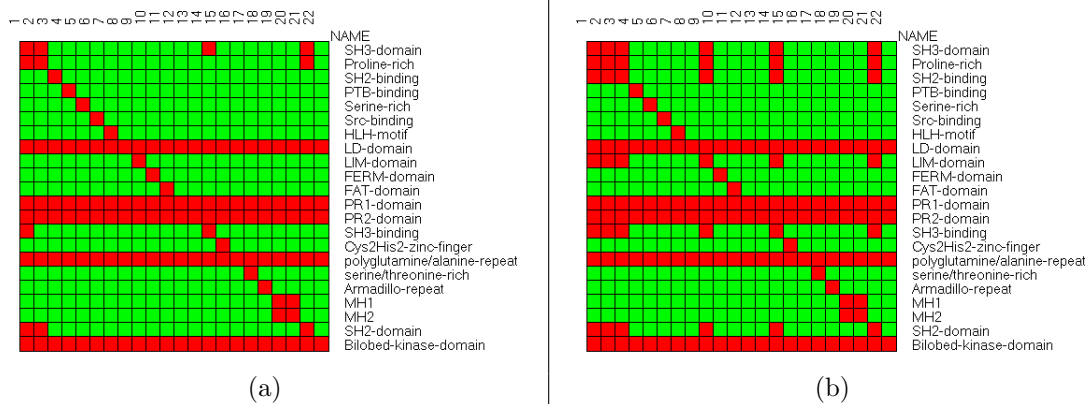


Figure 3. Protein Domains Association Heat Maps: (a) Direct (b) Transitive

count are as follows: *SHP-1*, *FAK*, *SHP-2*, *Akt*, *Nck*, *PI3K*, *c-Abl*, *CD3*, *Cas*, *SLP-76*, *Nef*, *IRS-1*, *PYK2*, *SHIP*, *E1*.

For *SH3-Binding - MH2*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency count are as follows: *FAK*, *p38*, *Nck*, *c-Abl*, *Nef*, *Cas*, *activin*, *Akt*, *p62*, *IL-2*, *CD3*, *PI3K*, *p47phox*, *c-Cbl*, *NF-kappaB*.

*INTERSECTION along a path and UNION across all paths (LAUA)* : For *SH3-Domain - SH2-binding*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency count are as follows: *SHP-1*, *FAK*, *SHP-2*, *Akt*, *Nck*, *PI3K*, *c-Abl*, *CD3*, *Cas*, *SLP-76*, *Nef*, *IRS-1*, *PYK2*, *SHIP*, *E1*.

For *SH3-Binding - MH2*, top 15 members of the  $P^{SGraph}$  set arranged in the decreasing order of frequency count are as follows: *FAK*, *p38*, *Nck*, *c-Abl*, *Nef*, *Cas*, *activin*, *Akt*, *p62*, *IL-2*, *CD3*, *PI3K*, *p47phox*, *c-Cbl*, *NF-kappaB*.

The identification of *SHP-1* as a potential member of the mechanosome network was unexpected and prompted a re-evaluation of our original hypothesis [3, 17]. *SHP-1* is a Src-homology 2 domain (SH2)-containing protein tyrosine phosphatase (PTP) expressed in numerous cells (see [18] for review). This protein suppresses osteoclastogenesis [19, 20], perhaps in part by attenuating cell adhesion [21]. Additionally, we have preliminary data suggesting a role for *Nmp4* in osteoblast adhesion. Therefore, on the basis of our in silico results and laboratory data we have proposed that the interactions between *Nmp4*, *p130Cas*, and *SHP-1* form a motif within the mechanosome network that regulates adhesion-mediated differentiation in osteoblasts and osteoclasts (and perhaps numerous other cell types).

The significant frequency of both *FAK* (focal adhesion kinase) and *PYK2* (protein tyrosine kinase 2) in both the "SH3 domain + SH2-binding domain" and the "SH3-binding domain + MH2 domain" search results are consistent with the fact that these two proteins are very similar in their domain arrangement with conservation of proline-rich regions, a 60% sequence identity



in the central kinase domain, and identical positions of four tyrosine phosphorylation sites [22]. *FAK* and *PYK2* appear to play roles in osteoblast mechanotransduction [23, 24]; *PYK2* mediates the organization of the osteoclast cytoskeleton and osteoclast adhesion [25, 26, 27]. The role of *FAK* in osteoclasts is less clear but most data point to an adhesion-related function in these cells [27].

## 5. CONCLUSIONS

Helping bone biologists visualize possible biological pathways and generate likely new hypotheses concerning novel interactions through multi-level text mining using transitive closure and maximal network flow offers a new method to help find a cost-effective treatment to bone diseases such as osteoporosis. Of particular interest in our text mining analysis are those pairs for which the literature fails to indicate a direct association but our transitive text mining indicates an association. Such interesting pairs are located in the first-level text mining analysis. We used such interesting pairs to carry out the second-level text mining analysis which glean further interesting outcomes in the form of proteins in the *mechanosome* network. The direct association matrix generated in our work suggests that there is a strong relationship in the literature between *NMP4* and *PTH*, *p130Cas*, *zyxin*, *actin*, and *beta-catenin* in decreasing order. Our work with maximum network flow suggests that there may be a more buried and weak relationship between *NMP4* and almost all of the terms. 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.

## REFERENCES

1. J. Compston, "Osteoporosis: social and economic impact," *Radiol Clin North Am.*, vol. 48, pp. 477–482, 2010.
2. S. E. Saag, K. G. and S. e. a. Boonen, "Teriparatide or alendronate in glucocorticoid-induced osteoporosis," *The New England Journal of Medicine*, vol. 20, p. 20282039, 2007.
3. P. Childress, A. Robling, and J. Bidwell, "Nmp4/ciz: Road block at the intersection of pth and load," *Bone*, 2009.
4. P. Srinivasan and B. Libbus, "Mining medline for implicit links between dietary substances and diseases," *Bioinformatics*, vol. 20, pp. i290–i296, 2004.
5. S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, pp. 1057–1063, 2010.
6. L. Sbordone, C. Sbordone, N. Filice, G. Menchini-Fabris, M. Baldoni, and P. Toti, "Gene clustering analysis in human osseous remodeling," *J Periodontol*, vol. 80, pp. 1998–2009, 2009.
7. K. Seki and J. Mostafa, "Discovering implicit associations between genes and hereditary diseases," *Pacific Symposium on Biocomputing*, vol. 12, pp. 316–327, 2007.
8. L. Tari, S. Anwar, S. Liang, J. Hakenberg, and C. Baral, "Synthesis of pharmacokinetic pathways through knowledge acquisition and automated reasoning," *Pacific Symposium on Biocomputing*, vol. 15, pp. 465–476, 2010.



9. Y. Garten, N. Tatonetti, and R. Altman, "Improving the prediction of pharmacogenes using text-derived drug-gene relationships," *Pacific Symposium on Biocomputing*, vol. 15, pp. 305–314, 2010.
10. V. K. Gajendran, J. Lin, and D. P. Fyhrie, "An application of bioinformatics and text mining to the discovery of novel genes related to bone biology," *Bone*, vol. 40, pp. 1378–1388, 2007.
11. V. Narayanasamy, S. Mukhopadhyay, M. Palakal, and D. Potter, "Mining transitive associations among biological objects from text," *Journal of Biomedical Sciences*, vol. 11, pp. 864–873, 2004.
12. <http://mips.helmholtz-muenchen.de/proj/ppi/>.
13. J. Rothblatt, P. Novick, and T. Stevens, *Guidebook to the Secretory Pathway*. Oxford University Press Inc., 1994.
14. L. R. Ford and D. R. Fulkerson, "Maximal flow through a network," *Canadian Journal of Mathematics*, vol. 8, pp. 399–404, 1956.
15. J. Edmonds and R. Karp, "Theoretical improvements in algorithmic efficiency for network flow problems," *Journal of the ACM*, vol. 19, pp. 248–264, 1972.
16. H. Jin, R. van't Hof, O. Albagha, and S. Ralston, "Promoter and intron 1 polymorphisms of *colla1* interact to regulate transcription and susceptibility to osteoporosis," *Hum Mol Genet*, vol. 18, pp. 2729–2738, 2009.
17. F. Pavalko, S. Norvell, D. Burr, C. Turner, R. Duncan, and J. Bidwell, "A model for mechanotransduction in bone cells: the load-bearing mechanosomes," *Cell Biochem.*, vol. 80, pp. 104–112, 2003.
18. U. Lorenz, "Shp-1 and shp-2 in t cells: two phosphatases functioning at many levels," *Immunol Rev*, vol. 228, pp. 342–359, 2009.
19. E. van Beek, T. de Vries, L. Mulder, T. Schoenmaker, K. Hoeben, T. Matozaki, G. Langenbach, G. Kraal, V. Everts, and T. van den Berg, "Inhibitory regulation of osteoclast bone resorption by signal regulatory protein alpha," *FASEB*, vol. 23, pp. 4081–4090, 2009.
20. K. Aoki, E. Didomenico, N. Sims, K. Mukhopadhyay, L. Neff, A. Houghton, M. Amling, J. Levy, W. Horne, and R. Baron, "The tyrosine phosphatase shp-1 is a negative regulator of osteoclastogenesis and osteoclast resorbing activity: increased resorption and osteopenia in *me(v)/me(v)* mutant mice," *Bone*, vol. 25, pp. 261–267, 1999.
21. S. Granot-Attas and A. Elson, "Protein tyrosine phosphatases in osteoclast differentiation, adhesion, and bone resorption," *Eur J Cell Biol*, vol. 87, pp. 479–490, 2008.
22. S. Mitra, D. Hanson, and D. Schlaepfer, "Focal adhesion kinase: in command and control of cell motility," *Nat Rev Mol Cell Biol*, vol. 6, pp. 56–68, 2005.
23. S. Young, R. Gerard-O'Riley, M. Harrington, and F. Pavalko, "Activation of *nf-kappab* by fluid shear stress, but not *tnf-alpha*, requires focal adhesion kinase in osteoblasts," *Bone*, vol. 47, pp. 74–82, 2010.
24. A. Guignandon, N. Boutahar, A. Rattner, L. Vico, and M. Lafage-Proust, "Cyclic strain promotes shuttling of *pyk2/hic-5* complex from focal contacts in osteoblast-like cells," *Biochem Biophys Res Commun*, vol. 343, pp. 407–414, 2006.
25. T. Miyazaki, A. Sanjay, L. Neff, S. Tanaka, W. Horne, and R. Baron, "Src kinase activity is essential for osteoclast function," *J Biol Chem*, vol. 279, pp. 17 660–17 666, 2004.
26. P. Lakkakorpi, A. Bett, L. Lipfert, G. Rodan, and T. Duong le, "Pyk2 autophosphorylation, but not kinase activity, is necessary for adhesion-induced association with *c-src*, osteoclast spreading, and bone resorption," *J Biol Chem*, vol. 278, pp. 11 502–11 512, 2003.
27. W. Xiong and X. Feng, "Pyk2 and *fak* in osteoclasts," *Front Biosci*, pp. 1219–1226, 2003.