Evaluating Joint Modeling of Yeast Biology Literature and Protein-Protein Interaction Networks

Ramnath Balasubramanyan and **Kathryn Rivard** and **William W. Cohen**
School of Computer Science
Carnegie Mellon University
rbalasub,krivard,wcohen@cs.cmu.edu

Jelena Jakovljevic and **John Woolford**
Deparment of Biological Sciences
Carnegie Mellon University
jelena,jw17@andrew.cmu.edu

Abstract

Block-LDA is a topic modeling approach to perform data fusion between entity-annotated text documents and graphs with entity-entity links. We evaluate Block-LDA in the yeast biology domain by jointly modeling PubMed® articles and yeast protein-protein interaction networks. The topic coherence of the emergent topics and the ability of the model to retrieve relevant scientific articles and proteins related to the topic are compared to that of a text-only approach that does not make use of the protein-protein interaction matrix. Evaluation of the results by biologists show that the joint modeling results in better topic coherence and improves retrieval performance in the task of identifying top related papers and proteins.

1 Introduction

The prodigious rate at which scientific literature is produced makes it virtually impossible for researchers to manually read every article to identify interesting and relevant papers. It is therefore critical to have automatic methods to analyze the literature to identify topical structure in it. The latent structure that is identified can be used for different applications such as enabling browsing, retrieval of papers related to a particular sub-topic etc. Such applications assist in common scenarios such as helping a researcher identify a set of articles to read (perhaps a set of well-regarded surveys) to familiarize herself with a new sub-field; helping a researcher to stay abreast with the latest advances in his field by identifying relevant articles etc.

In this paper, we focus on the task of organizing a large collection of literature about yeast biology to enable topic oriented browsing and retrieval from the literature. The analysis is performed using topic modeling(Blei et al., 2003) which has, in the last decade, emerged as a versatile tool to uncover latent structure in document corpora by identifying broad topics that are discussed in it. This approach complements traditional information retrieval tasks where the objective is to fulfill very specific information needs.

In addition to literature, there often exist other sources of domain information related to it. In the case of yeast biology, an example of such a resource is a database of known protein-protein interactions (PPI) which have been identified using wetlab experiments. We perform data fusion by combining text information from articles and the database of yeast protein-protein interactions, by using a latent variable model — Block-LDA (Balasubramanyan and Cohen, 2011) that jointly models the literature and PPI networks.

We evaluate the ability of the topic models to return meaningful topics by inspecting the top papers and proteins that pertain to them. We compare the performance of the joint model i.e. Block-LDA with a model that only considers the text corpora by asking a yeast biologist to evaluate the coherence of topics and the relevance of the retrieved articles and proteins. This evaluation serves to test the utility of Block-LDA on a real task as opposed to an internal evaluation (such as by using perplexity metrics for example). Our evaluaton shows that the joint model outperforms the text-only approach both in topic co-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 155–162,
Montréal, Canada, June 8, 2012.

herence and in top paper and protein retrieval as measured by precision@10 values.

The rest of the paper is organized as follows. Section 2 describes the topic modeling approach used in the paper. Section 3 describes the datasets used followed by Section 4 which details the setup of the experiments. The results of the evaluation are presented in Section 5 which is followed by the conclusion.

2 Block-LDA

The Block-LDA model (plate diagram in Figure 1) enables sharing of information between the component on the left that models links between pairs of entities represented as edges in a graph with latent block structure, and the component on the right that models text documents, through shared latent topics. More specifically, the distribution over the entities of the type that are linked is shared between the block model and the text model.

The component on the right, which is an extension of the LDA models documents as sets of "bags of entities", each bag corresponding to a particular type of entity. Every entity type has a topic wise multinomial distribution over the set of entities that can occur as an instance of the entity type. This model is termed as Link-LDA(Nallapati et al., 2008) in the literature.

The component on the left in the figure is a generative model for graphs representing entity-entity links with an underlying block structure, derived from the sparse block model introduced by Parkkinen et al. (2009). Linked entities are generated from topic specific entity distributions conditioned on the topic pairs sampled for the edges. Topic pairs for edges (links) are drawn from a multinomial defined over the Cartesian product of the topic set with itself. Vertices in the graph representing entities therefore have mixed memberships in topics. In contrast to Mixed-membership Stochastic Blockmodel (MMSB) introduced by Airoldi et al. (2008), only observed links are sampled, making this model suitable for sparse graphs.

Let K be the number of latent topics (clusters) we wish to recover. Assuming documents consist of T different types of entities (i.e. each document contains T bags of entities), and that links in the graph are between entities of type t_l, the generative process is as follows.

1. Generate topics: For each type $t \in 1, \ldots, T$, and topic $z \in 1, \ldots, K$, sample $\beta_{t,z} \sim \text{Dirichlet}(\gamma)$, the topic specific entity distribution.

2. Generate documents. For every document $d \in \{1 \ldots D\}$:

- Sample $\theta_d \sim \text{Dirichlet}(\alpha_D)$ where θ_d is the topic mixing distribution for the document.
- For each type t and its associated set of entity mentions $e_{t,i}$, $i \in \{1, \cdots, N_{d,t}\}$:
  - Sample a topic $z_{t,i} \sim \text{Multinomial}(\theta_d)$
  - Sample an entity $e_{t,i} \sim \text{Multinomial}(\beta_{t,z_{t,i}})$

3. Generate the link matrix of entities of type t_l:

- Sample $\pi_L \sim \text{Dirichlet}(\alpha_L)$ where π_L describes a distribution over the Cartesian product of the set of topics with itself, for links in the dataset.
- For every link $e_{i1} \rightarrow e_{i2}$, $i \in \{1 \cdots N_L\}$:
  - Sample a topic pair $\langle z_{i1}, z_{i2} \rangle \sim \text{Multinomial}(\pi_L)$
  - Sample $e_{i1} \sim \text{Multinomial}(\beta_{t_l,z_{i1}})$
  - Sample $e_{i2} \sim \text{Multinomial}(\beta_{t_l,z_{i2}})$

Note that unlike the MMSB model, this model generates only realized links between entities.

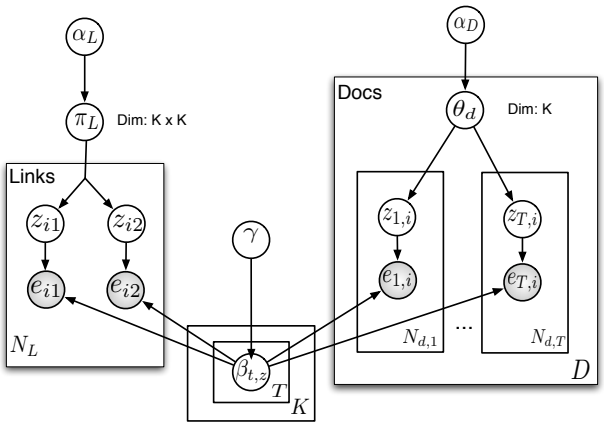
Given the hyperparameters α_D, α_L and γ, the joint distribution over the documents, links, their topic distributions and topic assignments is given by

$$p(\pi_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{e}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_D, \alpha_L, \gamma) \propto \tag{1}$$

$$\prod_{z=1}^{K} \prod_{t=1}^{T} \text{Dir}(\beta_{t,z} | \gamma_t) \times$$

$$\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha_D) \prod_{t=1}^{T} \prod_{i=1}^{N_{d,t}} \theta_d^{z_{t,i}^{(d)}} \beta_{t,z_{t,i}^{(d)}}^{e_{t,i}} \times$$

$$\text{Dir}(\pi_L | \alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{t_l,z1}^{e_{i1}} \beta_{t_l,z2}^{e_{i2}}$$



α_L - Dirichlet prior for the topic pair distribution for links
α_D - Dirichlet prior for document specific topic distributions
γ - Dirichlet prior for topic multinomials
π_L - multinomial distribution over topic pairs for links
θ_d - multinomial distribution over topics for document d
$\beta_{t,z}$ - multinomial over entities of type t for topic z
$z_{t,i}$ - topic chosen for the i-th entity of type t in a document
$e_{t,i}$ - the i-th entity of type t occurring in a document
z_{i1} and z_{i2} - topics chosen for the two nodes participating in the i-th link
e_{i1} and e_{i2} - the two nodes participating in the i-th link

Figure 1: Block-LDA

A commonly required operation when using models like Block-LDA is to perform inference on the model to query the topic distributions and the topic assignments of documents and links. Due to the intractability of exact inference in the Block-LDA model, a collapsed Gibbs sampler is used to perform approximate inference. It samples a latent topic for an entity mention of type t in the text corpus conditioned on the assignments to all other entity mentions using the following expression (after collapsing θ_D):

$$p(z_{t,i} = z | e_{t,i}, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{e}^{\neg \mathbf{i}}, \alpha_D, \gamma) \quad (2)$$
$$\propto (n_{dz}^{\neg i} + \alpha_D) \frac{n_{zte_{t,i}}^{\neg i} + \gamma}{\sum_{e'} n_{zte'}^{\neg i} + |E_t|\gamma}$$

Similarly, we sample a topic pair for every link conditional on topic pair assignments to all other links after collapsing π_L using the expression:

$$p(\mathbf{z_i} = \langle z_1, z_2 \rangle | \langle e_{i1}, e_{i2} \rangle, \mathbf{z}^{\neg \mathbf{i}}, \langle \mathbf{e_1}, \mathbf{e_2} \rangle^{\neg \mathbf{i}}, \alpha_L, \gamma) \quad (3)$$
$$\propto \left(n_{\langle z_1, z_2 \rangle}^{L \neg i} + \alpha_L \right) \times$$
$$\frac{\left(n_{z_1 t_l e_{i1}}^{\neg i} + \gamma\right)\left(n_{z_2 t_l e_{i2}}^{\neg i} + \gamma\right)}{\left(\sum_e n_{z_1 t_l e}^{\neg i} + |E_{t_l}|\gamma\right)\left(\sum_e n_{z_2 t_l e}^{\neg i} + |E_{t_l}|\gamma\right)}$$

E_t refers to the set of all entities of type t. The n's are counts of observations in the training set.

- n_{zte} - the number of times an entity e of type t is observed under topic z
- n_{zd} - the number of entities (of any type) with topic z in document d
- $n_{\langle z_1, z_2 \rangle}^{L}$ - count of links assigned to topic pair $\langle z_1, z_2 \rangle$

The topic multinomial parameters and the topic distributions of links and documents are easily recovered using their MAP estimates after inference

using the counts of observations.

$$\beta_{t,z}^{(e)} = \frac{n_{zte} + \gamma}{\sum_{e'} n_{zte'} + |E_t|\gamma}, \quad (4)$$

$$\theta_d^{(z)} = \frac{n_{dz} + \alpha_D}{\sum_{z'} n_{dz'} + K\alpha_D} \text{and} \quad (5)$$

$$\pi_L^{\langle z_1, z_2 \rangle} = \frac{n_{\langle z_1, z_2 \rangle} + \alpha_L}{\sum_{z_1', z_2'} n_{\langle z_1', z_2' \rangle} + K^2\alpha_L} \quad (6)$$

A de-noised form of the entity-entity link matrix can also be recovered from the estimated parameters of the model. Let B be a matrix of dimensions $K \times |E_{t_l}|$ where row $k = \beta_{t_l,k}$, $k \in \{1, \cdots, K\}$. Let Z be a matrix of dimensions $K \times K$ s.t $Z_{p,q} = \sum_{i=1}^{N_L} \mathbf{I}(z_{i1} = p, z_{i2} = q)$. The de-noised matrix M of the strength of association between the entities in E_{t_l} is given by $M = B^T Z B$.

In the context of this paper, de-noising the protein-protein interaction networks studied is an important application. The joint model permits information from the large text corpus of yeast publications to be used to de-noise the PPI network and to identify potential interactions that are missing in the observed network. While this task is important and interesting, it is outside the scope of this paper and is a direction for future work.

3 Data

We use a collection of publications about yeast biology that is derived from the repository of scientific publications at PubMed®. PubMed® is a free, open-access on-line archive of over 18 million biological abstracts and bibliographies, including citation lists, for papers published since 1948. The subset we work with consists of approximately 40,000 publications about the yeast organism that have been curated in the Saccharomyces Genome Database (SGD) (Dwight et al., 2004) with annotations of proteins that are discussed in the publication. We further restrict the dataset to only those documents that are annotated with at least one protein from the protein-protein interactions databases described below. This results in a protein annotated document collection of 15,776 publications. The publications in this set were written by a total of 47,215 authors. We tokenize the titles and abstracts based on white space, lowercase all tokens and eliminate stopwords. Low frequency (< 5 occurrences) terms are also eliminated. The vocabulary that is obtained consists of 45,648 words.

The Munich Institute for Protein Sequencing (MIPS) database (Mewes et al., 2004) includes a hand-crafted collection of protein interactions covering 8000 protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated.

Finally, we use another dataset of protein-protein interactions in yeast that were observed as a result of wetlab experiments by collaborators of the authors of the paper. This dataset consists of 635 interactions that deal primarily with ribosomal proteins and assembly factors in yeast.

4 Setup

We conduct three different evaluations of the emergent topics. Firstly, we obtain topics from only the text corpus using a model that comprises of the right half of Figure 1 which is equivalent to using the Link-LDA model. For the second evaluation, we use the Block-LDA model that is trained on the text corpus and the MIPS protein-protein interaction database. Finally, for the third evaluation, we replace the MIPS database with the interaction obtained from the wetlab experiments. In all the cases, we set K, the number of topics to be 15. In each variant, we represent documents as 3 sets of entities i.e. the words in the abstracts of the article, the set of proteins associated with the article as indicated in the SGD database and finally the authors who wrote the article. Each topic therefore consists of 3 different multinomial distributions over the sets of the 3 kinds of entities described.

Topics that emerge from the different variants can possibly be assigned different indices even when they discuss the same semantic concept. To compare topics across variants, we need a method to determine which topic indices from the different variants correspond to the same semantic concept. To obtain the mapping between topics from each variant, we utilize the Hungarian algorithm (Kuhn, 1955) to solve the assignment problem where the cost of aligning topics together is determined using the Jensen-Shannon divergence measure.

Once the topics are obtained, we firstly obtain the proteins associated with the topic by retrieving the

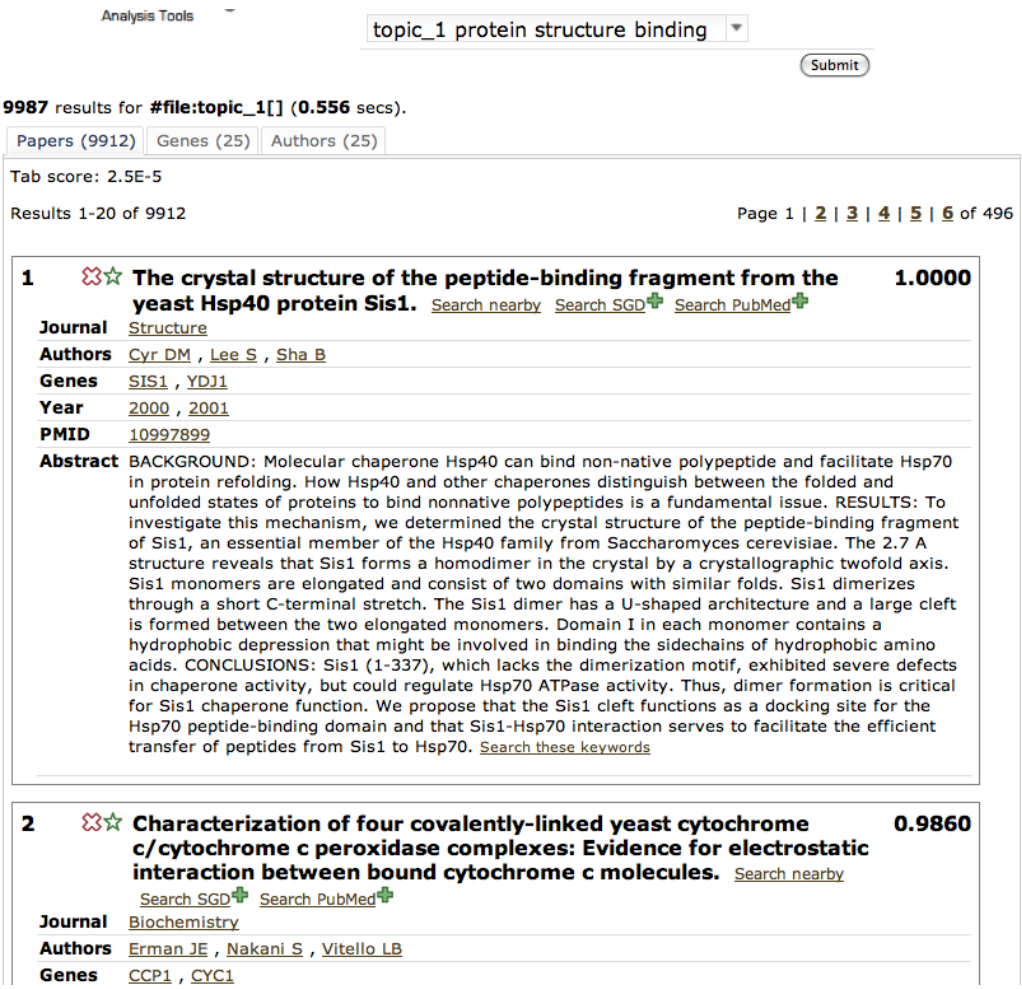


Figure 2: Screenshot of the Article Relevance Annotation Tool

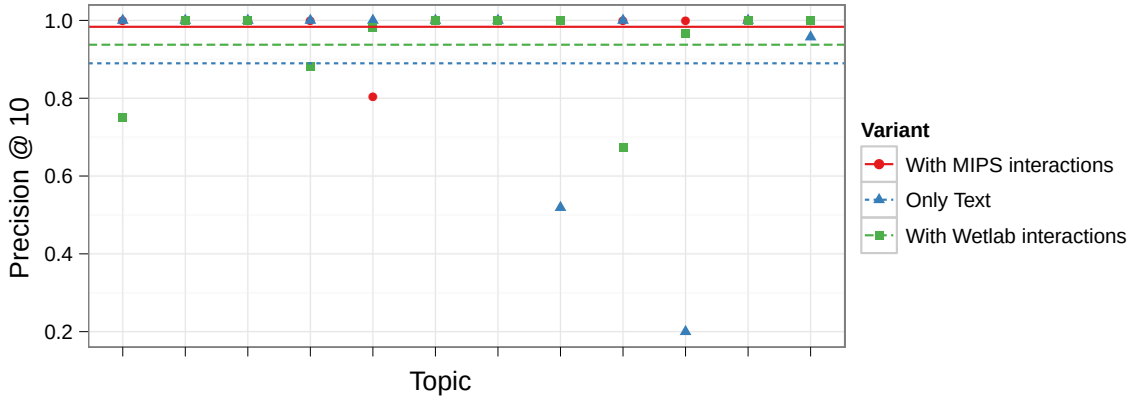
Variant	Num. Coherent Topics
Only Text	12 / 15
Text + MIPS	13 / 15
Text + Wetlab	15 / 15

Table 1: Topic Coherence Evaluation

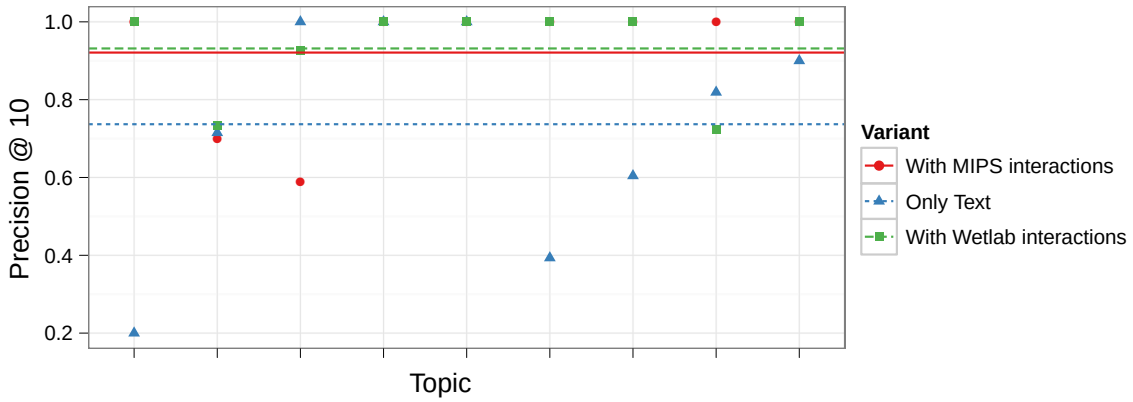
top proteins from the multinomial distribution corresponding to proteins. Then, the top articles corresponding to each topic is obtained using a ranked list of documents with the highest mass of their topic proportion distributions (θ) residing in the topic being considered.

4.1 Manual Evaluation

To evaluate the topics, a yeast biologist who is an expert in the field was asked to mark each topic with a binary flag indicating if the top words of the distribution represented a coherent sub-topic in yeast biology. This process was repeated for the 3 different variants of the model. The variant used to obtain results is concealed from the evaluator to remove the possibility of bias. In the next step of the evaluation, the top articles and proteins assigned to each topic were presented in a ranked list and a similar judgement was requested to indicate if the article/protein was relevant to the topic in question. Similar to the topic coherence judgements, the process was repeated for each variant of the model. Screenshots of the tool used for obtaining the judgments can be seen in Figure 2. It should be noted that since the nature of the topics in the literature considered was highly technical and specialized, it was impractical to get judgements from multiple annotators.



(a) Article Retrieval



(b) Protein Retrieval

Figure 3: Retrieval Performance Evaluation (Horizontal lines indicate mean across all topics)

To evaluate the retrieval of the top articles and proteins, we measure the quality of the results by computing its precision@10 score.

5 Results

First we evaluate the coherence of the topics obtained from the 3 variants described above. Table 1 shows that out of the 15 topics that were obtained, 12 topics were deemed coherent from the text-only model and 13 and 15 topics were deemed coherent from the Block-LDA models using the MIPS and wetlab PPI datasets respectively.

Next, we study the precision@10 values for each topic and variant for the article retrieval and protein retrieval tasks, which is shown in Figure 3. The plots also show horizontal lines representing the mean of the precision@10 across all topics. It can be seen from the plots that for both the article and protein retrieval tasks, the joint models work better than the text-only model on average. For the article retrieval task, the model trained with the text + MIPS resulted in the higher mean precision@10 whereas for the protein retrieval task, the text + Wetlab PPI dataset returned a higher mean precision@10 value. For both the protein retrieval and paper retrieval tasks, the improvements shown by the joint models using either of the PPI datasets over the text-only model (i.e. the Link LDA model) were statistically significant at the 0.05 level using the paired Wilcoxon sign test. The difference in performance between the

Topic: Protein Structure & Interactions	
Top articles using Publications Only	**Top articles using Block-LDA with Wetlab PPI**
* X-ray fiber diffraction of amyloid fibrils.	* X-ray fiber diffraction of amyloid fibrils.
* Molecular surface area and hydrophobic effect.	* Scalar couplings across hydrogen bonds.
* Counterdiffusion methods for macromolecular crystallization.	* Dipolar couplings in macromolecular structure determination.
* Navigating the ClpB channel to solution.	* Structure of alpha-keratin.
* Two Rippled-Sheet Configurations of Polypeptide Chains, and a Note about the Pleated Sheets.	* Stable configurations of polypeptide chains.
* Molecular chaperones. Unfolding protein folding.	* The glucamylase and debrancher of S. diastaticus.
* The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure.	* A study of 150 cases of pneumonia.
* Unfolding and hydrogen exchange of proteins: the three-dimensional ising lattice as a model.	* Glycobiology.
* Packing of alpha-helices: geometrical constraints and contact areas.	* The conformation of thermolysin.
Topic: DNA Repair	
Top articles using Publications Only	**Top articles using Block-LDA with Wetlab PPI**
* Passing the baton in base excision repair.	* Telomeres and telomerase.
* The bypass of DNA lesions by DNA and RNA polymerases.	* Enzymatic photoreactivation: overview.
* The glucamylase and debrancher of S. diastaticus.	* High-efficiency transformation of plasmid DNA into yeast.
* DNA replication fidelity.	* The effect of ultraviolet light on recombination in yeast.
* Base excision repair.	* T-loops and the origin of telomeres.
* Nucleotide excision repair.	* Directed mutation: between unicorns and goats.
* The replication of DNA in Escherichia Coli.	* Functions of DNA polymerases.
* DNA topoisomerases: why so many?	* Immortal strands? Give me a break.

Table 2: Sample of Improvements in Article Retrieval

two joint models that used the two different PPI networks were however insignificant which indicates that there is no observable advantage in using one PPI dataset over the other in conjunction with the text corpus.

Table 2 shows examples of poor results of article retrieval obtained using the publications-only model and the improved set of results obtained using the joint model.

5.1 Topics

Table 3 shows 3 sample topics that were retrieved from each variant described earlier. The table shows the top words and proteins associated with the topics. The topic label on the left column was assigned manually during the evaluation by the expert annotator.

Conclusion

We evaluated topics obtained from the joint modeling of yeast biology literature and protein-protein interactions in yeast and compared them to topics that were obtained from using only the literature. The topics were evaluated for coherence and by measuring the mean precision@10 score of the top articles and proteins that were retrieved for each topic. Evaluation by a domain expert showed that

Topic	Top Words & Proteins
Protein Structure & Interactions (Publications Only)	**Words:** protein structure binding residues domain structural beta complex atp proteins alpha interactions folding structures form terminal peptide helix model interaction bound domains molecular changes conformational **Proteins:** CYC1 SSA1 HSP82 SUP35 HSP104 HSC82 SSA2 YDJ1 URE2 KAR2 SSB1 SSA4 GCN4 SSA3 SSB2 PGK1 PDI1 SSC1 HSP60 STI1 SIS1 RNQ1 SEC61 SSE1 CCP1
DNA Repair (Using MIPS PPI)	**Words:**dna recombination repair replication strand single double cells mutations stranded induced base uv mutants mutation homologous virus telomere human type yeast activity telomerase mutant dna_polymerase **Proteins:** RAD52 RAD51 RAD50 MRE11 RAD1 RAD54 SGS1 MSH2 RAD6 YKU70 REV3 POL30 RAD3 XRS2 RAD18 RAD2 POL3 RAD27 YKU80 RAD9 RFA1 TLC1 TEL1 EST2 HO
Vesicular Transport (Using Wetlab PPI)	**Words:**membrane protein transport proteins atp golgi er atpase membranes plasma_membrane vesicles cells endoplasmic_reticulum complex fusion ca2 dependent translocation vacuolar intracellular yeast lipid channel hsp90 vesicle **Proteins:** SSA1 HSP82 KAR2 PMA1 HSC82 SEC18 SSA2 YDJ1 SEC61 PEP4 HSP104 SEC23 VAM3 IRE1 SEC4 SSA4 SEC1 PMR1 PEP12 VMA3 VPH1 SSB1 VMA1 SAR1 HAC1

Table 3: Sample Topics

the joint modeling produced more coherent topics and showed better precision@10 scores in the article and protein retrieval tasks indicating that the model enabled information sharing between the literature and the PPI networks.

References

Edoardo M. Airoldi, David Blei, Stephen E. Fienberg, and Eric P. Xing. 2008. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, September.

Ramnath Balasubramanyan and William W. Cohen. 2011. Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *SDM*, pages 450–461. SIAM / Omnipress.

David. M Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Selina S. Dwight, Rama Balakrishnan, Karen R. Christie, Maria C. Costanzo, Kara Dolinski, Stacia R. Engel, Becket Feierbach, Dianna G. Fisk, Jodi Hirschman, Eurie L. Hong, Laurie Issel-Tarver, Robert S. Nash, Anand Sethuraman, Barry Starr, Chandra L. Theesfeld, Rey Andrada, Gail Binkley, Qing Dong, Christopher Lane, Mark Schroeder, Shuai Weng, David Botstein, and Michael Cherry J. 2004. Saccharomyces genome database: Underlying principles and organisation. *Briefings in bioinformatics*, 5(1):9.

Harold W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Hans-Werner Mewes, C. Amid, Roland Arnold, Dmitrij Frishman, Ulrich Gldener, Gertrud Mannhaupt, Martin Mnsterktter, Philipp Pagel, Normann Strack, Volker Stmpflen, Jens Warfsmann, and Andreas Ruepp. 2004. MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32:41–44.

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, Las Vegas, Nevada, USA. ACM.

Juuso Parkkinen, Janne Sinkkonen, Adam Gyenge, and Samuel Kaski. 2009. A block model suitable for sparse graphs. In *Proceedings of the 7th International Workshop on Mining and Learning with Graphs (MLG 2009)*, Leuven. Poster.