BMC Bioinformatics

BioMed Central

Research article

Open Access

Identifying biological concepts from a protein-related corpus with a probabilistic topic model

Bin Zheng, David C McLean Jr and Xinghua Lu*

Address: Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC 29405, USA

Email: Bin Zheng - zheng@musc.edu; David C McLean - mcleandc@musc.edu; Xinghua Lu* - lux@musc.edu

* Corresponding author

Published: 08 February 2006

*BMC Bioinformatics*2006, **7**:58 doi:10.1186/1471-2105-7-58

Received: 14 September 2005

Accepted: 08 February 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/58

Abstract

Background: Biomedical literature, e.g., MEDLINE, contains a wealth of knowledge regarding functions of proteins. Major recurring biological concepts within such text corpora represent the domains of this body of knowledge. The goal of this research is to identify the major biological topics/concepts from a corpus of protein-related MEDLINE© titles and abstracts by applying a probabilistic topic model.

Results: The latent Dirichlet allocation (LDA) model was applied to the corpus. Based on the Bayesian model selection, 300 major topics were extracted from the corpus. The majority of identified topics/concepts was found to be semantically coherent and most represented biological objects or concepts. The identified topics/concepts were further mapped to the controlled vocabulary of the Gene Ontology (GO) terms based on mutual information.

Conclusion: The major and recurring biological concepts within a collection of MEDLINE documents can be extracted by the LDA model. The identified topics/concepts provide parsimonious and semantically-enriched representation of the texts in a semantic space with reduced dimensionality and can be used to index text.

Background

An important task of bioinformatics research is to acquire and represent biomedical knowledge in computable form so that it can be efficiently stored, retrieved, and used for discovery of new knowledge. For example, the Gene Ontology (GO) Consortium [1] and the Gene Ontology Annotation (GOA) project [2] are dedicated to the task of representing biological knowledge with the controlled vocabulary of GO terms. Knowledge of protein functions serves as a cornerstone of modern biomedical knowledge. Much of such knowledge is contained in the form of free text in biomedical literature. A more compressed and accessible representation of this same knowledge is contained in bibliographic databases, e.g., MEDLINE. In addition to current manual annotation efforts, needs for automatic knowledge acquisition and representation exist, and a critical step of this process is to extract biological concepts from free text.

The task of automatic knowledge acquisition from free text is usually addressed within the frameworks of the natural language processing (NLP), information extraction (IE), and information retrieval (IR) techniques [3-5], which has been wide applied in bioinformatics setting, as reviewed in [6-9]. Recent trend in text mining is to acquire deeper semantic information from text, e.g., semantic information has be used to cluster genes [10] and evaluate the functional coherence of a group of genes [11-13].

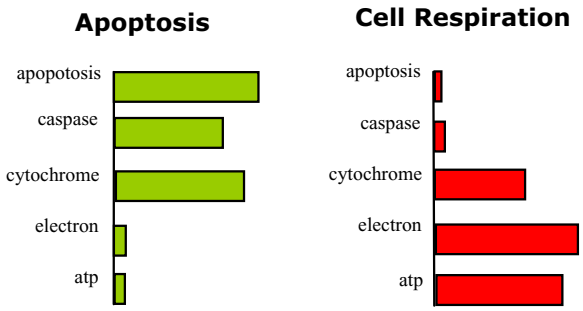

Figure 1
Representing concepts with word distributions. Two hypothetic topics are depicted. The bar lengths indicate the word usage preference in form of probability.

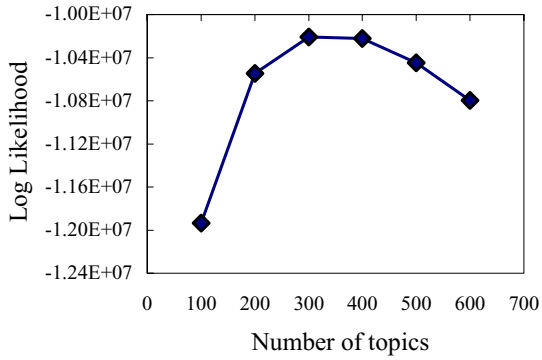

Figure 2
Bayesian model selection. The means of approximated evidence for different models are plotted; standard error bars are within the symbols.

Extracting semantic information from free text requires the capability of effectively dealing with the uncertainties commonly associated with human language. To this end, probabilistic semantic analyses serve as promising approaches for handling such uncertainties and performing semantically enriched text mining.

In this paper, we report extraction of semantic topics/concepts from a corpus of MEDLINE titles and abstracts using a probabilistic topic model, the LDA model [14,15]. The goal was to identify the major and recurring concepts that represent the major knowledge domains of protein functions. Furthermore, extraction of the semantic contents of a document provides a parsimonious and concise representation of that text. Such information can be used for efficient indexing, information retrieval, and protein annotation.

Results

Representing semantic topics with a probabilistic topic model

In a scientific article, a scientist will refer to multiple real world objects and/or concepts, thus a paper usually consists of multiple topics/subjects, e.g., a paper may discuss a protein located in *mitochondria* and involved in the cellular process of *apoptosis*. When discussing objects or concepts, the author will choose certain words to convey the semantic meaning. For instance, when discussing the topic *mitochondria*, words like 'electron,' 'cytochrome,' and 'ATP' are commonly used, while words like 'apoptosis,' 'programmed,' 'death,' and 'caspase' are commonly used to discuss the concept of *apoptosis*. Thus a document can be treated as a mixture of words from multiple topics. The LDA model represent such a notion by explicitly encoding multi-topicality of a document with a topic-composition variable and then simulating the "generation" of words by accordingly mixing words from topics, which are represented as multinomial distributions over a vocabulary, i.e., a word-usage pattern. Figure 1 shows how a topic can be represented as word-usage pattern in a probabilistic topic model. Given a corpus of text documents, the LDA model is capable of extracting the topics by statistical inference as described in the Methods section.

Training of LDA model

The LDA model was applied to extract the semantic topics from a corpus of MEDLINE titles and abstracts downloaded from the GOA project website as described in the Methods section. The training of an LDA model requires specification of the number of topics for the models, an issue of interest from both semantic analysis and statistical learning view points. From a semantic analysis point of view, this is equivalent to determining the granularity of abstraction of the concepts that can be used to summarize the semantic contents of the corpus. From the statistical learning point of view, this is equivalent to select among the models with different complexity. A Bayesian model selection framework was employed to determine the "optimal" number of topics based on the posterior probability of a model, $p(M \mid \mathbf{w})$. To perform the Bayesian model selection, samples of the latent semantic topics, $\mathbf{z}$, were collected for a model with a given number of topics, T, and the approximate the posterior probabilities were calculated according to equation (7) and plotted (Figure 2). The model with 300 topics had the highest approximated marginal likelihood and was thus used for the analyses reported in this paper.

Evaluating semantic topics

A trained LDA model returns estimated distributions of the following parameters and latent variables: (1) the word-usage distribution, ϕ_t, for each topic; (2) the latent

Table 1: The ten most common topics from a trained LDA model

Topic #	Topic words
51	receptor coupl ligand agonist subtype pharmacolog antagonist orphan adrenerg desensit
156	kinas phosphoryl serin threonin pkc autophosphoryl casein akt catalyt ste20
136	cerevisia saccharomyc strain yeast plasmid multicopi lacz floccul auxotroph gal1
67	Famili member belong multigen subfamily mrg Dalton cabp28k heterogen transmembran
154	patient syndrom diseas disord autosom inherit recess ref caus clinic
124	cdna librari clone probe screen isol lambda obtain oligonucleotid gtl1
37	neuron axon migrat motor glial spinal cord neurit dendrite outgrowth
229	mutant defect doubl phenotyp fail rescu restor impair pleiotrop unable
112	exon intron genom kb flank region span upstream bp start
172	nuclear nucleu export cytoplasm nuclei pore ran hnrnp envelop import

topic labeling z_i for each word w_i; and (3) the topic-composition distribution θ_d for each document. The parameter vector ϕ_t is a distribution representing a word-usage pattern for the topic *t*. High probability words of each ϕ_t can be thought as the words frequently used to discuss the topics. In Table 1, the 10 most commonly observed topics and their high probability words of the trained LDA model are listed. The topics are sorted in descending order according to the number of words assigned to them in the corpus. High probability words of these topics constitute clusters of words that coherently convey biological concepts. For example, topic # 51 reflects the concept of *ligand-activated receptors*, and the topic # 156 is related to *serine/threonin kinase activity*. Because the LDA model attempts to capture the major topics that can be used to "generate" the data, the concepts extracted by this model should reflect the recurring themes of the corpus. Indeed, when multiple models with 300 topics were trained with different random-number seeds, similar major topics were extracted although the index of the topics differed among the models. Thus, the topics listed in Table 1 do reflect common biological themes in our corpus.

Inferring the semantic content of a text

The instantiated latent variables $\mathbf{z}_d$ indicates the semantic contents of the document. For the text in the training data set, the topic contents for each document were returned as the estimated latent variables $\mathbf{z}_d$ of the trained model. For a newly observed text, the topic contents can be inferred by invoke the sampling algorithm with the estimated parameters as described in the Methods section. Figure 3 shows an example of a MEDLINE abstract, in which topic assignment for the words were inferred using a trained LDA model. This abstract discusses a protein referred to as apoptosis inducing factor (AIF), a mitochondrial protein that induces apoptosis. In this figure, the inferred semantic topic for each word (excluding "stop" words) is shown as the superscript numbers next to it. The abstract is associated with the following GO terms: (1) GO:0008630, DNA damage response, signal transduction resulting in induction of apoptosis; (2) GO:0009055, electron carrier activity; (3) GO:0005739, mitochondrion; and (4) GO:0006309, DNA fragmentation during apoptosis. In Figure 3, two major topics, # 73 and # 147, are the dominant topics of the abstract. Topic # 73 is related to the *mitochondrion* and topic # 147 reflects the concept *of apoptosis*. Interestingly, several words, which can belong to multiple topics depending on context, were found in the abstract, e.g., "space" and "outer." The LDA model has captured their common occurrence in the context of *mitochondrion* and correctly assigned these common words to this topic based on the context. With the inferred topics, this abstract can be readily indexed with these two major topics which agree well with the human GO annotations of this abstract. Furthermore, a document can also be indexed as a vector containing the counts of the words in each topic or with the normalized estimated $\hat{\theta}_d$, which be treated as a vector in the space spanned by the topics. Such representation effectively projects the document from the high dimensional vocabulary space onto the reduced-dimensionality of topic space. Such information could be used to automatically index the text.

Assessing biological relevance of topics

The LDA model simulates the "generation" of a corpus. By its generative nature, it will incorporate topics needed to capture the common characteristics in the corpus. However, some common features may not be necessarily relevant to biology but merely reflect the linguistic feature of the corpus. To determine the biological relevance of topics, we further inspected the high probability words and

assigned a biological relevance score, ranging from 0 (indicating no biological relevance) to 5 (representing strong biological relevance) to each topic. A histogram of the assigned biological relevance scores (Panel A of Figure 4) indicates that most topics/concepts extracted from this corpus were biologically relevant, with only a fraction with biological relevance scores equal to zero, indicating no biological relevance.

Each MEDLINE abstract from the GOA corpus was associated with one or more GO terms, providing an opportunity to study the relationship between the semantic topics extracted by the LDA model and the GO annotations. The correlation between the semantic topic and the GO annotation can be quantified by mutual information (MI) between the latent topic and the annotated GO terms. MI is a symmetric, non-negative quantity that measures the relevance (amount of information) of one variable with respect to another variable, which equals zero if and only if the variables are independent. Since GO terms are designed to represent biological objects/concepts, the topics highly relevant to biological objects/concepts should have high MI with some GO terms, while the topics irrelevant to biology should have low MI values for topic-GO association. Indeed, as shown in Figure 4, the topics rated low relevance have very low MI with any GO terms, while topics with high relevance have the highest topic-GO MI (Panel B). However, there were some topics that were assigned high relevance scores but had low MI with GO terms. This disparity was likely due to the way the MI for a topic-GO association was calculated in this study, which specifies that, if a document was annotated with a GO term *g*, every word in the document was considered as annotated with that GO term. This method was adopted due to the lack of supervised training data specifying which words in a document were responsible for the GO annotations. MI calculated under this assumption is skewed for the relatively uncommon topics in the corpus. Nonetheless, the MI of topic-GO association serves as a criterion of evaluating the biological relevance of a topic. When a topic had a high MI value for a topic-GO association, it usually reflected a coherent biological concept. Interestingly, a topic with low biological relevance did not mean that it was not a coherent semantic concept. For example, topics # 224 and # 227 (Table 2) consisted of common English words that therefore had the lowest MI with any GO term. However, the topics did contain the words that constitute coherent semantic concepts, e.g., topic # 224 contains words related to the concept of *being unique*.

Mitochondria[73] play a key part[160] in the regulation[113] of apoptosis[147] (cell[200] death[147]). Their intermembrane[73] space[73] contains[131] several proteins[265] that are liberated[224] through the outer[73] membrane[219] in order[294] to participate[87] in the degradation[299] phase[209] of apoptosis[147]. Here we report[33] the identification[208] and cloning of an apoptosis[147]-inducing[147] factor[19], AIF[147], which is sufficient[3] to induce[147] apoptosis[147] of isolated[76] nuclei[191]. AIF[147] is a flavoprotein[73] of relative[122] molecular[177] mass[185] 57,000 which shares[168] homology[212] with the bacterial[213] oxidoreductases[73]; it is normally[122] confined[123] to mitochondria[73] but translocates[166] to the nucleus[191] when apoptosis[147] is induced[147]. Recombinant[279] aif[147] causes[141] chromatin[51] condensation[279] in isolated[76] nuclei[191] and large-scale[41] fragmentation[174] of dna[126]. It induces[147] purified[213] mitochondria[73] to release[5] the apoptogenic[147] proteins[265] cytochrome[73] c and caspase9[147]. Microinjection[217] of aif[147] into the cytoplasm[81] of intact[257] cells[200] induces[147] condensation[279] of chromatin[51], dissipation[292] of the mitochondrial[73] transmembrane[206] potential[64], and exposure[280] of phosphatidylserine[68] in the plasma[219] membrane[219]. None of these effects[257] is prevented[147] by the wide-ranging[132] caspase[147] inhibitor[170] known[140] as zvad.fmk[172]. Overexpression[150] of bcl2[147], which controls[113] the opening[101] of mitochondrial[73] permeability transition[209] pores[191], prevents[147] the release[5] of aif[147] from the mitochondrion[73] but does not affect[257] its apoptogenic[147] activity[23]. These results[150] indicate[144] that aif[147] is a mitochondrial[73] effector[147] of apoptotic[147] cell[200] death[147].

Figure 3

Semantic analysis for a MEDLINE abstract (PMID 9989411). The topics associated with the words were inferred by the LDA model and are shown as the superscript number next to the words. The words from the topics # 73 and # 147 are highlighted with blue and red colors, respectively.

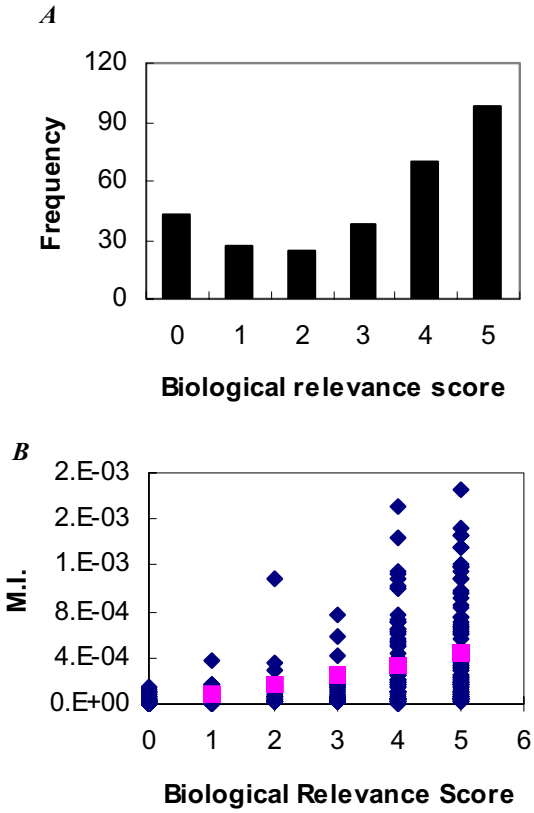

Figure 4
Determining the biological relevance of the topics. ***Panel A***. Histogram of human assigned biological relevance scores. A score of 0 indicates no biological relevance, while scores of 1 through 5 indicate increasingly relevant and coherent biological relevance. ***Panel B***. Relationship between the human assigned biological relevance score and the topic-GO MI.

Associating topic with GO terms

Studying the correlation between the topics and the GO terms also allowed the mapping of topics to the controlled vocabulary of GO terms, laying a foundation for possible future automatic annotation/indexing of MEDLINE abstracts with the GO terms. While annotating a gene product based on biomedical literature, a human curator needs to extract and summarize the semantic concepts of the literature, find a GO term that is semantically close to the concepts, and assign that GO term to the gene product. To identify the potential matching GO terms for each topic, the MI values for all observed topic-GO associations were calculated. Then, for each topic *t*, a GO term from each of the three GO categories with the highest MI value was treated as the candidate GO term matching the topic. Table 2 shows examples of associating the extracted semantic topics with the GO terms. The top 9 rows are the topic-GO associations with high MI values, while the bottom 2 rows are examples of topic-GO associations with low MI. When MI values for topic-GO associations were high, the definitions of the GO terms usually agreed well with the semantic concepts contained in the latent topics. Interestingly, the inference of the topics by the LDA model mimics the process of identifying the biologic concepts from the texts by a human curator; and determining the MI ("the strength") of topic-GO association mimics the process of mapping the biological concepts to the GO terms. Thus, mapping latent topics to GO terms potentially provides a means to automatically annotate a protein with GO terms based on the semantic concepts contained in the associated literatures.

Clustering proteins according to their functional descriptions

In a topic that strongly related to a specific biological object or process, i.e., when MI of topic-GO association was high, the names of the proteins involved in that process frequently appeared on the top of the word list for the topics. For example, topic # 156 in Table 2 is related to *threonine/serine phosphorylation* process, and the protein names 'pkc,' 'akt,' and 'ste20' were among the most frequent words of the topic, indicating that the LDA model was capable of clustering gene/protein names according to the concept of protein functions. Interestingly, clustering of these protein names did not require them to co-occur within the same documents. The LDA model was capable of clustering the gene/protein names simply based on their associations with some common key words of the biological concepts. This finding could be used as a tool to cluster genes with similar functions from different organisms based on their associated literatures. This finding also agrees with a previous study by Homayouni et al [10], in which proteins were represented as points in the vocabulary space based on their associated literature, and they were further projected onto a reduced-dimension semantic space constructed with the LSI techniques. The proteins with similar functions were form clusters within semantic space.

Discussion

Most biomedical knowledge is stored as free text in the biomedical literature, and the size of the biomedical literature is increasing rapidly. There is an urgent need for automatically acquiring and representing this body of knowledge in a computable form to facilitate the discovery of new knowledge, which requires the development of computational methods to extract knowledge from the text. The current state of the art of the text mining approaches have applied to biomedical literature and reported in several recent challenge evaluations, such as the KDD, the BioCreative, and the TREC [7,9,16]. However, most of these approaches are within the conven-

Table 2: Examples of topic-GO associations

Topic #	GO ID	MI	GO Category	GO Term	Most Frequent Topic Words
278	GO:0005730	0.001439	Component	nucleolus	ribosom rrna pre deplet process small nucleolar biogenesi accumul nucleolu
267	GO:0005681	0.001193	Component	spliceosome complex	splice altern pre snrnp mrna spliceosom u2 step sap snrna
105	GO:0005816	0.00119	Component	spindle pole body	microtubul spindl mitot tubulin kinetochor mitosi centrosom pole centromer bodi
236	GO:0006935	0.00186	Process	chemotaxis	lymphocyt macrophag chemokin monocyt neutrophil inflammatori leukocyt peripher mcp cd8
156	GO:0006468	0.001514	Process	protein amino acid phosphorylation	kinas phosphoryl serin threonin pkc autophosphoryl casein akt catalyt ste20
267	GO:0000398	0.001404	Process	nuclear mRNA splicing	splice altern pre snrnp mrna spliceosom u2 step sap snrna
156	GO:0004674	0.001148	Function	protein serine/threonine kinase activity	kinas phosphoryl serin threonin pkc autophosphoryl casein akt catalyt ste20
267	GO:0008248	0.001463	Function	pre-mRNA splicing factor activity	splice altern pre snrnp mrna spliceosom u2 step sap snrna
236	GO:0008009	0.001093	Function	chemokine activity	lymphocyt macrophag chemokin monocyt neutrophil inflammatori leukocyt peripher mcp cd8
224	GO:0015671	5.05E-06	Process	oxygen transport	ha uniqu characterist featur extens character typic possess unusu exhibit
227	GO:0015213	5.00E-06	Function	uridine transporter activity	function defin unknown perform wide thei tissu repres consist creat

tional NLP, IE, and IR framework, and the application of probabilistic or non-probabilistic semantic modeling of biomedical literature remains relatively sparse [10-12].

In this paper, we report the extraction of a set of semantic topics from a corpus of protein-related MEDLINE titles and abstracts with the LDA model. The key advantages of applying an LDA model to perform statistical semantic analysis includes, but is not necessarily limited to the following: (1) it model is capable of extracting major recurring themes from a corpus of text in a unsupervised manner; (2) the assumption that a document is a mixture of topics naturally simulates real world text and allows modeling of text at finer granularity; and (3) it can effectively resolve many ambiguities commonly association with natural language.

Recurring biological themes reflect knowledge domains

The LDA model identifies topics from a text corpus by capturing the covariance of the words and organizes the words that tend to co-occur into a structure that mimics a topic. The inference algorithm for the model is unsupervised, precluding the need of expensive, manually-annotated data. The generative nature of the LDA model ensures that the extracted topics/concepts reflect the recurring themes within the corpus. We used a well-annotated data set from the Uniprot database [17], thus the major topics identified from the corpus arguably reflect the major domains of our knowledge of proteins.

We applied a Bayesian model selection approach to determine the "optimal" number of topics for the purpose of model fitting. The Bayesian model selection favors the simplest model that explains data well [18]. With such a preference, many of the 300 topics in our results reflect the general themes of the corpus. However, the model is also capable of capturing strong co-occurrence patterns that are highly specific biological objects/concepts, as demonstrated in Table 2. As more training data become available, especially as full electronic texts of the biomedical literature become available, the Bayesian model selection can accommodate more complex models thus simulating the data with finer granularity. One limitation of the LDA model is that it requires a specified number of topics in order to model the data. However, it is a strong assumption to specify that a corpus is generated with a fixed number of topics, which may not be valid in the real world. To address this issue, recent development in the nonparametric approaches, such as the Dirichlet process based methods may be more reasonable to model the data without a specified number of topics, such as in the Dirichlet process related models [19-21].

In the LDA model, a topic is represented as a distribution reflecting the word-usage pattern. One key advantage of the LDA model is that the extracted topics correspond to real world objects or concepts that are readily understandable by people with domain knowledge. In comparison, another extensively studied semantic analysis approach, the latent semantic indexing (LSI) model [10,12,22-24], cannot recover understandable semantic topics from text. The LSI model also captures the covariance of the words from a collection of text and identifies the major directions of the covariance space. It applies the singular value decomposition (SVD) approach to identify the orthogonal directions of semantic space spanned by the word vec-

tors of the documents and uses major directions to represent the semantic space with a reduced rank. Thus, a document can be represented as a vector in a reduced-rank space spanned by few major directions – a process of indexing the document with respect to semantic directions. However, restricting the semantic directions to be orthogonal to each other, the LSI identifies the directions that may not correspond to any human-understandable topics, thus remaining "latent."

Semantic analysis and automatic indexing

As shown in Figure 3, the LDA model can be used to extract semantic contents of an abstract, indicating that the model should be useful for automatic document indexing and information retrieval. In comparison to conventional information retrieval by keyword indexing, semantic indexing by LSI has been demonstrated to be more accurate [5] due to the fact that semantic indexing allows retrieval of documents whose semantic contents align well with the semantic meanings of the query terms, without requiring occurrence of the exact query terms in the documents. Although not yet tested on as large a scale as the LSI, the LDA model should have similar indexing power due to the fact that the semantic concepts extracted by the LDA aligns well with human perception.

We have shown that many of the topics extracted by the LDA model can be mapped to the controlled vocabulary of GO terms, potentially serving as a means of automatically annotating a protein-related corpus. Currently, most GO annotations are manually performed by PhD level biologists at different centers of GO consortium. Although accurate and specific, manual annotation is labor-intensive and cannot be expected to keep up with the pace of growth in the biomedical literature. Automatic annotation of proteins based upon the biomedical literature is a growing and urgent task facing the bioinformatics community that motivated the specific tasks in the recent competitive evaluations [7,9,16]. Our results indicate that it is possible to extract salient biological concepts from a large amount of biomedical literature and map the concepts to the controlled vocabulary. Although the mapping between the latent topics from the LDA model to the GO terms may not provide annotations as specific as manual annotations, automatic annotation based on the LDA should provide general and consistent descriptions of a protein

Dealing with ambiguities of natural language

Human natural language is full of ambiguities confounding the results of contemporary NLP, IE, and IR techniques [3,4]. Most noticeably, the phenomena of polysemy and synonym need to be effectively addressed during NLP, IE, and IR. The LDA effectively handles the uncertainties and ambiguities caused by the polysemes and synonyms due to its probabilistic representation of the topics. The distributional representation of concepts allows the synonyms to be group into a common topic, while a polyseme can participate in multiple concepts. Such representation effectively captures the key relationship between the words and semantic concepts: the concept is conveyed by choice of words and sense of a word is dependent on context. The inference algorithm of the LDA model explicitly utilizes such relationships to infer the topic for a word, so that the semantic topics of synonyms and polysemes can be assigned based on the context of text. This capability makes the LDA model a powerful tool to enhance the performance of other NLP, IE and IR techniques for text mining. The result shown in Figure 3 serves as a good example of the capability of the LDA model to properly assign words to topics depending upon context. Note that the words "space" and "induce" are general words that fit into different semantic context, and the LDA algorithm correctly associated them with the concepts of *mitochondria* and *apoptosis*, respectively, based on the semantic context of the document.

Conclusion

In summary, we extracted a set of major semantic concepts from a protein-related corpus of text words from MEDLINE titles and abstracts by applying the LDA model. The identified concepts are semantically coherent, and most of them are biologically relevant. The extracted biological topics reflect the major knowledge domains of current knowledge of protein function contained in the corpus. The semantic content of a document can be inferred from a text and used for automatically indexing the text. Future directions will be explored to extend the current approach or to develop new techniques for extracting biological concepts of finer granularity and combining semantic analyses with conventional NLP, IE, and IR techniques to map the topics to the controlled vocabulary.

Methods

Data set

The protein annotation data of the Uniprot database (Version 22, October 2004) was downloaded from the GOA project [2] web site of the European Bioinformatics Institute. In this data set, each protein was annotated with one or more GO annotations. Many annotation entries contained references to PubMed identification (PMID) numbers, presumably these annotations resulted from reading the literature indexed by the PMID. All the PMIDs and their associated GO terms were extracted from the Uniprot data set. The extracted data contained 6,565 unique GO accession numbers (GOID) and 25,005 unique PMIDs. The MEDLINE entries indexed by these PMID were downloaded from the National Center for Biotechnology Information (NCBI) using the Entrez E-utility

service, and their titles and abstracts were extracted. These MEDLINE text data were preprocessed as follows: (1) common words from a standard English "stop words" list were removed; (2) words were stemmed using Porter's stemmer [25]; (3) words that appeared fewer than 5 times in the corpus were discarded. The processed data set is referred to as GOA corpus and contained the preprocessed MEDLINE text words and associated GO annotations. After preprocessing, the vocabulary of the corpus consisted of 25,143 unique terms.

LDA model

Model specification

The LDA model is a probabilistic topic model [14,15,26]. It is a hierarchical generative model that simulates the process of writing a text. Let the corpus $C = \{d_1, d_2, ..., d_D\}$ be a set of documents, where D denotes the number of documents in the corpus; a document $d = (w_1, w_2,..., w_{Nd})$ consists of a sequence of words; and w be a word that takes a value from the vocabulary $\{v_1, v_2, ..., v_V\}$. Let T be the number of topics of a LDA model and V be the size of the vocabulary of the corpus. The LDA model simulates the generation of a document with following stochastic processes:

• For each document, sample a topic proportion vector $\theta = (\theta_1, \theta_2,..., \theta_T)'$ from a Dirichlet distribution with parameter α: $\theta \sim Dir(\theta \mid \alpha)$. This is equivalent to an author deciding what topics to include in the paper.

• For each word in the document, sample a topic z according to multinomial distribution governed by θ: $z \sim Multi(z \mid \theta)$. This can be thought as assigning a word to a topic.

• Conditioning on z, sample a word w according multinomial distribution with parameter ϕ_z: $w \sim Multi(w \mid \phi_z, z)$. This corresponds to picking words to represent the concept.

• The parameter ϕ_t with $t \in \{1,2,...,T\}$, is a V-dimension vector that defines the multinomial word distribution of a topic. It is distributed as Dirichlet with parameter β: $\phi_t \sim Dir(\phi_t \mid \beta)$.

The probabilistic directed acyclic graphical representation of the LDA model is shown in Figure 5 in plate notation [27]. In a probabilistic graph, nodes represent random variables and edges represent the probabilistic relationship, i.e., the conditional probability, between the variables. The shaded and un-shaded nodes represent the observed and unobserved variables, respectively. Each rectangular plate represents a replica of the data structure; the number at the bottom right of each plate indicates the number of the replicates. In this graph, each document is associated with a topic composition variable θ and total of N_d replicates of topic variable z and word w. The graph also shows that there are T topic word distributions.

Statistical learning

Given the observed documents, the learning task is to infer the topic-composition θ_d for each document; the topic variable, z_i, for each word,w_i, within the document; and the word distribution ϕ_t for each topic t. The exact inference of these unobserved variables is intractable. A Markov chain Monte Carlo (MCMC) [28] inference algorithm by Griffiths and Steyvers [15] was adopted to perform approximate inference. Let $\mathbf{z}$ denote a vector of the instances of all latent topic variables and $\mathbf{w}$ denote a vector of all the observed words of the corpus. The algorithm concentrates on the joint probability $p(\mathbf{w}, \mathbf{z})$ and applies Gibbs sampling to instantiate the latent topic variable for each word. Gibbs sampling is a technique to generate samples from a complex posterior distribution $p(\mathbf{z} \mid \mathbf{w})$ by iteratively sampling and updating each component variable z_i according to the conditional distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{w})$, where $\mathbf{z}_{-i}$ denotes the current instantiation of all the latent topic variables except z_i, and $\mathbf{w}$ denotes the vector of all observed words of the corpus. The Gibbs sampling procedure follows these steps: (1) randomly initialize the latent variables $\mathbf{z}$; (2) each element z_i of $\mathbf{z}$ is iteratively sampled and updated; (3) repeat step (2) until the Markov chain converges to the target posterior distribution $p(\mathbf{z} \mid \mathbf{w})$ ("burn in"); and (4) samples of $\mathbf{z}$ are collected from the Markov chain. The conditional distribution $p(z_i \mid \mathbf{z}_{-i}, \mathbf{w})$ is defined as follows:

$$p\left(z_i = j \middle| z_{-i}, w\right) \propto \frac{n_{-ij}^{(w_i)} + \beta}{n_{-ij}^{(\cdot)} + V\beta} \times \frac{n_{-ij}^{(d_i)} + \alpha}{n_{-i}^{(\cdot)} + T\alpha} \qquad (1)$$

In equation (1), $n_{-ij}^{(w_i)}$ denotes the count of the words in the corpus that are indexed by w_i and assigned to the topic j, excluding the word w_i; $n_{-ij}^{(\cdot)}$ is the count of all words assigned to the topic j, excluding the word w_i; $n_{-ij}^{(d_i)}$ is the count of words assigned to the topic j in document d_i that contains topic variable z_i, excluding w_i; $n_{-i}^{(d_i)}$ stands for the count of all the words in that document excluding w_i; and α, β, V and T were defined previously. During training of the LDA model, the values for the corpus level parameters were set as follows: $\alpha = 1$, $\beta = 0.1$.

Equation (1) has an intuitive explanation for how the inference algorithm determines the topic label z_i for the observed word w_i. The first term on the right side indicates the likelihood of observing word w_i if its topic $z_i = j$, e.g.,

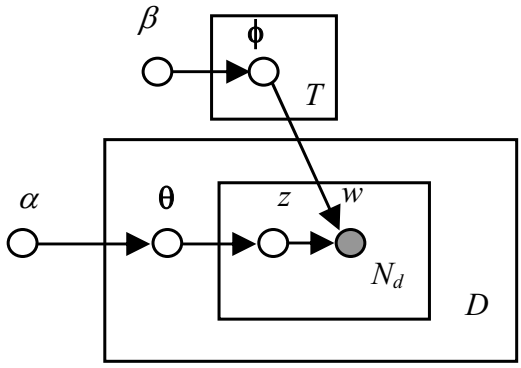

Figure 5
A directed acyclic graphical representation of the LDA model in plate notation.

the likelihood of observing word "death " if the topic is *apoptosis*. The second term specifies the likelihood that a word in the document belongs to topic *j*, based on the context of the document. In plain English, the second term would read: "the word w_i more likely belongs to topic *j* if many other words in the document belong to the topics *j*." For example, the word "death" is more likely to belong to the topic *apoptosis*, if there are other words in the document, such as "apoptosis," "programmed," and "cell," belonging to the same topic.

Once the vector of the latent topics **z** is instantiated by sampling, the parameters governing the posterior distribution of θ and ϕ can be estimated analytically as follows:

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha} \qquad (2)$$

$$\hat{\phi}_j^{(v)} = \frac{n_j^{v} + \beta}{n_j^{(\cdot)} + V\beta} \qquad (3)$$

where n_j^d is the number of words assigned the topic *j* in the document *d*; $n_\cdot^{(d)}$ is the total number of words in the document *d*; $n_j^{(v)}$ stands for number of times a word indexed by *v* belongs to the topic *j*; and $n_j^{(\cdot)}$ denotes total number of words assigned to the topic *j*.

Inference for new data

A trained model can be used to infer the latent topic variables **z** and estimate θ_d for a newly observed document. This is achieved by sampling **z** from the posterior distribution with MCMC by invoking Equation (1). During the sampling, the first term of equation (1) is replaced with the previously learned $\hat{\phi}$ from equation (3), and only the counts in the second terms are updated.

Model Selection

One objective of model training is to allow the model to fit the data well while avoiding over fitting. From a statistical learning point of view, this is a model selection problem that can be addressed within a Bayesian model selection framework to select the optimal model $\hat{M}$ that has the highest posterior probability conditioning on the observed data **w** as follows:

$$\hat{M} = \arg\max_M p(M \mid w), \qquad (4)$$

$$p(M \mid w) = \frac{p(w \mid M)p(M)}{p(w)} \qquad (5)$$

Assuming an uninformative prior distribution $p(M)$ for the models, the model selection was determined by the evidence (marginal likelihood) $p(\mathbf{w} \mid M)$, which can calculated by integrating out the latent parameters and variables:

$$p(w \mid M) = \int_\varphi \int_\theta \sum_z p(w, z, \phi, \theta \mid M) d\theta d\phi. \qquad (6)$$

The summation and integration in the equation (6) was intractable. Instead, a Monte Carlo approximation for this quantity was employed [15]. With the parameters α and β fixed, the difference between the model M_l and M_k is the number of the topics T_l and T_k. For a model with a given number of topics, *T*, the evidence $p(\mathbf{w} \mid M)$ was approximated as follows: 40 samples of latent variable vectors, $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{40}\}$, were collected from 4 randomly initialized Markov chains according equation (1). Then, the conditional probability $p(\mathbf{w} \mid \mathbf{z}, M)$ for each sample **z** was evaluated by analytically integrating out ϕ:

$$p(w \mid z, M) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \right)^T \prod_{j=1}^{T} \frac{\prod_{i=1}^{v} \Gamma(n_j^{(i)} + \beta)}{\Gamma(n_j^{(\cdot)} + V\beta)} \qquad (7)$$

The evidence $p(\mathbf{w} \mid M)$ was approximated with the harmonic means of the sample conditional probabilities $p(\mathbf{w} \mid \mathbf{z}, M)$ [29]. The selection among the models with differ-

ent T was carried out based on the approximated evidence.

Mutual information

MI is a symmetric, non-negative quantity that measures the amount of information one variable contains with respect to another variable, and it equals zero if and only if the variables are independent. The MI between a latent topic and a GO term was calculated as follows:

$$I(A_g, L_t) = \sum_{A_g, L_t} p(A_g, L_t) \log \frac{p(A_{g,}, L_t)}{p(A_g) p(L_t)} \qquad (8)$$

where $I(A_{g,} L_t)$ is the mutual information between the annotation of a word with GO term g and labeling the word with topic t; A_g and L_t are binary variables indicating whether a word is annotated with the GO term g and assigned to the topic t, respectively. The topic labeling of a word was determined according to the inferred latent variable samples **z**. We specified that each word within a given document was annotated with a GO term g if the document was annotated with the term g. Note that this is a strong assumption, which may skew the MI value for some uncommon topics. The joint and marginal probabilities in equation (8) were estimated empirically by counting the events

Authors' contributions

BZ performed data collection, processing and model training experiments. DCM carried out results evaluation. XL conceived, directed the study and implemented the LDA inference program.

Acknowledgements

XL is partially supported by the Medical University of South Carolina cardiovascular COBRE grant from NIH/NCRR (5 P20 RR016434-04) and NIH/NLM 5T15LM007438-03. DCM is supported by the NLM training grant 5T15-LM007438-02. The authors would like to thank Drs. Mark Steyvers, Alan Aronson, Chengxiang Zhai, and anonymous reviewers for their discussions and suggestions.

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25(1):**25-29.
2. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R: **The Gene Ontology Annotation (GOA) Database – an integrated resource of GO annotations to the UniProt Knowledgebase.** In *Silico Biol* 2004, **4(1):**5-6.
3. Manning CD, Schutze H: **Foundation of statistical natural language processing.** Cambridge, MA: MIT Press; 1999.
4. Jurafsky D, Martin JH: **Speech and language processing.** Upper Saddle River, NJ: Prentice Hall; 2000.
5. Baeza-Yates R, Ribeiro-Neto B: **Modern Information Retrieval.** Pearson Education Limited and ACM Press; 1999.
6. Hirschman L, Park JC, Tsujii J, Wong L, Wu CH: **Accomplishments and challenges in literature data mining for biology.** *Bioinformatics* 2002, **18(12):**1553-1561.
7. Hersh WR, Bhuptiraju RT, Ross L, Johnson P, Cohen AM, Kreamer DF: **TREC 2004 genomics track overview.** *Text Retrieval Conference (TREC) 2004* 2004.
8. Krallinger M, Valencia A: **Text-mining and information-retrieval services for molecular biology.** *Genome Biol* 2005, **6(7):**224.
9. Hirschman L, Yeh A, Blaschke C, Valencia A: **Overview of BioCreAtIvE: critical assessment of information extraction for biology.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S1.
10. Homayouni R, Heinrich K, Wei L, Berry MW: **Gene clustering by latent semantic indexing of MEDLINE abstracts.** *Bioinformatics* 2005, **21(1):**104-115.
11. Raychaudhuri S, Altman RB: **A literature-based method for assessing the functional coherence of a gene group.** *Bioinformatics* 2003, **19(3):**396-401.
12. Khatri P, Done B, Rao A, Done A, Draghici S: **A semantic analysis of the annotations of the human genome.** *Bioinformatics* 2005, **21(16):**3416-3421.
13. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
14. Blei DM, Ng AY, Jordan MI: **Latent Dirichlet Allocation.** *Journal of Machine Learning Research* 2003, **3:**993-1022.
15. Griffiths TL, Steyvers M: **Finding scientific topics.** *Proc Natl Acad Sci U S A* 2004, **101(Suppl 1):**5228-5235.
16. Yeh AS, Hirschman L, Morgan AA: **Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup.** *Bioinformatics* 2003, **19(Suppl 1):**i331-339.
17. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LS: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2005, **33(Database):**D154-159.
18. MacKay DJC: **Information theory, inference and learning algorithms.** Cambridge, UK: Cambridge University Press; 2003.
19. Teh YW, Jordan MI, Beal MJ, Blei DM: **Hierarchical Dirichlet Processes.** *Advances in Neural Information Processing Systems (NIPS) 17: 2005* 2005.
20. Yu K, Yu S, Tresp V: **Dirichlet enhanced latent semantic analysis.** *Workshop on Artificial Intelligence and Statistics AISTAT 2005* 2005.
21. Blei DM, Jordan MI: **Variational methods for the Dirichlet process.** *Proceedings of the 21st International Conference on Machine Learning (ICML): 2004* 2004.
22. Deerwester S, Dumais ST, Landauer TK, Furnas GW, Harshman RA: **Indexing by latent semantic analysis.** *J Am Soc Inf Sci* 1990, **41:**391-407.
23. Berry MW, Drmac Z, Jessup ER: **matrices, vector spaces, and information retrieval.** *SIAM Review* 1999, **41(2):**335-362.
24. Ding CHQ: **A Probabilistic Model for Latent Semantic Indexing.** *J Am Soc Inf Sci Tech* 2005, **56:**.
25. Porter MF: **An algorithm for suffix stripping.** *Program* 1980, **14(3):**130-137.
26. Hofmann T: **Probabilistic Latent Semantic Indexing.** *the 22nd International Conference on Research and Development in Information Retrieval (SIGIR'99):1999* 1999.
27. Buntine W: **Operations for learning with graphical models.** *Journal of Artifical Intelligence Research* 1994, **3:**993.
28. Andrieu C, Freitas Nd, Doucet A, Jordan MI: **An Introduction to MCMC for Machine Learning.** *Machine Learning* 2003, **50(1–2):**5-43.
29. Kass RE, Raftery AE: **Bayes Factors.** *J Am Stat Assoc* 1995, **90:**773-795.