34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Biological pathway discovery through text mining and data integration

Chi Zhang, Rajesh Chowdhary, and Jinfeng Zhang*

Abstract— **Biological pathways are becoming increasingly important in our understanding of biological processes and discovering treatment for diseases. Constructing a pathway requires the knowledge of the set of proteins that are involved in the pathway. Much of this information is obtained through manual annotations of the literature. However, manual annotation of pathway related information is very time and resource consuming and can hardly catch up with the ever increasing publications in biomedical science. In addition, information often resides in different places making integrative analysis of and computations on such data more challenging. In this study, we integrate data from different sources, including manually annotated databases and literature. We further discover new pathway-protein associations that have not been documented in databases before using a knowledge discovery system, integrated bio-entity network, we proposed recently. Through manual verification of some discovered examples, we show that our method can effectively found new pathway-protein associations. The tool we developed in this study can be used to assist human annotations of pathway related information and also helpful for biologists who study certain pathways.**

I. Introduction

A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, turn genes on and off, or spur a cell to move. The molecules that make up biological pathways interact with signals, as well as with each other, to carry out their designated tasks. In recent years, researchers found that for many diseases, instead of targeting on one particular protein, it is more effective to target on one or a few pathways. Identifying what genes, proteins and other molecules are involved in a biological pathway can provide clues about what goes wrong when a disease strikes.

Taking cancer as an example, until recently, based on the observation of one type of leukemia[1] many had hoped that most types of cancers were driven by a single genetic error and could be treated by designing drugs to target those specific errors. Unfortunately, the one-target, one-drug approach has not held up for most other types of cancer. Recent projects that deciphered the genomes of cancer cells have found an array of different genetic mutations that can lead to the same cancer in different patients. This complex view can be simplified by looking at which biological pathways are disrupted by the genetic mutations. Rather than designing dozens of drugs to target dozens of mutations, drug developers could focus their attentions on just two or three biological pathways.

The pathway-centric approach requires scientists to understand the players involved in the pathways, such as proteins/genes and small molecules, and how they interact with one another. In the past, through laboratory studies of cultured cells and various organisms researchers have discovered many important biological pathways and the proteins involved in them. Such scattered information need to be put together for each pathway and this process is called pathway construction or building [2].

Pathway building has been performed by individual groups studying a network of interest as well as by large bioinformatics consortia (e.g., pathway interaction database [3], the Reactome Project [4] and KEGG database [5]) and commercial entities (e.g., Ingenuity Systems) through manually annotating of the literature. In recent years, automatic information extraction methods have also been developed to extract pathway related information [6, 7, 8, 9, 10, 11].

Most of current information extraction methods use co-occurrence of bio-entity names in abstracts to infer the relationship between bio-entities. Here, bio-entity has a broad definition and includes proteins/genes, small molecules, pathways, diseases and GO terms, which can all be related to a particular pathway. Co-occurrence approach has several drawbacks. Firstly, it can produce a large number of false positives especially for molecular interactions. Even using co-occurrences in the same sentence for information extraction, protein-protein interaction (PPI) extractions can suffer from large false positive rates [12]; Secondly, co-occurrence relies on counts of co-occurrences between two bio-entities to establish the statistical significance of the relationship. Some newly discovered relationships can have very low count, which will be missed by the method, although they may well be those a user wants to find out; Finally, an issue with co-occurrence and many other current methods is that they only extract PPI information, which is not linked to the corresponding pathways they are involved into. As a result, the current databases [3, 4, 5] document a rather small number of proteins related to their pathways.

In this study, we integrate information from several sources including manually annotated databases and literature. The databases include both databases for pathway

Chi Zhang is with the Department of computer science, Florida State University, Tallahassee, FL 32306 USA (e-mail: cz06@fsu.edu).

Rajesh Chowdhary is with Marshfield Clinic, Marshfield, WI 54449 USA. (e-mail: chowdhary.rajesh@mcrf.mfldclin.edu).

Jinfeng Zhang is with the Department of Statistics, Florida State University, Tallahassee, FL 32306 USA (corresponding author, phone: 850-228-3897, fax: 850-644-5271, e-mail: jinfeng@stat.fsu.edu).

related information and databases documenting PPIs. In addition, we extract PPIs from a large number of PubMed abstracts using a recently develop PPI extraction method [12, 13]. All the information is integrated to a structured form, called integrated bio-entity network (IBN) [14], which allows us to discover new relationships from the integrated information. We show with some examples that our method can effectively find proteins that have not been associated with pathways previously. A web interface was also built to allow users to find the proteins related to a pathway of their interest, including both known and predicted proteins.

II. METHODS

In this study, we integrate several methods we developed in the past and apply them to the problem of discovering new proteins associated with pathways. Here we briefly describe these methods. Interested readers can refer to our earlier publications for details [1-4].

A. Protein-protein interaction extraction from literature

Overview. Our protein-protein interaction (PPI) extraction method is based on a concept called triplets, which contains a word describing the interaction relationship (called interaction word) and two protein names in a sentence. Triplets can be extracted from sentences and then classified to be either true or false by a Bayesian network (BN) based machine learning method [1]. To tag the protein names in sentences, we used a protein name dictionary, which has now been extended to contain more than seven million protein names [4]. A dictionary of interaction words has also been built manually and extended recently [3]. The interaction word dictionary now contains more than 1000 words. To extract pathway related information we further built a pathway name dictionary using the names obtained from KEGG pathway database [5], Reactome [6], and pathway interaction database [7].

Features. To infer the classes of the triplets (as true or false), we manually selected the features that we believe are related to the language rules people use to describe PPIs. In the current method, we used 20 features [1, 3]. Each feature is assumed to capture information/signals associated with certain grammar or language rules that describe PPI relationships. In addition to the above features, we also added some new features based on part-of-speech tagging results using natural language processing (NLP) techniques in a more recent study [3]. For these features, the publicly available Stanford part-of-speech tagger [8] is used to tag the sentences.

An ensemble approach for PPI extraction. Through triplet features, we learn the language rules related to PPIs using three difference machine learning methods, Bayesian network (BN) [1], mixture of logistic models (ML) [3] and support vector machine (SVM). To construct the ensemble predictor, we fitted a logistic model from the predictions of individual methods. The cross-validation performance of the ensemble and individual methods on benchmark datasets showed that the three machine learning methods perform similarly to one another and the ensemble approach performs better than the individual methods overall [3].

B. Data integration

We collected data from several sources. Proteins related certain pathways are obtained from pathway interaction database [7] and Reactome database [6]. Totally, we have 269 pathways with 12651 associated proteins. For protein–protein interactions, BioGRID [9], EBI IntAct [10] and NCBI Gene database [11] are used, with 303,093 total interactions. We extracted 652,236 interactions from PubMed abstracts with an estimated number of 130,000 true cases [4]. The heterogeneous data is integrated into a structured form, called integrated bio-entity network (IBN) [4], where the nodes or vertices are bio-entities (including both proteins and pathways) and edges are their relationships. Using IBN, we can easily perform information retrieval. An integrated molecular interaction database (IMID) was built recently (integrativebiology.org) using the structured information in IBN [12]. We can also discovery new relationships that have not been reported in literature or documented in database before [4]. We designed algorithms for knowledge discovery through IBN. In this study, we modified one of them and applied it to discover new proteins associated with pathways (pathway-protein associations).

C. Knowledge discovery algorithms

The probabilities of the relationships between any two vertices that are not connected by an edge in IBN can be calculated using the probabilities of existing edges. Any edge, representing a relationship between two bio-entities, has a probability assigned to it. For relationships obtained from manually annotated databases, the probabilities are 1. For relationships extracted from literature, the probabilities are given by the extraction method. When multiple instances are extracted for one particular relationship (i.e. several mentions of the same interaction between two proteins) from the literature, the probability is calculated as $p = 1 - \prod(1 - p_i)$, where p_i is the probability of instance i, and each p_i is assumed to be independent to one another. We have designed two algorithms for automatic knowledge discovery using IBN, breadth-first search with pruning (BFSP) and most probable path (MPP) [4].

Breadth-first search with pruning (BFSP) algorithm. To search for all indirectly connected vertices from a given vertex we perform a modified breadth-first search (BFS) algorithm [13], breadth-first search with pruning (BFSP), starting from the vertex. Here we are only interested in vertices whose relationships to i have probabilities greater than a threshold value, p_c, or have a maximum of d_0 edges away from i. The additional pruning step aims to only include those significant relationships in the search result, which is essential in large scale knowledge discovery.

```
procedure BFSP(graph G, node i)
create a queue Q
enqueue vertex i onto Q
mark vertex i and set d_i = 0
while Q is not empty
```

```
dequeue a vertex v from Q
for each unmarked neighbour w of v
    if w is not marked
        d_w = d_v + 1
        p_{i,w} = p_{i,v} × p_{v,w} × p_d
```

/* $p_{i,w}$ is the probability for the relationship between node i and w, $p_{i,v}$ is the probability between node i and v, $p_{v,w}$ is the probability for node v and w, and p_d is a parameter to model the uncertainty when inferring relationships through indirect edges */

```
            if p_{i,w} > p_c or d_w ≤ d_0
```

/* p_c is the threshold for selecting more relevant relationships */

```
                mark w
                enqueue w onto Q
```

In the above procedure the probability p_d is used to model the uncertainty when inferring relationships through indirect edges. In principle, this probability can be learned from data and does not have to be a constant.

III. RESULTS

We performed BFSP for all the 269 pathways in our dataset by setting $d_0 = 2$, $p_d = 0.8$ and $p_c = 0.4$ (see BFSP procedure). The BFSP algorithm will not visit any vertices which are more than two edges apart from any of the pathways. There are totally 12651 manually annotated proteins directly associated with these pathways (47 proteins per pathway on average). In addition to these known associations, we have found 220,084 new pathway-protein associations (818 new proteins per pathway on average). If we only count manually annotated protein-protein interactions, there are 177,779 new pathway-protein associations with 661 new proteins per pathway on average.

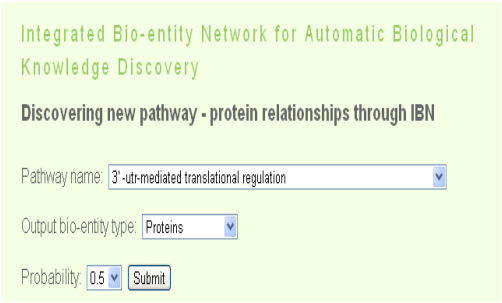

Figure 1. The interface of web server for pathway-related protein discovery.

Protein name	Pathway-protein association link	Molecule name	Molecular interaction link	Probability
ITGB7	link1			
ITGB7	link1	EED	link2	1
ITGB7	link1	FLNA	link2	1
ITGB7	link1	MADCAM1	link2	1
ITGB7	link1	FN1	link2	1
ITGA4	link1			
ITGA4	link1	EED	link2	1
ITGA4	link1	PXDN	link2	1
ITGA4	link1	CD81	link2	1
ITGA4	link1	PXDN	link2	1

Figure 2. Output table.

Taking insulin pathway as an example, protein TRB3, which is not annotated to be related to insulin pathway, is found to be related through AKT, whose association with insulin pathway is documented in database. The interaction between TRB3 and AKT is extracted from [24], where it is mentioned that TRB3 disrupts insulin signaling by binding to AKT. Another example is inhibitor kappaB kinase (IKK), which contributes to insulin resistance by phosphorylating protein IRS-1 [25], a protein that has been annotated to be associated with insulin pathway. It is worth noting that manual verification of the discovered associations is generally difficult. This is due to not only the difficulty of finding evidence from the vast amount of literature, but also the fact that even evidences cannot be found, we cannot rule out that the association can still be true.

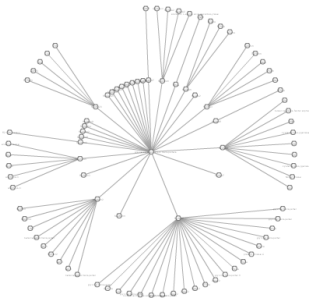

Figure 3. Network for amine compound SLC transporters pathway.

We implemented a publicly accessible web interface for users to perform pathway-related protein discovery at http://stat.fsu.edu/~jinfeng/IBN.html. The interface is simple with three optional dropdown boxes (Figure 1). In the first dropdown box, the user is expected to select the pathway he/she would like to search for. The second dropdown box is the type of molecules that will be included in the output. So far, we have only proteins implemented while working on small molecules. The third dropdown box is the probability cutoff for the PPIs that the user wants to include in the output. A probability of 1 means only manually annotated PPIs will be used in knowledge discovery.

The output information is given in a table as shown in Figure 2. The first column contains proteins manually annotated to be associated with the pathway, called directly-related proteins. Clicking those proteins will bring the user to the corresponding page of the protein on UniProt database [16]; The second column gives the links showing the direct associations between proteins and the pathway taken from Reactome or pathway interaction database; The third column lists the molecules associated with the pathway through the directly-related proteins and these proteins are called indirectly-related proteins. Again clicking them will bring the user to UniProt database; The fourth column contains

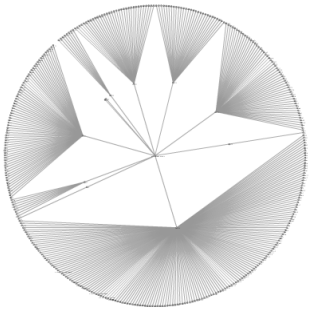

Figure 4. Network for ALK2 signaling events pathway.

the links showing the evidence of the interaction between the directly-related protein the indirectly-related protein. The links point to the corresponding PubMed abstracts, and the interaction may come directly from the abstract or from the text in the article; and the fifth column shows the probabilities of the interactions. Users can also plot the network associated with the pathway. An example is shown in Figure 3 for amine compound SLC transporters pathway drawn by Cytoscape [26]. The center of the network is the pathway. Proteins on the smaller circle are directly-related proteins. Proteins on the larger circle are the indirectly-related proteins that are discovered through directly-related proteins. Figure 4 is the network for the pathway, ALK2 signaling events. This network is different from the one in Figure 3 in that there are many more indirectly-related proteins than the directly-related proteins, indicating that those proteins in ALK2 signaling events pathway are studied much more than those proteins in amine compound SLC transporters pathway.

IV. CONCLUSION AND DISCUSSION

In this study, we applied a recently developed knowledge discovery system, integrated bio-entity network (IBN), to discovery new pathway-protein associations. We found that the approach is able to find a large number of new associations and some of the discovered associations are manually verified to be true relationships. Manual verification of the data we generated in this study may add a significant number of new pathway-protein associations to our current knowledge base.

Although we have found a large number of pathway-protein associations that have not been documented in databases before, it is likely many of them are false. We discuss several future directions to further improve the accuracy of the knowledge discovery. Firstly, there is still a considerable room for improvement in PPI extraction [14-25]. This is an area where even moderate improvement will see clear benefit in the downstream knowledge discovery; Secondly, extraction of the interaction words associated with the PPIs and the direction of the interaction can be very useful in determining whether the proteins are actually related to the pathway [19, 26]. For example, those proteins that regulate or affect those directly-related proteins are more likely to be related to the pathway than those proteins that are regulated or affected by the directly-related proteins; Thirdly, other extraction methods such as those extract protein function information or protein-disease, protein-pathway information [27-29] can be combined with the current framework to enhance the accuracy of the discovery. A particular relationship is more likely to be true if it is supported by multiple sources of information.

REFERENCES

[1] Chowdhary, R., J. Zhang, and J.S. Liu, *Bayesian inference of protein-protein interactions from biological literature.* Bioinformatics, 2009. **25**(12): p. 1536-42.

[2] G. A. Viswanathan, J. Seto, S. Patil, G. Nudelman, and S. C. Sealfon, "Getting started in biological pathway construction and analysis," *PLoS Comput Biol,* vol. 4, p. e16, Feb 2008.

[3] Bell, L., J. Zhang, and X. Niu, *Mixture of logistic models and an ensemble approach for extracting protein-protein interactions.* ACM-BCB, 2011: p. 371-375.

[4] Bell, L., et al., *Integrated bio-entity network: a system for biological knowledge discovery.* PLoS One, 2011. **6**(6): p. e21474.

[5] Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101.

[6] Vastrik, I., et al., *Reactome: a knowledge base of biologic pathways and processes.* Genome Biol, 2007. **8**(3): p. R39.

[7] Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D674-9.

[8] Tarcea, V.G., et al., *Michigan molecular interactions r2: from interacting proteins to pathways.* Nucleic Acids Res, 2009. **37**(Database issue): p. D642-6.

[9] Stark, C., et al., *BioGRID: a general repository for interaction datasets.* Nucleic Acids Res, 2006. **34**(Database issue): p. D535-9.

[10] Aranda, B., et al., *The IntAct molecular interaction database in 2010.* Nucleic Acids Res, 2010. **38**(Database issue): p. D525-31.

[11] Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2007. **35**(Database issue): p. D26-31.

[12] Balaji, S., et al., *IMID: integrated molecular interaction database.* Bioinformatics, 2012. **28**(5): p. 747-9.

[13] Cormen, T.H., et al., *Introcution to algorithms*. 2009: The MIT Press.

[14] Huang, M., et al., *Mining physical protein-protein interactions from the literature.* Genome Biol, 2008. **9 Suppl 2**: p. S12.

[15] Tikk, D., et al., *A Comprehensive Benchmark of Kernel Methods to Extract Protein-Protein Interactions from Literature.* PLoS Computational Biology, 2010. **6**(7): p. e1000837.

[16] Saetre, R., et al., *Extracting Protein Interactions from Text with the Unified AkaneRE Event Extraction System.* Ieee-Acm Transactions on Computational Biology and Bioinformatics, 2010. **7**(3): p. 442-453.

[17] Bui, Q.C., S. Katrenko, and P.M. Sloot, *A hybrid approach to extract protein-protein interactions.* Bioinformatics, 2010.

[18] Bjorne, J., et al., *Complex event extraction at PubMed scale.* Bioinformatics, 2010. **26**(12): p. i382-i390.

[19] Giles, C.B. and J.D. Wren, *Large-scale directional relationship extraction and resolution.* BMC Bioinformatics, 2008. **9 Suppl 9**: p. S11.

[20] Devignes, M., et al., *BioRegistry : automatic extraction of metadata for biological database retrieval and discovery.* International Journal on Metadata, Semantics and Ontologies, 2010. **5**: p. 184-189.

[21] Krallinger, M., et al., *Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge.* Genome Biol, 2008. **9 Suppl 2**: p. S1.

[22] Kim, S., J. Yoon, and J. Yang, *Kernel approaches for genic interaction extraction.* Bioinformatics, 2008. **24**(1): p. 118-26.

[23] Kim, S., et al., *PIE: an online prediction system for protein-protein interactions from text.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W411-5.

[24] Sanchez-Graillet, O. and M. Poesio, *Negation of protein-protein interactions: analysis and extraction.* Bioinformatics, 2007. **23**(13): p. i424-32.

[25] Rodriguez-Penagos, C., et al., *Automatic reconstruction of a bacterial regulatory network using Natural Language Processing.* BMC Bioinformatics, 2007. **8**: p. 293.

[26] Wren, J.D., et al., *Knowledge discovery by automated identification and ranking of implicit relationships.* Bioinformatics, 2004. **20**(3): p. 389-98.

[27] Yilmaz, S., et al., *Gene-disease relationship discovery based on model-driven data integration and database view definition.* Bioinformatics, 2009. **25**(2): p. 230-6.

[28] Mottaz, A., et al., *Mapping proteins to disease terminologies: from UniProt to MeSH.* BMC Bioinformatics, 2008. **9 Suppl 5**: p. S3.

[29] Koike, A., Y. Niwa, and T. Takagi, *Automatic extraction of gene/protein biological functions from biomedical text.* Bioinformatics, 2005. **21**(7): p. 1227-36.