A Language Modeling Text Mining Approach to the Annotation of Protein Community

Xiaodan Zhang, Daniel D. Wu, Xiaohua Zhou, and Xiaohua Hu
{Xiaodan.zhang, daniel.wu, Xiaohua.zhou}@drexel.edu, thu@ischool.drexel.edu
College of Information Science and Technology
Drexel University
3141 Chestnut, Philadelphia, PA 19104

***Abstract*-This paper discusses an ontology based language modeling text mining approach to the annotation of protein community. Communities appear to play an important role in the functional properties of complex networks. Being able to annotate the identified the community structure in a biological network can help us to understand better the structure and dynamics of biological systems. Traditional method such as Gene Ontology (GO) provides information about the functionality of gene products, but they are not enough to annotate community as for only limited number of proteins in the database, limited protein properties available for annotation and the inability to annotate a group of gene products as a whole. Thus, we present an ontology based mixture language model approach to annotate protein community. Compared to traditional method, we have the following three advantages. First, biomedical literature mining brings much richer information than existed gene databases. Second, the mixture language model can help "purify" the document by eliminating some background noise. Third, using domain ontology, we extract biological concept and concept pairs from abstracts. Biological concept is more meaningful than word or multi-word phrases. Moreover, using concept pairs can deliver much more information and serve as evidence of annotation results. We test our approach on four communities SAGA-SRB, CCR-NOT, RFC and ARP2/3, detected from dataset of interactions for Saccharomyces cerevisae from the General Repository for Interaction Datasets (GRID). Annotation results provide a very coherent indication of functionality of each community.**

I. INTRODUCTION

Proteins are important players in executing the genetic program. When carrying out a particular biological function, or serving as molecular building blocks for a particular cellular structure, proteins rarely act alone. Rather, biological complexity is encapsulated in the structure and dynamics of the combinatorial interactions among proteins (as well as other biological molecules) at different levels, ranging from biochemical pathways to ecological phenomena [1]. Therefore, one of the key challenges in this post genomic era is to understand these complex molecular interactions that confer the structure and dynamics of a living cell.

Community structure is an important property common to many networks. Although there is no formal definition for the community structure in a network, it often loosely refers to the gathering of vertices into groups such that the connections within groups are denser than between groups [2]. The study of community structure in a network is not new. It is closely related to the graph partitioning in graph theory and computer science and the hierarchical clustering in sociology [3]. Recent years have witnessed an intensive activity in this field partly due to the dramatic increase in the scale of networks being studied. Many algorithms [4-12] for finding communities in networks have been proposed. They can be roughly classified into two categories, divisive and agglomerative. The divisive approach takes the route of recursive removal of vertices (or edges) until the network is separated into its components or communities, whereas the agglomerative approach starts with isolated individual vertices and joins together small communities.

In the context of biological networks, communities might represent structural or functional groupings, and can be synonymous with molecular modules, biochemical pathways, gene clusters, or protein complex. Because communities are believed to play a central role in the functional properties of complex networks [3], the ability to detect and annotate communities in networks could have practical applications. Especially growing community from a given seed makes it much more practical value for biologists when they have knowledge of a certain protein and wish to discover and study its community. Thus, we detect and annotate community grown from a given seed protein for this purpose. In this paper, our algorithm verified through MIPS, has been proved to be an efficient approach to detect "good" community.

Being able to annotate the identified the community structure in a biological network can help us to understand better the structure and dynamics of biological systems. Community annotation, i.e. the evaluation and annotation of identified communities, is especially critical in the study of biological networks, because not only it's important to understand the biological networks but also the identified communities must be biological relevant for them to be useful.

The task of annotating a protein community is in essence to summarize the functionality concepts shared by most of community members. The ideal situation is that we know functionality of every protein in a community and then extract protein property information common to the community. However, it's a very challenging task, especially when to realize it automatically, as it's dependent on many factors such as the property information available for each protein, the technique of information extraction and summarization, and the quality of detected community. Although conventional methods such as using GO [15], MIPS[16] functional and complex catalogue may help identify functionality of group members, they have many limitations. Existed databases such as GO only contain information of limited number of proteins.

Although they are growing through intensive manual work based on published experiment results, it's hard to catch up with not only the high-through put experiments but also the outgrowing number of published experiment results. Besides, many of functional explanations might not have been verified. There are many proteins whose information can not be found in these databases but may exist in published literature. Moreover, it's controlled vocabulary and only allows annotating genes and their products with only a limited set of attributes [15]. Thus, biologists often need to find information about genes whose function is not described in the genome databases [30].

Therefore, a tool to automatically extract functional information for a community from biomedical literature becomes very necessary. However, to accomplish the task, two problems should be solved: how to represent a document; how to mine topical terms from documents. There are many approaches that have been developed to overcome the difficulties. Most existed works relies on statistical based methods and heuristics to summarize a gene or group of genes using key terms or sentences. Hu [31] uses information gained based method to extract key phrases from PubMed abstracts and then apply mutual reinforcement principle to generate key phrases and sentences for the given gene cluster. The problem with the method is that pure information gain based approach to extract phrases from biomedical literature appears to be weak to solve synonym and polysemy problems in biomedical concepts. Popescu and colleagues [32] addresses the problem of constructing a functional summarization of groups of gene products that are found by clustering a database of such products annotated by GO. The method builds the "most representative term" for each group of gene products using a fuzzy similarity measure to find highest frequency in the description of the gene products. However, the work is purely dependent on GO database, thus it has no use for proteins not presenting in the database. MedMesh summarizer [33] treats each gene related documents as a category and utilizes statistical methods to extract topical terms from Mesh terms across the collection of groups of genes. The problem of this approach is that it relies on MeSH terms that may not enough to serve the function annotation for given cluster and is based on only heuristic of combination of different statistical methods. There are also works generating summary for a target gene. Ling and colleagues [30] tries to summarize a gene through first retrieving relevant articles and then extracting the most informative sentences from the retrieved articles to generate a structured gene summary. They develop a special tokenizer for gene products extraction. Sentences (that contain certain genes) are scored according to their category relevance score, document relevance score and location score. The limitations with the work are its dependence on the high-quality data in FlyBase and too many heuristic methods.

For most of existed works, less attention has been paid to terms extraction, which is very essential for biomedical literature mining. Biomedical named entity recognition is a very challenging task. For example, it's very difficult to recognize gene names from biomedical literature as for the following reasons: newly defined genes, long descriptive gene names, synonyms, and lexical variations. Traditional machine learning method can be easily over fitted. To handle this problem, the best way is to utilize domain knowledge to help extract biomedical named entity. Thus, how to adapt domain ontology to existed information extraction method and summarization technique would be very beneficial.

Hence, we initially query PubMed using proteins and their aliases of a community, then apply a dictionary based concept extraction module MaxMatcher trained from Universal Medical Language System (UMLS) [34] to extract biomedical concepts from abstracts for a community, and then develop a mixture language model text mining method to automatically extract key concepts or concept pairs from multi-document collections to form a sensible biological explanation to the identified communities. MaxMatcher is designed for extracting biomedical concepts instead of multi-word phrases or individual words, because concept is more representative and powerful than multi-word phrases or individual words. For example, C0020538 is a concept about the symptom of hypertension in UMLS; it represents a set of synonymous terms including high blood pressure, hypertension, and hypertensive disease. In comparison with individual words, a concept is more meaningful; in comparison with multi-word phrases, a concept well solves polysemy and synonymy problems.

Mixture language model [35, 36, 37] has been well studied and applied in information retrieval which is proved to be a solid method for giving consistently higher precision and recall. It can automatically interpolate between the model that generates topical concepts and the background model that generate concepts common to the whole collection, which helps to remove background noise for summarization process.

Compared to other methods, the advantages of our approach many folds: first, domain ontology trained MaxMatcher can target biomedical concepts more precisely; second, the mixture language model with ontology support is more suitable for biomedical domain than methods based on unigram models; third, biomedical concept is more meaning full than word and multi-word phrase and using concept pairs can deliver much more information and serve as evidence of annotation results; fourth, to annotate the community as a whole would provide chances for understanding the biological meaning of some proteins within the community whose functions are still unknown because proteins usually share functions with those proteins which interact with them mostly; last, our approach of annotating protein community using textual data (PubMed) can serve as an extension to traditional method of annotating a single protein such as GO.

II. Detection of the Community from a Seed Protein

In this section, we briefly introduce our protein community detection algorithm, referred to as CommBuilder, which is a graph-based detection algorithm, previously developed by our research group [13, 14].

Due to the complexity and modularity of biological networks, it is more feasible computationally to study a

community containing one or a few proteins of interest. A protein-protein interaction network is modeled as a simple graph. Each vertex of the graph represents a protein and each edge represents an interaction between the two proteins connected by it. An undirected graph, G = (V, E), is comprised of two sets, vertices V and edges E. An edge e is defined as a pair of vertices (u, v) denoting the direct connection between vertices u and v. The graphs we use in this paper are undirected, unweighted, and simple -meaning no self-loops or parallel edges.

For a subgraph $G' \subset G$ and a vertex i belonging to G', we define the in-community degree for vertex i, $K_i^{in}(G')$, to be the number of edges connecting vertex i to other vertices belonging to G' and the out-community degree, $K_i^{out}(G')$, to be the number of edges connecting vertex i to other vertices that are in G' but do not belong to G. CommBuilder adopted the quantitative definitions of community defined by Radicchi and colleagues [22]. In this definition, a subgraph G' is a community in a strong sense if for each vertex i in G', its in-community degree is greater than out-community degree. More formally, G' is a community in a strong sense if

$$K_i^{in}(G') > K_i^{out}(G'), \forall i \in G', G' \subset G \quad (1)$$

In a weak sense if the sum of all degrees within G' is greater than the sum of all degrees from G' to the rest of the graph, i.e., G' is a community in a weak sense if

$$\sum_i K_i^{in}(G') > \sum_i K_i^{out}(G'), i \in G', G' \subset G \quad (2)$$

CommBuilder accepts a seed protein s, and then gets its neighbors, finds the core of the community to build, and finally expands the core to obtain the eventual community. The two major components of CommBuilder are *FindCore* and *ExpandCore*. In fact, *FindCore* performs a naïve search for maximum clique from the neighborhood of the seed protein by recursively removing vertices with the lowest in-community degree until all vertices in the core set have the same in-community degree.

The algorithm performs a breadth first expansion in the core expanding step. It first builds a candidate set containing the core and all vertices adjacent to each vertex in the core. It then adds to the core a vertex that either meets the quantitative definition of community in a strong sense or the fraction of in-community degree over a relaxed affinity threshold f of the size of the core. The affinity threshold is 1 when the candidate vertex connects to each of vertices in the core set. This threshold provides flexibility when expanding the core, because it is too strict requiring every expanding vertex to be a strong sense community member. For details of algorithm, please refer to our previous works [13, 14].

Once we identify a protein community from the interaction network, we compare the community membership with the MIPS catalogues to evaluate the identified community, with details deferred to the Experimental Results section. In the following section, we propose a biomedical literature mining approach using mixture language models for the annotation of a given protein community.

III. Protein Community Annotation through Biomedical Literature Mining

Annotation of a protein community using ontology (e.g., GO) or biomedical literature is in essence a technique of summarization of properties of community group members. The ideal situation is that we check the functionality of each protein in the community through GO or MIPS and then extract functional concepts common to most group members to serve as the annotation of the community. However, this is barely practical because of the limitation of these databases. Let's take GO as an example. The GO project has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. However, GO allows us to annotate genes and their products with only a limited set of attribute [15]. For example, GO does not allow us to describe genes in terms of which cells or tissues they're expressed in, which developmental stages they're expressed at, or their involvement in disease, while these information may be available in the according biomedical literature and can be really important to biological scientists. Moreover, GO uses a controlled vocabulary which may not be enough to describe the properties of proteins. The worst of all is that there are many proteins that have not been annotated because manually annotating proteins can hardly catch up with the outgrowing gene products related biomedical literature. It can be very normal that there are several proteins in a community that have not been annotated in the database but appeared in the literature. These proteins may affect the annotation of whole group.

It should be noted that there may be some proteins in a community that are not indexed in both literature and gene databases (GO). Although our summarization approach can provide guidance to predict functionality of those unknown proteins, if most of community members are not indexed by literature, it will be beyond our approach, because our focus is how to apply summarization technique to annotate a group of proteins using according biomedical literature. However, it can be our future work to annotate a community by integrating functional prediction information using other resources or methods for those unknown proteins.

While using literature to annotate a community has a lot of advantages over existed gene databases, it also brings background noise. For a document or a document set, terms of interest are those topical and informative terms. For those terms that are very common according to the whole collection or database are of not interested. One simple example of background noise is stop word. Even though counting on some statistical methods such as tf*idf, z-score, the combination of mean, stand deviation information and so on may help remove some of background noise, but it's not a generative model and involves a lot heuristics. Therefore, we introduce a generative mixture language model to solve this problem.

Language modeling is a technique initially used for speech recognition. Ponte and Croft introduced language modeling approach to text retrieval in [35]. The relative simplicity and effectiveness of the language modeling approach, together with the fact that it leverages statistical methods that have been developed in speech recognition and other areas, make it an attractive framework for not only text retrieval but also theme detection across collections by Zhai [37]. By interpolating with a collection model, each unigram document model can reduce the effect of the background noise. The mixture language model (i.e.,the mixture of the unigram document model and the collection model) has demonstrated its effectiveness in information retrieval [36] and shown the potential on multi-document summarization [37].

However, unigram language model without domain ontology support has limitation in biomedical domain. For example, as for protein functionality, with unigram language model, it's very difficult to differentiate functionality concepts from other concepts. In this paper, we fit UMLS domain ontology in Zhai's mixture language model [36]. In this way, we can detect most common functionality concepts shared by each community, while filtering out most of non-related concepts.

A. The Mixture Language Model

A mixture language model proposed in [36] is used to generate topical concepts for each protein community. Suppose we have built a collection of document denoted as *D* for a given community (see Section 3.2 for details). We then assume each concept in the collection *D* is generated either by the community theme model θ_c or the background model θ_b. That is, where α is a coefficient accounting for the background noise

$$p(w \mid D) = (1-\alpha)p(w \mid \theta_c) + \alpha p(w \mid \theta_b) \quad (3)$$

Under this simple mixture model, the log-likelihood of generating the collection *D* is

$$\log p(D \mid \theta_c, \theta_b) = \sum_w c(w, D) \log((1-\alpha)p(w \mid \theta_c) + \alpha p(w \mid \theta_b)) \quad (4)$$

The protein community theme model can be estimated using the Expectation Maximization (EM) algorithm with the following update formulas:

$$\hat{p}^{(n)}(w) = \frac{(1-\alpha)p^{(n)}(w \mid \theta_c)}{(1-\alpha)p^{(n)}(w \mid \theta_c) + \alpha p(w \mid \theta_b)} \quad (5)$$

$$p^{(n+1)}(w \mid \theta_c) = \frac{c(w, D)\hat{p}^{(n)}(w)}{\sum_i c(w_i, D)\hat{p}^{(n)}(w_i)} \quad (6)$$

where $c(w, D)$ is the frequency count of concept *w* in *D*. The choice of the background model will affect the estimation of the community theme model. The concepts generated by the background model are often those frequently occurring in the background collection. Therefore, if we choose the collection D itself as the background collection, many topical concepts of the protein community will be falsely treated as background concepts and excluded from the community theme model. To overcome this limitation, we randomly download ten percentages of Medline abstracts published in 2005 as the background collection referred to as *C*. Using the maximum likelihood estimate, we obtain the background model as follows:

$$p(w \mid \theta_B) = \frac{c(w, C)}{\sum_{w'} c(w', C)} \quad (7)$$

where $c(w, C)$ is the frequency count of concept *w* in *C*.

B. Document Retrieval and Concept Extraction

We manually collected the alias of each protein in the community and submit it together with the protein name (see table 1) to the PubMed search engine. In this way, we collected Medline abstracts for each protein community. Here we treat documents for each protein community as a separate collection.

The dictionary-based approach is frequently used to extract protein function concepts from biomedical literatures. Its major advantage over the feature-based approach is that it not only recognizes concept names, but also identifies unique concept identities, which is very helpful to solve synonym and polysemy problems and thus to improve summarization accuracy. However, the dictionary-based approach is critiqued by its low extraction recall due to the variation of biological concept names. As a result, we develop MaxMatcher [39], a dictionary-based biological concept extraction system, which well handles the term variations using approximate dictionary lookup, and achieves good extraction recall. The basic idea of this approach is to capture the significant words instead of all words to a particular concept. MaxMatcher uses UMLS as the dictionary and extracts biological concepts of 135 semantic types.

In particular, we propose a relative significance score measure. Suppose a concept (*c*) has *n* concept names denoted as $s_1, \ldots, s_n$, respectively. Let *N(w)* denotes the number of concepts whose variant names contain word *w*, and let w_{ji} denotes the *i*-th word in the *j*-th variant name of the concept, the significance of *w* to the concept is defined as follows:

$$I(w, c) = \max\{ I(w, s_j) \mid j \le n\} \quad (8)$$

where :

$$I(w, s_j) = \begin{cases} 0 & w \notin s_j \\ \dfrac{1/N(w)}{\sum_i 1/N(w_{ji})} & w \in s_j \end{cases}$$

We use UMLS Metathesaurus 2005AA version [34] as the dictionary to train the significance score of each word to biological concepts containing that word. For details, please refer to our previous work [39] [41].

Definition 1 A ***concept (w)*** is a unique meaning in a domain. It represents a set of synonymous terms in the domain. For example, *C0020538* is a concept about the disease of hypertension in UMLS Metathesaurus; it also represents a set of synonymous terms including *high blood pressure*, *hypertension*, and *hypertensive disease*. Therefore, concept-based term extraction helps to relieve the synonym and polysemy problems in biomedical literature, where a term (e.g., a gene or a protein) might have many synonyms while also representing different concepts in different context [41].

In order to validate the results of concept-based annotation, we also represent documents as a set of concept pairs and run a concept-pair language model to find out topical concept pairs

for each protein community. Because of the limit of the precision of the state-of-the-art retrieval approaches, a good part of the retrieved documents may not be relevant to the identified protein community. Thus, the top-ranked topical concepts in the community theme model may be the functional descriptions of proteins outside the given community. For this reason, if we do find certain topical concept has a strong relationship with proteins in the given community in natural language statement, we will be more confident on the annotation result.

Definition 2 A ***concept pair (t)*** is defined with two order-free components as in $t(w_i, w_j)$, where w_i and w_j are two concepts related to each other syntactically and semantically. The implementation of the syntactic and semantic relationships between two concepts is determined by specific applications [40].

A pair of two concepts will be extracted if they meet the following three requirements: (1) they appear in the same clause of an English sentence; and (2) their semantic types are compatible according to the domain ontology. For example, two proteins could be semantically compatible in UMLS (e.g., protein-protein interaction).

Example: *A recent epidemiological study (C0002783, research activity) revealed that obesity (C0028754, disease) is an independent risk factor for periodontal disease (C0031090, disease).*

Concept Index: *C0002783, C0028754, C0031090*

Topic Signature Index: *(C0028754, C0031090)*

In the above example, the underlined phrases are extracted concept names followed by the corresponding concept ID and semantic type. Obesity and periodontal disease is treated as a concept pair while the concept epidemiological study has no relationships with other concepts because it is in a separate clause.

IV. Experiment Results

In this section, we evaluated both the protein community detection algorithm (i.e. CommBuilder) and the community annotation approach. Although the detection algorithm is not the focus of this paper, we still evaluate its results for the two following reasons. First, the annotation result depends on the quality of the detected community; if the detected community itself is not coherent, the resulting annotation will not make sense. Second, to understand the functional properties of the detected community will help us judge the quality of the annotation.

A. Document Retrieval and Concept Extraction

We downloaded a dataset of interactions for Saccharomyces cerevisae from the General Repository for Interaction Datasets (GRID) [17]. The GRID database contains all published large-scale interaction datasets as well as available curated interactions such as those deposited in BIND [18] and MIPS [16]. The yeast dataset we downloaded has 4,907 proteins and 17,598 interactions. Four communities are identified from the interaction network by our detection program using one protein as seed. The members of each community are listed in table 1. The seed protein is selected randomly from the "core" protein set.

TABLE I
THE FOUR PROTEIN COMMUNITIES DETECTED BY COMMBUILDER. THE ALIAS FOR EACH PROTEIN IS MANUALLY COLLECTED OFFLINE. AS FOR SPACE, WE ONLY LIST THE ALIAS NAME OF PROTEIN.

SAGA/SRB		CCR4-NOT		RFC	ARP2/ARP3
ADA2*	SRB5†	CCR4*†	UBR1	POL12†	SKT5†
CSE2†	SRB6†	POL12†	NOT3*†	MAP2	MNN2
GCN5*	SRB7†	PKC1	MOB1*	RFC5*†	CHS3†
HFI1*	SRB8	CDC39*†	PRI1†	MRC1†	ARC40*†
MED11†	SSN2	COP1†	YAK1	RAD24†	RVS161†
MED2†	SSN3	DHH1*	CBF1	POL32†	ARP2*†
MED4†	TAF1	CDC36*†	CAF17*	RFC2*†	RVS167†
MED6†	TAF10*	TFP1†	PRI2†	RAD27†	SAP155†
MED7†	TAF11	CCT6†	STM1	CSM3†	SLT2†
MED8†	TAF12*	RVB1	CSI1	CTF18†	ARC15*†
NGG1*	TAF13	YAP6	POL1†	POL1†	BCK1†
PGD1†	TAF2	ARH1	CAF40†	TOF1†	ARP3*
ROX3†	TAF3	SRB4	POP2*†	RFC3*†	ARC19*†
SGF29*	TAF5*	MOT2*†	HRT1	RFC4*†	YLR111w
SPT15	TAF6*	SEC27†	STI1	ELG1†	ARC18*
SPT20*	TAF7	MPT5	STD1	RFC1*†	END3†
SPT3*	TAF8	GCN1†	TFC7†	CTF4†	SLA2†
SPT7*	TAF9*	PIL1	RET3		CLA4†
SRB2†	TRA1*	DBF2*	RVB2†		ARC35*
SRB4†		CAF130†	NOT5*†		BRO1

For the SAGA/SRB community: Proteins that belong to SAGA complex listed in MIPS complex catalogue database are indicated by (*) and those belonging to SRB complex are indicated by (†). For CCR4-NOT, Proteins belonging to CCR4-NOT complex listed in MIPS are indicated by (*) and proteins considered to be involved in transcription and DNA/chromatin structure maintenance are indicate by (†). For The RFC community. proteins belonging to RFC complex listed in MIPS are indicated by (*) and proteins listed in the functional category of DNA recombination and DNA repair or cell cycle checkpoints by MIPS are indicated by (†). For The ARP2/ARP3 community. Proteins belonging to ARP2/3 complex listed in MIPS are indicated by (*) and proteins listed in the functional category of budding, cell polarity, and filament formation by MIPS are indicated by (†).

As for better evaluate our annotation result, we would like to discuss the properties of each community using MIPS in detail.

The first community is identified using TAF6 as seed. TAF6 is a component of the SAGA complex which is a multifunctional co-activator that regulates transcription by RNA polymerase II [23]. The SAGA complex is listed in MIPS complex catalogue as a known cellular complex consisting of 16 proteins. As shown in Table I, the community identified by our algorithm contains 39 members, including 14 of the 16 SAGA complex proteins listed in MIPS (indicated by an asterisk in the Alias column). The community also contains 14 of 21 proteins listed in MIPS as Kornberg's mediator (SRB) complex. The rest of the proteins in the community are either TATA-binding proteins or transcription factor IID (TFIID) subunits or SRB related. TFIID is a complex involved in initiation of RNA polymerase II transcription. SAGA and TFIID are structurally and functionally correlated, make overlapping contributions to the expression of RNA

polymerase II transcribed genes [23]. SRB complex is a mediator that conveys regulatory signals from DNA-binding transcription factors to RNA polymerase II [24]. In addition, 27 of the top 50 potential co-complex proteins (9 of the top 10), not including the seed proteins, predicted by our module are in the identified community.

However, the most striking commonality among proteins within this community is revealed by the MIPS functional category each of these proteins belongs to. With only one exception (YCL010c), they all belong to functional category 11.02.03 (mRNA synthesis). More specifically, most of them are either in the category of 11.02.03.01 (general transcription activities) or 11.02.03.04 (transcriptional control). Therefore, we may annotate the first identified community with the MIPS functional category 11.02.03 (mRNA synthesis).

The second community is discovered using NOT3 as seed. NOT3 is a known component protein of the CCR4-NOT complex which is a global regulator of gene expression and involved in such functions as transcription regulation and DNA damage responses. MIPS complex catalogue lists 5 proteins for NOT complex and 13 proteins (including the 5 NOT complex proteins) for CCR4 complex. The NOT community identified is composed of 40 members. All 5 NOT complex proteins listed in MIPS and 11 of the 13 CCR4 complex proteins are members of the community. POL1, POL2, PRI1, and PRI2 are members of the DNA polymerase alpha (I) – primase complex, as listed in MIPS. RVB1, PIL1, UBR1, and STI1 have been grouped together with CCR4, CDC39, CDC36, and POP2 by systematic analysis [25]. The community also contains 20 out of 26 proteins of a complex that probably is involved in transcription and DNA/chromatin structure maintenance [26]. Not surprisingly, most of the proteins in this community are in the MIPS category of 10.01 (DNA processing), 10.03 (cell cycle), and/or 11.02 (RNA synthesis).

The third community is identified by using RFC2 as the seed (Table I). RFC2 is a component of the RFC (replication factor C) complex, the "clamp loader", which plays an essential role in DNA replication and DNA repair. The community identified by our algorithm has 17 members. All five proteins of RFC complex listed in MIPS complex catalogue database are members of this community, as shown in Table 1. This community also includes the top 8 ranked proteins predicted by our module. All but one (YBL091c) protein in this community belongs to the MIPS functional category of 10.01 (DNA processing), with majority of the proteins in the category of 10.01.03 (DNA synthesis and replication).

We use ARP3 as seed to identify the last community (figure 2). ARP2/ARP3 complex acts as multi-functional organizer of actin filaments. The assembly and maintenance of many actin-based cellular structures likely depend on functioning ARP2/ARP3 complex [27]. The identified community contains all 7 proteins of the ARP2/ARP3 complex listed in MIPS (Table I). Not including the seed (ARP3), these proteins represent the top 6 ranked proteins predicted by our module. Out of the 20 proteins in this community, one is in the MIPS functionally unclassified category (99), 14 in the category of budding, cell polarity and filament formation (43.01.03.05), one in cytoskeleton-dependent transport (20.09.14), one in vacuolar transport (20.09.13), and 15 in the category of cell wall (42.01) and/or cytoskeleton (42.04).

B. Evaluation of the Annotation Approach

The annotation of a community is through the following text mining procedure.

1) Relevant documents are retrieved from PubMed for each protein community. See the details of the retrieval in Section 3.2

2) For each community, we use MaxMatcher to extract biological concepts and concept pairs from retrieved abstracts and then index concepts and concept pairs separately.

3) The mixture language model algorithm is then applied to extract topical concepts and topical concept pairs of a protein community. We find setting background coefficient to 0.9 makes the best biomedical sense for protein community annotation. And this is reasonable because most terms are generated according the whole PubMed database collection;

4) UMLs semantic types (including Biologic Function, Physiologic Function, Cell Function, Molecular Function, Genetic Function, Pathologic Function, Cell or Molecular Dysfunction, Chemical Viewed Functionally, Functional Concept) are used to filter out those non functionality concepts. In UMLs ontology, each term has a unique concept ID and each concept ID has one or several semantic types. Based on this, concepts or pairs without functionality semantic types are filtered out in the last step; In particular, for concept pair, we require one is functionality concept, the other is gene or protein name(with semantic type as "Amino Acid, Peptide, or Protein" or "Gene or Genome"). This helps target protein's functionality.

5) Finally, the top k concepts or concept pairs ranked based on the mixture language model probability are chosen to serve as the functional annotation for the corresponding protein community. In practice, top 15 concepts and concept pairs are selected for each protein community (see table III).

TABLE II
COMMUNITY AND ITS DOWNLOADED DOCUMENT SETS

Community	# of document
SAGA/SRB	4064
CCR4-NOT	5885
RFC	2313
ARP2/ARP3	827

We obtained lists of top 25 functional concepts and concept pairs for each of the identified communities (Table III). These concepts and pairs provide a very coherent indication of functionality of each community. For example, the top concepts and pairs for the SAGA-SRB community are predominantly related to transcription initiation (activation); the top concepts and pairs for the RFC community are mainly about cell cycle and DNA damage. One main disadvantage of summarization is that it's hard to evaluate result. However, our concept pair based annotation not only performs well as functional annotation, but also serves as evidence to verify concept based annotation result. For example, for community ARP2/ARP3, the concept pair based annotation successfully targets the functionality of ARP2 protein (the functional

information of SLT2 and SLA2 are also extracted), such as regulation of actin polymerization, actin nucleation, and actin monomer binding, which is very consist with the community function (see table I). These pair information can be very in some way helps verify our concept based experiment results. For example concepts such as regulation of actin, and actin nucleation are verified as functional annotation concepts by APR2 protein functional concept pairs in table 1. So, it can serve a reasonable evidence to evaluate the concept based annotation results, automatically making the black box summarization technique more transparent.

TABLE III:
THE ANNOTATION OF THE FOUR DETECTED PROTEIN COMMUNITY. WE USE ";" TO SEPARATE CONCEPTS OR CONCEPT PAIRS. FOR CONCEPT PAIRS, WE USE "/" TO SEPARATE TWO CONCEPTS. FOR CONCEPTS, THE SEMANTIC TYPE HAS BEEN CHECKED WHETHER IT'S FUNCTIONALITY CONCEPT. FOR CONCEPT PAIR, ONE IS FUNCTIONAL CONCEPT AND THE OTHER IS PROTEIN. CONCEPTS AND CONCEPT PAIRS ARE RANKED ACCORDING TO THE GENERATIVE PROBABILITY OF THE COMMUNITY THEME MODEL

Top 15 topical concepts	**Top 15 topical concept pairs**
SAGA/SRB Community	
Transcription; Transcription Activation; protein protein interaction; Repeat; formation of translation preinitiation complex; transcription initiation; Transactivation; mutant; cell assembly; Nucleotide Excision Repair; DNA binding; cell expansion; DNA Repeat Expansion; histone acetylation; N-terminal binding;	TATA-Binding Protein/Transcription Activation; TATA-Binding Protein/transcription initiation; TATA-Binding Protein/formation of translation preinitiation complex; TATA-Binding Protein/DNA binding; TATA-Binding Protein/Transactivation; RNA Polymerase B/transcription initiation; CAG-2/DNA Repeat Expansion; TFIID/formation of translation preinitiation complex; XPB/Nucleotide Excision Repair; RNA Polymerase B/formation of translation preinitiation complex; Transcription Activation/TFIID; Gene Expression/TATA-Binding Protein; Transcription Activation/Taf protein; formation of translation preinitiation complex/TFIIB; Transcription Activation/hGCN5 gene product;
CCR4-NOT Community	
Transcription; Transcription Activation; gene complementation; derepression; deadenylation-dependent mRNA decay; rRNA transcription; Protein Splicing; S Phase; transcription initiation; transcription initiation factor activity; formation of translation preinitiation complex; chromosome loss; DNA synthesis; Chromosome Segregation; Site-Directed Mutageneses; Genetic Epistasis; Nucleotide Excision Repair;	upstream binding factor/rRNA transcription; Transactivation/CBF1 protein, human; GCN4 protein, S cerevisiae/derepression; Gene Expression/CCR4 gene; Amino Acids/derepression; transcription initiation/Pol I; Transcription Factors, TFI/transcription initiation factor activity; Protein Splicing/TFP1 protein, S cerevisiae; Gene Expression/MTH1 protein, human; Gene Expression/Hxt protein, mouse; Transcription Activation/CBF1 protein, human; transcription initiation/Transcription Factors, TFI; Gene Expression/CBF1 protein, human; CCR4 gene/deadenylation-dependent mRNA decay; Gene Expression/COP1 protein, human;
RFC Community	
Transcription; immunoreactivity; mutant; cell assembly; Cell Cycle; DNA Damage; S Phase; Mitoses; DNA damage checkpoint; replication; Repeat; PHF; Recombination; DNA synthesis; ribosomal RNA; Exostosis;	S Phase/Cds1 kinase; Cds1 kinase/DNA Damage; DNA Damage/rad24 protein; DNA Damage/MEC1 protein, S cerevisiae; dynein/MAP2 gene; DNA Damage/Rad17 protein; Exostosis/MAP2 gene; DNA Damage/rad9 protein; phosphokinase/MAP2 gene; EGF/MAP2 gene; DNA Damage/CHK2 protein, human; DNA Damage/CHK1; upstream binding factor/rRNA transcription; Cell Cycle/POL1 protein, S cerevisiae; Microtubule Proteins/dynein; replication factor A/DNA Repair;
ARP2/ARP3	
mutant; defects; Endocytosis; DELETION; Localization; Growth; **regulation of actin polymerization;** Increased Cell Wall Integrity; Cell Cycle; gene complementation; cell assembly; Cytokineses; Organization; Disruption; cell growth;	**regulation of actin polymerization/Arp2 protein,** human; **Endocytosis/nfo protein, E coli; SLT2 protein, S cerevisiae/calcineurin;** **Endocytosis/Clathrin;** **SLT2 protein, S cerevisiae/Increased Cell Wall Integrity; Arp2 protein, human/actin nucleation; actin monomer binding/Arp2 protein,** human; Mitoses/GIN4 protein, S cerevisiae; PKC1 protein, S cerevisiae/Increased Cell Wall Integrity; ubiquitin/DOA4 protein, S cerevisiae; Cytokineses/Cla4p; **p38/SLT2 protein, S cerevisiae;** CHS/CHS; **protein kinase cascade/SLT2 protein**, S cerevisiae;

V. CONCLUSIONS

In this paper, we present an ontology based language modeling text mining approach to the annotation of protein community, which provides a different perspective in evaluating and annotating the identified protein communities. The method is composed of four steps: retrieved documents from PubMed for each community as a separate collection; extract and index concept and concept pairs for each collection with UMLS support; use mixture language model to rank concepts and concept pairs for each collection, according to the probability that they belong to the common theme of the collection; choose top k ranked concepts and concept pairs respectively for each collection, which serve as functional annotation of the according protein community. The algorithm can automatically extract the top high probability biomedical concepts and concept pairs that form an indicative common theme for an identified community.

Compared to traditional method such as GO or MIPS, our approach has following advantages. First, biomedical literature mining brings much richer information than existed gene databases. Second, the mixture language model can help "purify" the document by eliminating some background noise. Third, using domain ontology UMLS, we extract biological concept and concept pairs from abstracts. Biological concept is more meaningful than word or multi-word phrases. Moreover, using concept pairs can deliver much more information, put

more control on extracted contents and serve as evidence of annotation results.

Experiment results demonstrate our approach's usefulness and potential in helping elucidate the biological relevance of these communities. In future, we will consider integrating domain knowledge to make structured summary for a given community.

ACKNOWLEDGEMENT

This work is supported in part by NSF Career grant (NSF IIS 0448023), NSF CCF 0514679, PA Dept of Health Tobacco Settlement Formula Grant (No. 240205 and No. 240196), and PA Dept of Health Grant (No. 239667).

REFERENCES

[1]. Barabasi, A.-L. and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5: 101–114.

[2]. Girvan, M. and Newman, M.E.J. (2002). Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. 99: 7821-7826.

[3]. Newman, M.E.J. (2003). The Structure and Function of Complex Networks. SIAM Review 45(2): 167-256

[4]. Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. Bioinformatics 19(4): 532-538.

[5]. Wilkinson, D. and Huberman, B.A (2004). A Method for Finding Communities of Related Genes. Proc. Natl. Acad. Sci. U.S.A. 101(Suppl 1): 5241-5248.

[6]. Hashimoto,R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., and Dougherty, E.R. (2004). Growing genetic regulatory networks from seed genes. Bioinformatics 20(8): 1241–1247.

[7]. Flake, G. W., Lawrence, S. R., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of Web communities, IEEE Computer 35: 66-71.

[8]. Jansen, R., Lan, N., Qian, J., and Gerstein, M. (2002). Integration of genomic datasets to predict protein complexes in yeast. J. Struct. Functional Genomics 2: 71–81.

[9]. Bader, G.D. and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.

[10]. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., Li, G. and Chen, R. (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Res. 31: 2443–2450.

[11]. Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999). From molecular to modular cell biology. Nature 402: C47–C52.

[12]. Spirin, V. and Mirny, L.A. (2003). Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. U.S.A. 100: 12123–12128.

[13]. Wu D., Hu X., An Efficient Approach to Detecting a Protein Community from a Seed , 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'05), Nov. 2005, p1-7.

[14]. Wu D., Hu X., Mining and Analyzing the Topological Structure pf Protein-Protein Interaction Networks, 2006 ACM Symposium on Applied Computing (Bioinformatics Track), April 23-27, Dijon, Bourgogne, France.

[15]. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29

[16]. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: A database for genomes and protein sequences. Nucleic Acids Res. 30: 31–34.

[17]. Breitkreutz, B.-J., Stark, C. and Tyers, M. (2003). The GRID: The General Repository for Interaction Datasets. Genome Biology 4: R23.

[18]. Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 31(1): 248-250.

[19]. Rives, A.W. and Galitski, T. (2003). Modular organization of cellular networks. Proc. Natl. Acad. Sci. U.S.A. 100: 1128–1133.

[20]. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. Proc. Natl. Acad. Sci. U.S.A. 101: 2658-2663.

[21]. Asthana, S., King, O.D., Gibbons, F.D., and Roth, F.P. (2004). Predicting Protein Complex Membership Using Probabilistic Network Reliability. Genome Res. 14: 1170-1175

[22]. Batagelj, V. and Mrvar, A. (1998). Pajek: Program for large network analysis. Connections 21: 47–57.

[23]. Wu, P.Y., Ruhlmann, C., Winston, F., and Schultz, P. (2004). Molecular architecture of the S. cerevisiae SAGA complex. Mol. Cell 15: 199–208.

[24]. Guglielmi, B., van Berkum, N.L., Klapholz, B., Bijma, T., Boube, M., Boschiero, C., Bourbon, H.M., Holstege, F.C.P., and Werner, M. (2004). A high resolution protein interaction map of the yeast Mediator complex. Nucleic Acids Res. 32: 5379–5391.

[25]. Ho, Y., et al (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180 – 183.

[26]. Gavin, A.-C., et al (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141 – 147.

[27]. Machesky, L.M. and Gould, K.L. (1999). The Arp2/3 complex: a multifunctional actin organizer. Curr. Opin. Cell Biol. 11: 117 – 121.

[28]. Hashimoto,R.F., Kim, S., Shmulevich, I., Zhang, W., Bittner, M.L., and Dougherty, E.R. (2004). Growing genetic regulatory networks from seed genes. Bioinformatics 20(8): 1241–1247.

[29]. Flake, G. W., Lawrence, S. R., Giles, C. L., and Coetzee, F. M. (2002). Self-organization and identification of Web communities, IEEE Computer 35: 66-71.

[30]. Ling, X., Jiang, J., He,X., Qiaozhu Mei, Chengxiang Zhai, and Bruce Schatz, Automatically Generating Gene Summaries from Biomedical Literature, Pacific Symposium on Biocomputing 11:40-51(2006)

[31]. Hu, X. "Integration of Cluster Ensemble and Text Summarization for Gene Expression Analysis," bibe, p. 251, Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE'04), 2004

[32]. Popescu, M. Keller, J.M. Mitchell, J.A. Bezdek, J.C., Intelligent Sensors, Sensor Networks and Information Processing Conference, pages 553- 558 (2004)

[33]. P. Kankar, S. Adak, A. Sarkar, K. Murari, K. and G. Sharma. "MedMeSH Summarizer: Text Mining for Gene Clusters", in the Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, 2002

[34]. UMLS, http://www.nlm.nih.gov/research/umls/

[35]. Ponte, J. and Croft, W. B. A language modeling approach to information retrieval. In 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pages 275-281, 1998.

[36]. Zhai, C. and Lafferty, J., Model-based feedback in the language modeling approach to information retrieval, Tenth International Conference on Information and Knowledge Management (CIKM 2001), 2001

[37]. Zhai, C. Velivelli, A. , Yu, B. A cross-collection mixture model for comparative text mining, Proceedings of ACM KDD 2004 (KDD'04), pages 743-748, 2004.

[38]. Dempster, A. P., Laird, N. M., & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, 39 (1), 1-38. 1977

[39]. Zhou, X., Zhang, X., Hu, X.,MaxMatcher: Biological Concept Extraction Using Approximate Dictionary Lookup", the Ninth Pacific Rim International Conference on Artificial Intelligence (PRICAI2006), Guilin, China.

[40]. Zhou X., Hu X., Zhang X., Lin X., Song I-Y., Context-Sensitive Semantic Smoothing for the Language Modeling Approach to Genomic IR, accepted in the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR 2006).

[41]. Zhou, X., Zhang, X., and Hu, X., "Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR", The 28th European Conference on Information Retrieval (ECIR' 2006), 10 - 12 April, 2006, London, UK, pp. 444-455.