BMC Bioinformatics

Research article

Statistical modeling of biomedical corpora: mining the Caenorhabditis Genetic Center Bibliography for genes related to life span

DM Blei*[1], K Franks[2], MI Jordan*[3,4] and IS Mian*[2]

Address: [1]Computer Science Department, Princeton University, Princeton, New Jersey 08540 USA , [2]Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720-8265, USA, [3]Department of Statistics, University of California Berkeley, Berkeley, California 94720, USA and [4]Department of EECS, University of California Berkeley, Berkeley, California 94720, USA

Email: DM Blei* - blei@cs.princeton.edu; K Franks - KFranks@lbl.gov; MI Jordan* - jordan@cs.berkeley.edu; IS Mian* - smian@lbl.gov

* Corresponding authors

Published: 08 May 2006

BMC Bioinformatics 2006, **7**:250 doi:10.1186/1471-2105-7-250

Received: 19 July 2005
Accepted: 08 May 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/250

Abstract

Background: The statistical modeling of biomedical corpora could yield integrated, coarse-to-fine views of biological phenomena that complement discoveries made from analysis of molecular sequence and profiling data. Here, the potential of such modeling is demonstrated by examining the 5,225 free-text items in the *Caenorhabditis* Genetic Center (CGC) Bibliography using techniques from statistical information retrieval. Items in the CGC biomedical text corpus were modeled using the Latent Dirichlet Allocation (LDA) model. LDA is a hierarchical Bayesian model which represents a document as a random mixture over latent topics; each topic is characterized by a distribution over words.

Results: An LDA model estimated from CGC items had better predictive performance than two standard models (unigram and mixture of unigrams) trained using the same data. To illustrate the practical utility of LDA models of biomedical corpora, a trained CGC LDA model was used for a retrospective study of nematode genes known to be associated with life span modification. Corpus-, document-, and word-level LDA parameters were combined with terms from the Gene Ontology to enhance the explanatory value of the CGC LDA model, and to suggest additional candidates for age-related genes. A novel, pairwise document similarity measure based on the posterior distribution on the topic simplex was formulated and used to search the CGC database for "homologs" of a "query" document discussing the life span-modifying *clk-2* gene. Inspection of these document homologs enabled and facilitated the production of hypotheses about the function and role of *clk*-2.

Conclusion: Like other graphical models for genetic, genomic and other types of biological data, LDA provides a method for extracting unanticipated insights and generating predictions amenable to subsequent experimental validation.

Background

In the design, analysis and interpretation of experiments, biomedical and clinical researchers encounter the problem of evaluating and summarizing prior knowledge on

the subject under investigation. Traditional solutions include examining articles in scientific journals, primarily via PubMed access to MEDLINE, and interrogating WWW-based sources such as Entrez Gene [1] and Online Mendelian Inheritance in Man (OMIM) [2]. Consider a scenario in which a researcher seeks enhanced knowledge about a protein implicated in aging. Typically, the steps involved in addressing this problem include interrogating structured data resources: searching protein sequence databases to identify homologs in other species, querying warehouses of genomic information to determine key non-coding regions and polymorphisms, examining collections of high-throughput molecular profiling data sets to ascertain genes with similar patterns of expression, probing ontologies such as the Gene Ontology (GO) [3] to uncover other genes with similar patterns of annotation, and so on. The entire procedure is accompanied by examination of the literature to determine, for example, classes of proteins mentioned in the same article as the putative gerontogene.

Advanced techniques and sophisticated tools for interacting with structured data are well known, widely available and include BLAST [4] for sequence databases, Ensembl [5] and the UCSC Browser [6] for genomes, GEO [7] for transcript profiles, and tools available from the GO that allow navigation of terms in the ontology. This is less true for the scientific literature. Given the time-consuming yet critical importance of synthesizing information in text corpora such as MEDLINE, the problem of making data interpretation a more systematic and automated endeavor is emerging as an important topic of research (see, for example, [8-10]). The development of strategies capable of providing a user the ability to assimilate and act upon information present in resources of structured and unstructured data remains an important goal.

The primary aim of biomedical text mining is the systematic analysis of document collections such as MEDLINE abstracts and full-text journal articles with the goal of generating useful and unanticipated scientific discoveries (for recent reviews of current methods and illustrative applications, see [11-13]). Examples of tasks addressed by text mining methods include identifying literature relevant to specific molecules, finding associations between genes and diseases, determining putative functions for proteins, and predicting regulatory networks.

A common approach to text mining is to treat the problem as one of natural language processing (NLP) [14]. NLP methods concentrate on the linguistic structure of documents and make explicit use of syntactic, relational, and ontological knowledge. In biology, such approaches [15] have been employed for information extraction: the task of ascertaining facts, relations, and entities in unstructured written language such as protein-protein interactions, protein subcellular location, and gene names. Tools based on these ideas include Textpresso [16] and Telemakus [17]. Elsewhere, NLP has been used in conjunction with LocusLink and GO to compare OMIM to MEDLINE [18].

Recently, nouns extracted from MEDLINE abstracts tagged with parts of speech were combined with knowledge from other sources, Principal Component Analysis, and means linkage clustering to find associations between genes and phenotypes [19]. In general, NLP can be effective in circumscribed domains where linguistic knowledge is available and the terminology evolves slowly, is consistent, largely unambiguous, and relatively simple. However, the paucity and incomplete nature of such information for biomedical corpora suggests that the full potential of text mining in biology remains unrealized.

An alternative to NLP is to frame the problem from the perspective of information retrieval (IR) [20]. Statistical IR methods explore large quantities of information and often involve capabilities for clustering, classifying, categorizing, summarizing, and detecting novel, similar and relevant objects. The most successful testaments to the real-world utility of IR techniques are Internet search engines. Thus, statistical IR models of biological document collections could reveal rich, complex, and previously unappreciated relationships. Such results would complement insights derived from analysis of molecular profiling, protein-protein interaction, gene knock-out, and similar types of data. With systematic deployment of IR tools, the interrogation of biomedical corpora could become as routine and indispensable a part of research as analyzing genomic and genetic data is today. The analogy is more than superficial and extends to the direct use of IR techniques such as singular value decomposition in bioinformatics (see for example [21]). Thus, the common mathematical foundations for algorithms that underpin IR and genome analysis make it possible to envision integrated procedures that combine primary biological data with biological corpora.

This work describes an application of statistical IR methodology to the analysis of a biomedical text corpus, the *Caenorhabditis* Genetic Center (CGC) Bibliography (Figure 1). The specific model at the heart of this study is the Latent Dirichlet Allocation (LDA) model [22], a hierarchical Bayesian model employed previously to analyze text corpora and to annotate images [23]. Recently, LDA has been used to extract and analyze the topics present in a document corpus consisting of articles published in the journal *Proceedings of the National Academy of Sciences* [24].

```
     Key: 4951
 Medline: 11696330
 Authors: Lim CS;Mian IS;Dernburg AF;Campisi J
   Title: C. elegans clk-2, a gene that limits life span, encodes a
          telomere length regulator similar to yeast telomere binding
          protein Tel2p.
Citation: Current Biology 11: 1706-1710 2001
    Type: ARTICLE
   Genes: clk-2
Abstract: An important quest in modern biology is to identify genes
          involved in aging. Model organisms such as the nematode
          Caenorhabditis elegans are particularly useful in this
          regard. The C. elegans genome has been sequenced [1], and
          single gene mutations that extend adult life span have been
          identified [2]. Among these longevity-controlling loci are
          four apparently unrelated genes that belong to the clk
          family [3-5]. In mammals, telomere length and structure can
          influence cellular, and possibly organismal, aging [6].
          Here, we show that clk-2 encodes a regulator of telomere
          length in C. elegans.
```

Figure 1
One of the 5,225 free-text items in the CGC Bibliography in its original form.

In general, IR methods assume that the order of words in a document can be neglected and view documents as "bags of words." The loss of information incurred by ignoring word order is offset by the ability to devise efficient computational algorithms that are viable for large corpora. Although there is no theoretical justification for casting a document in this manner, the practical benefits and utility of doing so are considerable. The LDA model considered here is a model for a corpus viewed as a collection of bags of words. It assumes that each word of each document is generated by one of several "topics"; each topic is associated with a different conditional distribution over a fixed vocabulary. The same set of topics is used to generate the entire set of documents in a collection but each document reflects these topics with different relative proportions. Thus, LDA is a mixture of mixtures model, *i.e.*, the mixture components are shared across all documents but each document exhibits different mixture proportions. As a generative probabilistic model, the LDA can handle unseen or novel data, *i.e.*, a document that was not one of the bag of words used to estimate the model.

The fundamental entities in LDA, random variables representing topics and words, are grouped together in such a way to form a corpus, *i.e.*, a group of groups of words. The hierarchical nature of the model stems from the fact that documents are modeled as probability distributions across topics, and topics are modeled as probability distributions across words. A notable virtue of LDA is that a given topic can occur with high probability in multiple documents, and that a given word can occur with high probability in multiple topics. Topics are treated as latent variables, namely entities that are not present explicitly in the data (a set of sequences of words), but are presumed to be present implicitly and are to be inferred by statistical analysis.

To analyze the CGC Bibliography, each item in the corpus was recast as a bag of words and the resultant data set of documents was used to estimate the parameters of three different statistical IR models. The predictive performance of the LDA model was better than that of two simpler bag of words models, a unigram model and a mixture of unigrams model, trained on the same data set. The potential of LDA in assisting biological studies was illustrated by considering the phenomenon of nematode aging. In order to illuminate the hidden factors permeating a corpus and captured by the topics discovered by a trained CGC LDA model, LDA topics were labeled via an automated process that assigned words from the CGC vocabulary (corpus-

based labels) and GO terms (ontology-based labels) to each topic. Examination of these labels indicated that the CGC topics captured meaningful and plausible facets of nematode biology. To investigate aging, topics whose corpus-based labels included many CGC words corresponding to the names of genes known to influence life span were identified. For the two topics with the greatest number of such CGC-based topic labels, novel candidates for age-related genes were equated with other CGC-based topic labels that corresponded to gene names (guilt-by-association). Finally, an LDA-based measure of pairwise document similarity was devised and used to address the problem of searching a database of documents to determine topic-space homologs of a query document. Inspection of the "document homologs" of the CGC item shown in Figure 1 resulted in enhanced understanding of the biology of the *clk-2* gene.

This work highlights the potential and utility of LDA in organizing and exploiting one type of widely available information resource, a collection of documents in the form of free or unstructured text. However, researchers are faced with a plethora of resources including images and structured data such as molecular sequences, transcript profiles, disease information, and so on. Thus, there is a compelling need for techniques and systems able to condense, integrate and present large amounts of disparate data to a user. This paper concludes with a discussion of how the family of probabilistic graphical models, of which LDA is a specific example, provides a framework for integrating heterogeneous data and thus meets this challenge.

Results

LDA outperforms mixture of unigrams, unigram and random models

In order to compare different models of text, a data set of *C. elegans* related documents was created. In particular, each CGC Bibliography free-text item was transformed into a bag of words yielding a corpus of $M = 5,225$ documents and a $V = 28,971$ word vocabulary.

The generalization performance of three statistical models was assessed: an LDA model (Figure 2), a mixture of unigrams model (right, Figure 3), and a "baseline" unigram model (left, Figure 3). A model was trained using 90% of the 5,225 documents in the CGC corpus and tested on the remaining 10%. LDA and mixture of unigrams models with $K = 5$, 10, 20, 50, and 100 latent topics were estimated; a single unigram model was estimated because such models harbor no notion of topic. The perplexity (inverse of the per-word likelihood) of the held-out test set of $J = 525$ documents (Equation 6) was computed for each trained model. Figure 4 shows the generalization performance of each model as a function of the number of latent topics. LDA has consistently smaller perplexity scores than the two extant models indicating better performance on unseen documents. Since the perplexity of 50- and 100-topic LDA's is low and similar, a latent space with 50 topics appears to provide a parsimonious description of the CGC corpus.

The ability of three specific models to retrieve a set of 842 aging-related documents in the collection of 5,225 CGC documents was assessed: a 50-topic LDA model estimated using all documents in the corpus, a 50-topic mixture of unigrams model estimated using all documents, and a model which ordered all documents randomly. For each model, an average precision/recall (PR) curve was constructed by computing the ranking of other documents given each aging-related document as a query, and the average F1 measure was computed. Figure 5 shows average PR curves for the three models. The average F1 measure (standard error) for the LDA model, the mixture of unigrams model and a random model is 0.30(2.86e – 06), 0.22(7.85e – 06), and 0.29(3.02e – 05) respectively. Although the F1 values for the LDA and random models are similar, the smaller standard error of the LDA model indicates its superiority to the random model. In addition, the average PR curves indicate that the LDA places more of the age-related documents higher up its rankings than the random model. Thus, of the three models investigated, LDA is best able to retrieve the set of related documents. Overfitting by the mixture of unigrams model results in a performance worse than the random model.

All subsequent discussion of an LDA model and/or a mixture of unigrams model pertain to a $K = 50$ topic model estimated using all $M = 5,225$ CGC documents in the corpus.

LDA latent topics embody concepts associated with nematode biology

A systematic strategy for clarifying the nature of the hidden factors permeating a corpus was devised and applied to a CGC LDA. Topic annotation (topic labeling) is defined as an automated process that creates a verbose (compact) description of an LDA topic. The method designed to annotate and label topics exploited the corpus-level parameter β (Figure 2). The $K \times V$ topic-word matrix β collates the multinomial distributions over the V words in the vocabulary that characterize the K topics. For a given LDA model of a particular corpus, the *kth* row specifies the topic-specific word distribution for topic k and an element, β_{kv}, denotes the likelihood of the vth word given the *kth* topic. For each of the $K = 50$ topics in the CGC LDA model, the $V = 28,971$ β_{kv} values were ordered and used to generate a word rank versus topic-specific word probability plot. In every case, the 500 top-ranked words accounted for most of the probability mass.

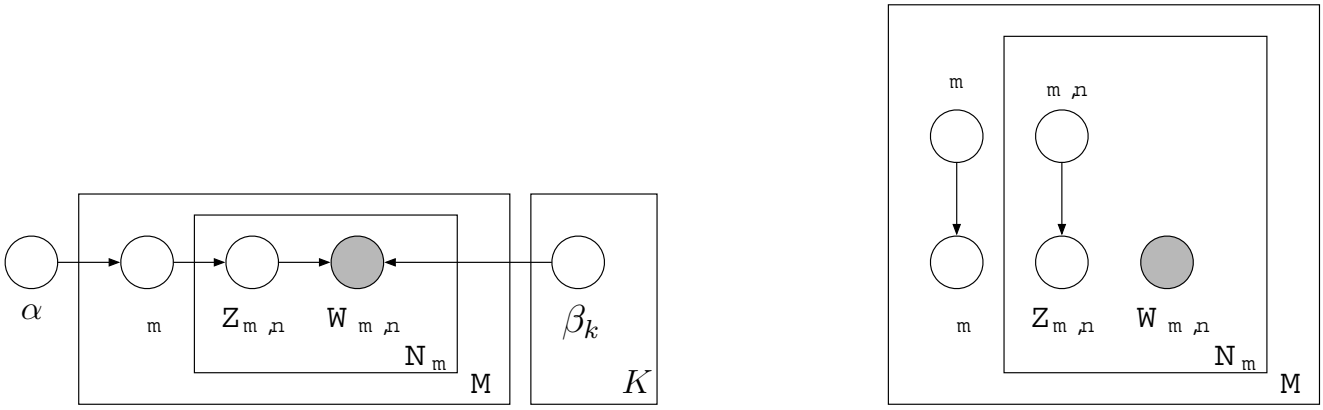


Figure 2
Graphical model representation of the LDA model (left) and the variational distribution used to approximate the posterior in LDA (right) [22]. LDA defines a distribution on a collection of documents in much the same manner that a profile hidden Markov model yields a distribution on a set of (biological) sequences [31]. The corpus depicted contains *M* documents and each is a sequence of *N* words. Open circles are parameters (α, β, γ, ϕ) or latent variables (θ, z). The shaded circle is the observed word variable (*w*) and boxes (plates) represent replicates. The Dirichlet parameter, α, and topic-word matrix, β, are corpus-level parameters sampled once in the process of generating a corpus. The topic proportions, θ, is a document-level variable sampled from αonce per document. The topic, z, is a word-level variable sampled from θ once for each word in a document. Formally, a *K*-topic LDA specifies a two-level probabilistic process that generates a document as follows, (i) a *K*-dimensional vector, θ, is chosen from the distribution $p(\theta|\alpha)$, and (ii) words are sampled repeatedly from the document-specific mixture distribution, $p(w|\theta)$. Exact inference and parameter estimation involve calculating the posterior distribution on a document $p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$. This is intractable because the latent variables are coupled via the edge between θ and z. The posterior can be approximated by computing the variational Dirichlet parameter γ and the variational multinomial parameter ϕ for each word in the document. The subscripts *m*, *n*, and *k* on a parameter (β, γ, ϕ) or variable (θ, z, *w)* donate the *m* th document, *nth* word and kth topic respectively. Note that the Dirichlet variable α is a distinct component of the probability model and not merely an expression of uncertainty about a parameter. This differs from profile hidden Markov models where a mixture of Dirichlet distributions is used as a prior for amino acid/nucleotide probability distributions.

Thus, these 500 high probability CGC words were designated topic annotation words (the same word from the *V*-wordvocabulary could annotate multiple topics).

Two different approaches were used to create labels for each topic. Corpus-based topic labels are topic annotation words that are unique to a topic and represent descriptors

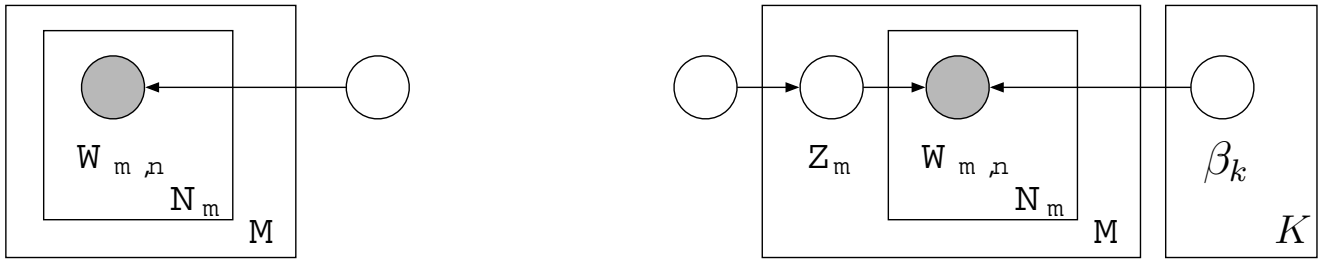


Figure 3
Graphical model representation of a mixture of unigrams model with *K* latent topics (right) and a unigram model (left). The corpus depicted contains *M* documents and each is a sequence of *N* words. Open circles represent latent variables (z)or parameters (β, θ). Each shaded circle is an observed word variable (*w*). Boxes (plates) represent replicates. The subscripts *m*, *n* abd *k* on a parameter (β, θ) or variable (z, *w*) donate the mth document, nth word and kth topic respectively.A mixture of unigrams generates all the words in a given document from exactly one topic, z. This differs from the LDA model where a single document can express multiple topics (Figure 2). Note that the naive Bayes model used to cluster transcript profiling data [41-43] has the same topology as the mixture of unigrams but the observed variables are continuous-valued expression measurements rather than discrete words.

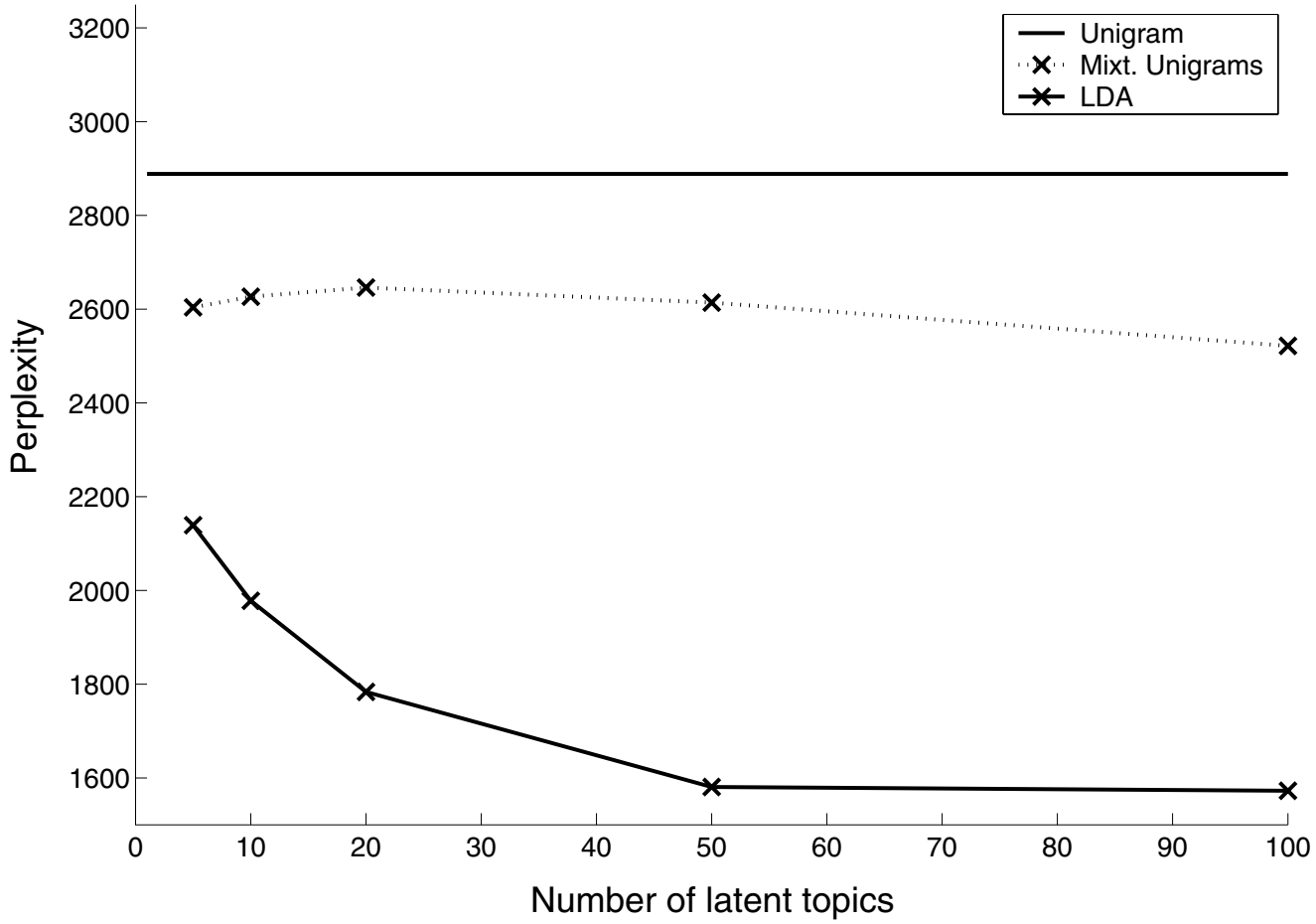


Figure 4
The perplexity of LDA, mixture of unigrams, and unigram models estimated and evaluated on the CGC corpus. The score of test documents is shown against the number of latent topics (the perplexity of the unigram is constant because this statistical model has no notion of latent topics).

applicable to only one topic. CGC-based topic labels were equated with topic annotation words that were not assigned to any of the other 49 topics, *i.e.*, the 50 sets of CGC-based topic labels formed disjoint subsets of words from the CGC vocabulary. Ontology-based topic labels are the outcome of filtering topic annotation words using an external knowledge source and represent descriptors applicable to one or more topic. The ontology exploited here was the GO. The relationship between GO controlled vocabulary terms can be depicted as a directed acyclic graph (DAG). Each node corresponds to a term from one of three aspects, for example, "exodeoxyribonuclease," "mitochondrial derivative" and "ethylene mediated signaling pathway" are exemplars of GO terms from the "Molecular Function," "Cellular Component" and "Biological Function" aspects respectively. The structure of the DAG underpinning the GO vocabulary defines semantic relationships amongst terms so that, for example, the node for the GO term "intracellular" is a parent of the node for the more specific GO term "nucleus." Recall that the topic annotation words for topic k are the 500 words from the CGC vocabulary that best characterize the topic. These 500 words were mapped to nodes in the GO DAG. A node where a topic annotation word coincided with a GO term was designated an explicit node. GO-based topic labels were equated with the GO terms for both explicit nodes and the children and grandchildren of explicit nodes.

Examination of the automatically generated CGC- and GO-based labels suggests that LDA topics capture meaningful and coherent facets of the molecular, cellular, and behavioral biology of *C. elegans*. Figure 6 shows results for four selected CGC topics (the results for all 50 topics are available in Additional file 1). The hidden factors permeating the CGC corpus include one pertaining to sexual

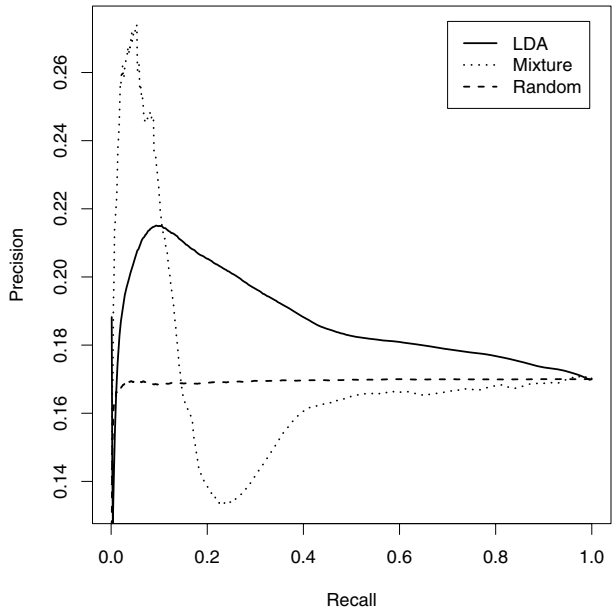


Figure 5

Precision/recall (PR) curves for three models of text (LDA, mixture of unigrams, random) and the task of retrieving a set of aging-related documents (842 CGC items that refer to one or more of the genes listed in Table 1). Precision is the fraction of documents in a list that are relevant (related to aging) whereas recall is the fraction of relevant documents in the list. For a desired level of recall, for example 70%, there is a corresponding precision. The graph shows average precision against average recall. Although each point is a mean of 842 pairs of precision and recall values, the standard error is negligible and so not depicted.

reproduction (Topic 6), chromosome structure and function (Topic 14), cell death (Topic 20) and locomotion (Topic 27).

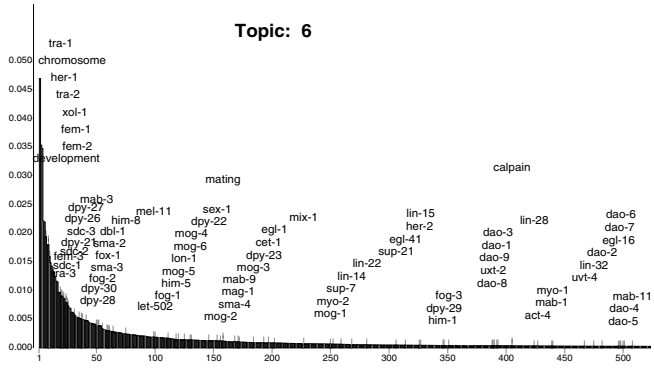
Ontology-based topic labels derived from structured knowledge for domains other than molecular and cellular biology are needed to clarify the nature of some CGC hidden factors. Figure 7 shows topics that have CGG- but no GO-based labels. Inspection of the CGC-based labels for Topics 3, 29, 38, and 41 suggest the presence of hidden factors that are concerned with scientific protocols and procedures that are independent of any biological question, and that allude to evolution.

Interrogation of LDA topics provides insights into genes influencing life span

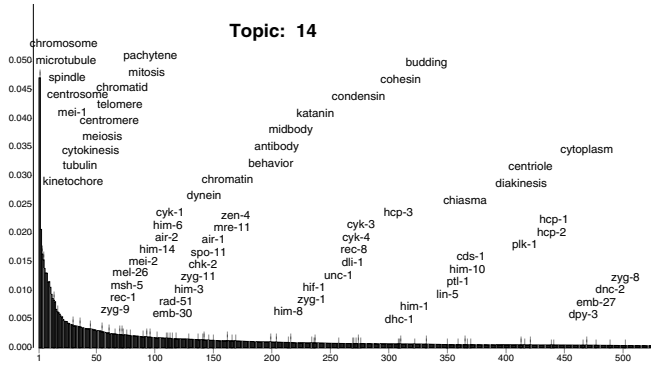
A guilt-by-association approach was devised to identify genes that may be involved in a phenomenon of interest and the procedure illustrated using genes implicated in modifying life span. A "gene word" is a word in the 28,971 word CGC vocabulary that corresponds to the name of a *C. elegans* gene. CGC-based topic labels are the ≤ 500 words in the vocabulary that best characterize a topic. If a number of the topic labels are gene words and most of the genes are known to be associated with a specific phenomenon, then the other gene words can be equated with genes likely to be involved in the same phenomenon. One factor influencing the biological insights that can be derived from this approach is the human curation component of the process used to create the CGC Bibliography, *i.e.*, the individual who defined the set of genes in the Genes record believed to be discussed in the Abstract record. In addition to limitations in the data used to estimate a statistical model of text, the LDA remains a model based on the simple bag of words representation of a document. While this LDA-based approach is not an automated method for formulating sophisticated and detailed hypotheses, it does highlight how a model that ignores syntax and semantics can organize information in a manner that provides a user the ability to exploit their background knowledge and enhance understanding of the subject in hand.

Table 1 lists the names of genes known to extend or shorten life span and designated aging-related gene words. Inspection of the two CGC LDA topics with the greatest number of CGC-based labels that are aging-related gene words suggests that *akt-1*, *akt-2* and *ges-1* may be associated with aging. Figure 8 shows these two topics, Topic 9 and Topic 12. In decreasing topic-specific word probability and with genes listed in Table 1 in bold, the Topic 9 CGC-based labels that are gene words are **age-1, clk-1, mev-1**, daf-18, **fer-15, clk-2, gro-1**, daf-23, akt-1, akt-2, **clk-3, pdk-1, rad-8**, sod-3, old-1, **ctl-1, tkr-1, daf-28**, sod-2, sir-2, daf-9, cln-3, ins-1, **age-2**, and **spe-10**. The genes depicted in a normal font have properties similar to known life span modifying genes such as dauer (daf) phenotypes (Table 1). For example, the Wormbase [25] annotations for *akt-1* and *akt-2* include "protein serine/threonine kinase" and "inhibition of both akt-1 and akt-2 leads to dauer-constitutive phenotype."

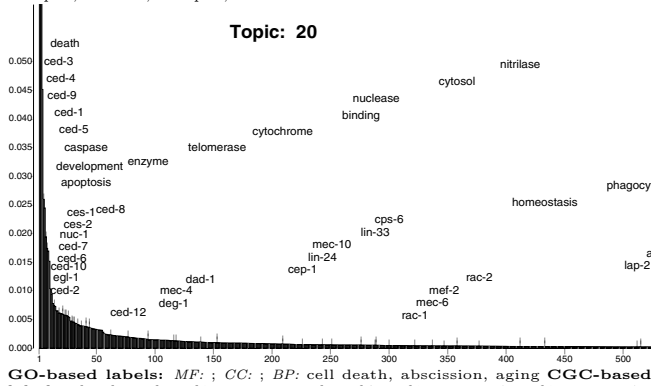
The mechanisms of action of putative gerontogenes suggested by different topics may not be identical. For Topic 12, the CGC-based labels that are gene words are ges-1, **osm-3, osm-5, che-2, osm-1**, lov-1, pkd-2, **daf-19, che-11**, unc-116, **che-13**, che-10, osm-10, che-1, che-14, pho-1, dyf-1, che-12, and pkd-1. *ges-1* is a gut-specific carboxylesterase, a molecular function not ascribed by Wormbase to life span modifying genes. Since many topic labels are associated with the osm phenotype, osmoregulation may be a feature that differentiates Topic 12 aging-related genes from those of Topic 9.



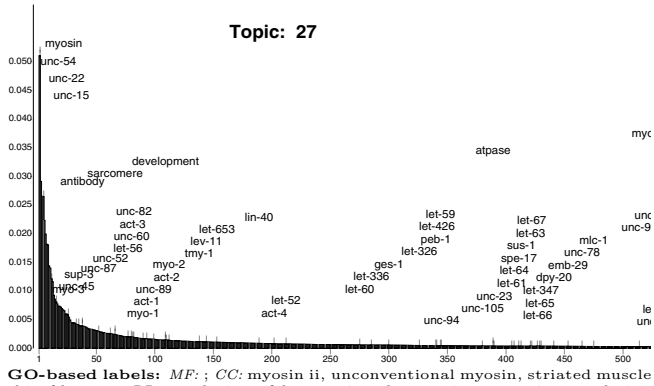
GO-based labels: *MF:* exodeoxyribonuclease; *CC:* ; *BP:* ethylene mediated signaling pathway, osmosensory signaling pathway via two-component system, two-component signal transduction system 'phosphorelay', negative regulation of transcription by pheromones, positive regulation of transcription by pheromones, regulation of transcription from pol i promoter mitotic, ethylene mediated signaling pathway 'induced systemic resistance', ethylene mediated signaling pathway 'jasmonic acid/ethylene dependent systemic resistance', mapkkk cascade 'osmolarity sensing', negative regulation of transcription from pol i promoter mitotic, positive regulation of transcription from pol i promoter mitotic, thermoregulation, fat body metabolism 'sensu insecta', fever, homoiothermy, negative regulation of transcription from pol ii promoter by pheromones, positive regulation of transcription from pol ii promoter by pheromones, abscission, aging, development of primary sexual characteristics, development of secondary sexual characteristics **CGC-based labels:** sex, tra1, dosage, her1, xx, compensation, sexual, xo, xol1, tra3, sdc1, fem, males, sdc2, dpy21, sdc3, dpy26, dpy27, sexe, mab3, dpy28, dpy30, hermaphrodites, fox1, sexspecific, dbl1, feminizing, fog, sexdetermination, copulatory, masculinization, sma, mel11, sexually, sdc, feminization, dimorphism, malespecific, hierarchy, masculinizing, mog5, transform, mog6, mog4, lf, dpy22, sex1, pp2c, mog2, sexes, mag1, mel, global, numerator, mog3, mndp8, stdp2, mndp10, dpy23, equalize, dsx, cet1, cpeb, coordinately, dimorphic, diploid, feminize, bmp, triploid, sexlinked, compensate, mix1, mndp9, mndp25, females, mndp27, bursa



GO-based labels: *MF:* ; *CC:* mitochondrial derivative, axonemal microtubule, centriole, telomerase holoenzyme complex, dynein; *BP:* adult feeding behavior, female meiosis i spindle assembly, female meiosis ii spindle assembly, meiotic spindle assembly, adult behavior **CGC-based labels:** air, mei1, cytokinesi, prophase, metaphase, kinetochore, furrow, anaphase, cytokinesis, centromere, aurora, crossover, chromatid, rec1, msh5, mel26, microtubules, mei2, zen, mitosis, him14, hcp, cyk, cohesion, air2, him6, cyk1, rad51, centrosomal, szt1, aster, him3, zyg, mei, holocentric, zyg11, midzone, chk2, mnt12, spo11, air1, synapsi, interphase, segregate, rec, mre11, breakage, zen4, mix, tau, pronuclear, chiasmata, hin1, depleted, crosslink, homozygote, promoting, poles, recombine, inversion, pronuclei, hif1, katanin, gammatubulin, dispensable, microtubuleassociated



GO-based labels: *MF:* ; *CC:* ; *BP:* cell death, abscission, aging **CGC-based labels:** death, ced, ced3, programmed, ced4, ced9, apoptosi, ced1, apoptotic, bcl2, ced5, caspase, ced2, ced10, engulfment, ced6, ced7, nuc1, corpse, ces2, cesces1, apoptosis, pcd, celldeath, ced8, ice, dying, phagocytosi, suicide, ced12, cebnip3, fhit, bh3, transglutaminase, engulfing, protect, p35, proapoptotic, dock180, cetwist, caspases, dad1, p53, e1b, csp, bcl, necrotic, telomerase, undergoing, interleukin1, killer, corpses, oligomerization, nit, converting, pathological, deaths, protective, flice, injury, necrosi, counterpart, cep1, apaf1, prodomain, lta, mac, damageinduced, lin24, homeostasi, tunel



GO-based labels: *MF:* ; *CC:* myosin ii, unconventional myosin, striated muscle thin filament; *BP:* regulation of locomotion, locomotion, negative regulation of locomotion, positive regulation of locomotion, microgametogenesis, abscission, aging **CGC-based labels:** myosin, thick, paramyosin, mhc, rod, unc45, filaments, sup3, unc87, e675, tropomyosin, let56, act3, unc82, myo1, heavychain, unc89, sdf2, e190, peb, act2, subfilament, musculature, chains, diameter, tmy1, lev11, assembled, mhca, let653, sdf9, lie, assemblage, nonunc54, filagenin, finestructure, um, globular, delete, lin40, tubule, organized, sdf8, myofibrillar, positioned, let52, sdf19, rigid, antimyosin, sdf7, waterston, hcdf1, backbone, structures, tubular, multifilament, myosins, crossbridge, c183, reactivity, sdf10, immunological, edf1

Figure 6

Results for four illustrative latent topics specified by a 50-topic CGC LDA model estimated from a corpus with 5,225 documents and a vocabulary of 28,971 words. Each panel shows results for a particular topic. The *y*-axis of the graph is topic-specific word probability (β_{kv})and words are arranged along the *x*-axis according to this likelihood. Only the 500 topic annotation words are plotted since the remaining words in the vocabulary have negligible probabilities. The words displayed explicitly are unigrams in the CGC vocabulary, including the names of *C. elegans* genes, and GO terms. The position of a word along the *x*-axis represents its rank; the staggering of words along the *y*-axis is not significant and is designed only to improve legibility. The graph legend lists two types of automatically-generated topic labels. CGC-based topic labels are a subset of the 50 × 500 topic annotation words that are unique to a topic and are words from the CGC vocabulary; these labels are ordered according to decreasing β_{kv} values. GO-based topic labels are the parents and grandparents GO terms of GO terms that are also topic annotation words. Only GO terms that occur four or more times are given and are listed in decreasing frequency (MF: molecular function; CC: cellular component, BP: biological process). A CGC-based label is unique to a topic whereas a GO-based label can be applied to one or more topic.

Exhibition of multiple latent topics by an LDA document reflects the complexity of issues discussed in documents

By virtue of its superior generalization performance and retrieval ability, an LDA model of the CGC corpus is a better statistical model of text than a mixture of unigrams model. A distinct advantage of LDA is that although both models are generative, an LDA document is the manifestation of many topics whereas a mixture of unigrams document is the product of only one topic. The subject matter of (biological) documents is rarely limited to a single area so the benefit of a CGC LDA is that words in a single document could come from, for example, a combination of Topic 6 (sexual reproduction) and Topic 14 (chromosome structure and function). The mixing of LDA topics in a CGC item was investigated by examining the document-specific, word-level parameter ϕ (Figure 2). The variational posterior topic probability $\phi_n(z_n = k)$ indicates the extent to which the nth word is associated with the kth topic. A value that is both large and significant is an indicator of the topic most likely to have generated the word.

The CGC item shown in Figure 1 is primarily a mixture of two topics. Figure 9 shows the topics most likely to have produced words in the document discussing the life span modifying *clk-2* gene (Table 1). Of the assigned words, 34 have posterior probabilities peaked on the aging-related Topic 9 (Figure 8) and 23 on the general purpose Topic 38 (Figure 7). Three words are allocated to Topic 7, two to Topic 19, two to Topic 13, and one to Topic 34.

Utility of LDA in formulating hypotheses: insights into clk-2 function

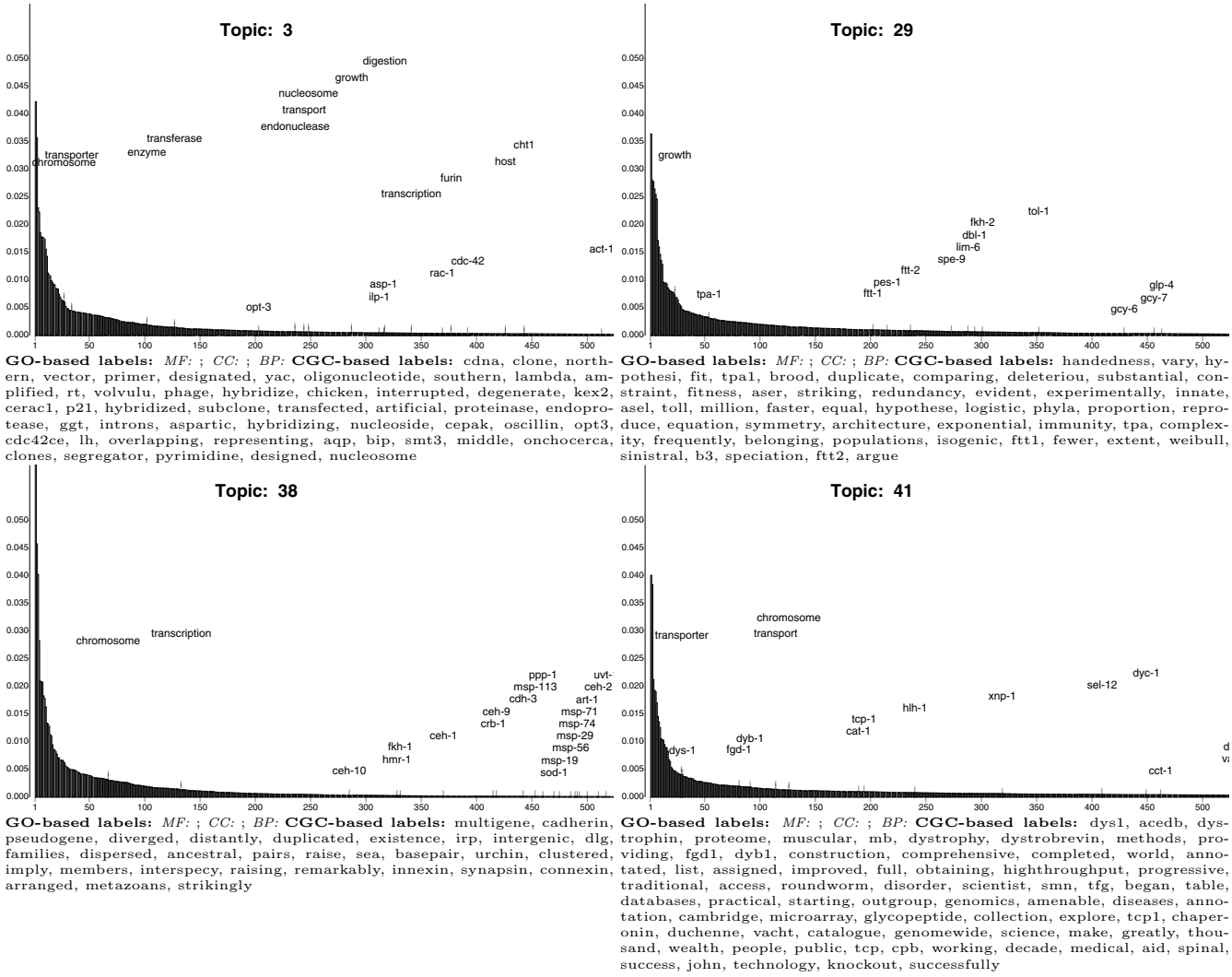
Searching a document database to identify homologs of a query document yields insights that can complement those obtained from sequence-, structure- and function-based analysis of genes and proteins. Prior studies of *clk-2* revealed that it encodes a sequence homolog of Tel2, a protein required for normal telomere length regulation in yeast (reviewed in [26]). To enhance knowledge of how *clk-2* might influence life span, homologs of a document discussing *clk-2* (Figure 1) were identified by computing the topic-space pairwise similarity score between this query q and every CGC document t (Equations 2 and 3). Although a given gene may appear in the Genes record of many CGC items, the results described below are based on analysis of document homologs of the single CGC item shown in Figure 1. Figure 10 shows the three most related items, CGC documents with the three largest $\mathcal{S}(q, t)$ values. Figure 11 shows the three topics most associated with them. The third best homolog is relatively uninformative: the text indicates a general review of aging mutants and Topic 7 labels are general words pertaining to life span.

LDA-based analysis leads to the hypothesis that *clk-2* may have a role in coordinating signals between the outside and inside of cells. Since the top two topic-space document homologs discuss nuclear receptors, *clk-2* may have a direct or indirect involvement in receptor biology. Topics 44 and 45 include the GO-derived labels "host cell plasma membrane," "regulation of fgf receptor signaling pathway" and "regulation of beta 2 integrin biosynthesis." Circumstantial evidence supports a possible role for clk-2 in signal transduction and tissue biology. FGF-2 regulates telomerase activity in human endothelial cells [27]. Integrins are cell surface receptors important in communication between the extracellular environment and the nucleus [28]. The suggestion that *clk-2* may influence telomere length via a mechanism not involving direct physical association with telomeres is plausible since a recent genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length identified genes with very diverse functions (overrepresented categories included DNA and RNA metabolism, chromatin modification, and vacuolar traffic) [29]. Recent experimental results support a connection between vacuolar protein-sorting genes and telomere length homeostasis [30].

Discussion

This study demonstrates how a specific statistical IR model, an LDA model, can be employed to infer the hidden factors permeating a biomedical text corpus and exploited to synthesize and organize information about complex biological phenomena. The results indicate that despite being estimated from a simple bag of words representation of items in the CGC Bibliography, the intra-document statistical structure captured by an LDA model is sufficient for the model to be used to enhance understanding of *C. elegans* biology. For example, analysis of the corpus-, document- and word-level parameters of a trained LDA model enabled the exploration and creation of hypotheses about known and putative nematode aging-related genes.

The CGC corpus studied here had $M = 5,225$ documents and a $V = 28,971$ word vocabulary. Estimating a 50-topic LDA model from such training data took 3 hours on a Macintosh Powerbook G4. It should be straightforward to estimate a model for larger corpora such as MEDLINE where the number of documents is many orders of magnitude greater ($M \sim 10^7$) and the vocabulary size is only one order of magnitude larger ($V \sim 10^5$). In estimating an LDA, the computational bottleneck is the variational E-step, *i.e.*, computing the posterior topic Dirichlet distribution for each document. Fortunately, this procedure can be parallelized because given a model, the posterior for each document can be assessed independently. Thus, it is feasible for the techniques described in this study to be



GO-based labels: *MF:* ; *CC:* ; *BP:* **CGC-based labels:** cdna, clone, northern, vector, primer, designated, yac, oligonucleotide, southern, lambda, amplified, rt, volvulu, phage, hybridize, chicken, interrupted, degenerate, kex2, cerac1, p21, hybridized, subclone, transfected, artificial, proteinase, endoprotease, ggt, introns, aspartic, hybridizing, nucleoside, cepak, oscillin, opt3, cdc42ce, lh, overlapping, representing, aqp, bip, smt3, middle, onchocerca, clones, segregator, pyrimidine, designed, nucleosome

GO-based labels: *MF:* ; *CC:* ; *BP:* **CGC-based labels:** handedness, vary, hypothesi, fit, tpa1, brood, duplicate, comparing, deleteriou, substantial, constraint, fitness, aser, striking, redundancy, evident, experimentally, innate, asel, toll, million, faster, equal, hypothese, logistic, phyla, proportion, reproduce, equation, symmetry, architecture, exponential, immunity, tpa, complexity, frequently, belonging, populations, isogenic, ftt1, fewer, extent, weibull, sinistral, b3, speciation, ftt2, argue

GO-based labels: *MF:* ; *CC:* ; *BP:* **CGC-based labels:** multigene, cadherin, pseudogene, diverged, distantly, duplicated, existence, irp, intergenic, dlg, families, dispersed, ancestral, pairs, raise, sea, basepair, urchin, clustered, imply, members, interspecy, raising, remarkably, innexin, synapsin, connexin, arranged, metazoans, strikingly

GO-based labels: *MF:* ; *CC:* ; *BP:* **CGC-based labels:** dys1, acedb, dystrophin, proteome, muscular, mb, dystrophy, dystrobrevin, methods, providing, fgd1, dyb1, construction, comprehensive, completed, world, annotated, list, assigned, improved, full, obtaining, highthroughput, progressive, traditional, access, roundworm, disorder, scientist, smn, tfg, began, table, databases, practical, starting, outgroup, genomics, amenable, diseases, annotation, cambridge, microarray, glycopeptide, collection, explore, tcp1, chaperonin, duchenne, vacht, catalogue, genomewide, science, make, greatly, thousand, wealth, people, public, tcp, cpb, working, decade, medical, aid, spinal, success, john, technology, knockout, successfully

Figure 7
CGC LDA topics that have no GO-based topic labels and capture hidden factors in the CGC corpus that pertain to the practical aspects of investigating biological mechanisms and processes. Topics are represented in the same manner as in Figure 6.

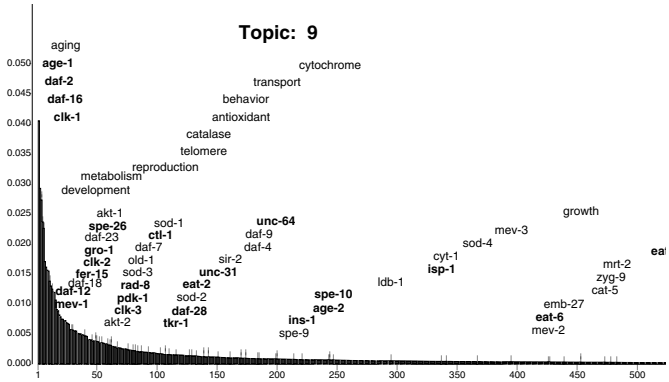
applied to other corpora and to address questions other than life span modification.

From an applications perspective, it is possible to envisage a scenario involving the creation of a library of LDA models where each constituent model was estimated from a user- and/or computationally-defined corpus of documents focused on a specific area such as "aging," "cancer," "yeast biology," "response to stress," "antibiotics," "HIV," "Parkinson's disease," "kinases," and so on. A query document would be compared to each subject-specific LDA in the collection as opposed to a single model as described here. This approach seeks to mirror a common strategy in sequence analysis whereby a query sequence is rated against a library of hidden Markov models (HMMs) [31] estimated for domains of interest, for example, the Pfam database of protein families [32]. Since seaching a database of probabilistic models of proteins in order to identify "remote" sequence homologs is known to be effective, an analogous approach could prove useful in retrieving distant document homologs.

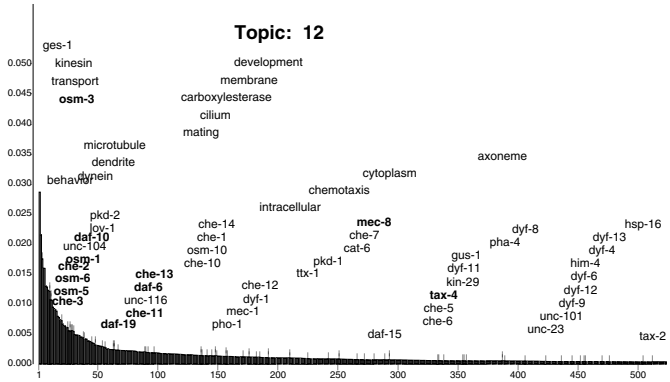
The ideas discussed in this work can be extended and improved in a variety of ways. Currently, the number of latent topics in an LDA model K is a user-defined parameter but recent research has examined the task of choosing K [33]. In particular, if the LDA is augmented with a non-parametric Bayesian prior known as a hierarchical Dirichlet process, both topic probabilities and the number of topics can be estimated from data. Under the hierarchical Dirichlet process prior, the number of topics grows as data are added to a collection []. The significance

Table 1: *C. elegans* genes known to extend or shorten life span. The list is taken from the Genes database of SAGEKE http://www.sageke.org

	Life span extension
Hsp-6	heat shock 70 protein
age-1	inositol/phosphatidylinositol kinase; signal transduction
age-2	AGEing alteration
che-2	G-beta-repeats
che-3	microtubule motor, dynein ATPase; microtubule-based movement
che-11	abnormal CHEmotaxis
che-13	abnormal CHEmotaxis
clk-1	ubiquinone biosynthesis
clk- 2	CLocK (biological timing) abnormality
clk-3	CLocK (biological timing) abnormality
daf-2	transmembrane receptor protein tyrosine kinase; phosphorylation, hydrogen transport, signalling
daf-6	abnormal DAuer Formation
daf-10	abnormal DAuer Formation
daf-12	steroid hormone receptor; transcription regulation
daf-19	DNA binding transcription factor; transcription regulation
daf-28	abnormal DAuer Formation
eat-1	EATing: abnormal pharyngeal pumping
eat-2	EATing: abnormal pharyngeal pumping
eat-3	EATing: abnormal pharyngeal pumping
eat-6	Na^+/K^+ ATPase α subunit; cation transport, metabolism
eat-13	EATing: abnormal pharyngeal pumping
glp-1	calcium ion binding; cell differentiation
gro-1	tRNA isopentenyltransferase; tRNA processing
ins-1	INSulin related
ins-18	insulin-like growth factor I like; hormone
isp-1	ubiquinol-cytochrome c reductase, Rieske iron-sulfur protein; electron transport
mec-8	nucleic acid binding; mechanosensory
mes-1	protein tyrosine kinase
osm-1	OSMotic avoidance abnormal
osm-3	kinesin
osm-5	aspartic-type endopeptidase; proteolysis and peptidolysis
osm-6	N-acetyllactosamine synthase
pdk-1	protein serine/threonine kinase
pgl-1	P GranuLe abnormality
rad-8	RADiation sensitivity abnormal/yeast RAD-related
sir-2.1	DNA binding; transcription regulation, chromatin silencing
spe-10	defective SPErmatogenesis
spe-26	MIPP repeats; defective SPErmatogenesis
tax-4	Cyclic-nucleotide-gated olfactory channel; potassium transport
tkr-1	G protein coupled receptor; signalling
unc-4	homeobox protein (otd subfamily); transcription regulation
unc-13	intracellular signaling cascade
unc-26	inositol/phosphatidylinositol phosphatase
unc-31	PH (pleckstrin homology) domain
unc-32	TJ6/proton pump
unc-64	syntaxin
unc-76	UNCoordinated
	Shortened life span
ctl-1	CaTaLase
daf-16	transcription factor
eat-7	EATing: abnormal pharyngeal pumping
fer-15	FERtilization defective (abnormal sperm)
mev-1	Succinate dehydrogenase cytochrome b chain; electron transport, tricarboxylic acid cycle



GO-based labels: *MF:* ; *CC:* mitochondrial derivative, kinetoplast, mitochondrial degradosome, microsome, telomerase holoenzyme complex; *BP:* adult feeding behavior, determination of adult life span, abscission, aging, drug metabolism, adult behavior **CGC-based labels:** age1, longevity, lifespan, clk1, ageing, insulin, mev1, insulinlike, longlived, daf18, superoxide, fer15, clk2, gro1, daf23, clk, akt1, dismutase, akt2, extended, consumption, clk3, pdk1, rad8, sod3, pten, old1, diapause, ctl1, aged, thermotolerance, caloric, igf, tkr1, gerontogene, daf28, agespecific, coq7, sod2, expectancy, ubiquinone, receptorlike, dramatically, carbonyl, sod, sir2, hx546, antioxidant, capacity, tension, paraquat, daf9, itt, lengthen, cln3, agedependent, pi3k, chronological, fashion, nondauer, postreproductive, ins1, replica, lived, slow, slowing, singlegene, age2, coenzyme, spe10



GO-based labels: *MF:* ; *CC:* axonemal microtubule, mitochondrial outer membrane, mitochondrial membrane, peroxisomal membrane, apical part of cell, basal part of cell; *BP:* immune cell chemotaxis, adult feeding behavior, ethylene mediated signaling pathway, osmosensory signaling pathway via two-component system, adult behavior, two-component signal transduction system 'phosphorelay', abscission, aging, neutrophil chemotaxis **CGC-based labels:** ges1, cilia, kinesin, esterase, amphid, osm3, pha, osm5, che2, ash, osm1, retrograde, ending, avoidance, sensilla, amphidial, ciliated, lov1, gutspecific, pkd2, phasmid, ift, gesl, dendrite, daf19, cargo, che, repellent, anterograde, ir, che11, unc116, che13, avoid, intraflagellar, wgatar, che10, cephalic, osm10, crystallin, che1, moving, cilium, che14, digestive, isoelectric, carboxylesterase, pho1, chaperonelike, cog, focusing, ciliary, rectum, tagged, deleted, dyf1, con, dendritic, che12, ase, perception, sorting, fluorescein, repellents, raft, polycystic, lip, dendrites, fitc, pkd1, adl, chemoreception, deirid, flagella, dimeric, chemotaxis

Figure 8
The two CGC LDA topics with the greatest numbers of aging-related gene words (CGC-based topic labels corresponding to the names of genes implicated in modifying life span). Each topic is represented the same manner as Figure 6. In the graph, words in bold are the gerontogenes listed in Table 1.

and practical importance of this feature is that when modeling scientific documents, the nature and size of the corpus is evolving constantly meaning that discovering topics is an ongoing task. Further studies are required to devise rigorous methods for determining a set of words best able to characterize a topic: the current approach is somewhat arbitary in that corpus-based topic labels were defined as the 500 words in CGC vocabulary with the highest values for the likelihood of the word given the topic.

In the LDA model, topics represent entities that are presumed to permeate a corpus and these latent variables are to be inferred by statistical analysis. These implicit concepts are assumed to be equally related to one another, *i.e.*, the topics are not organized in any way and form a "flat structure." It seems reasonble to believe that, for example, topics embodying the concepts of "DNA repair" and "chromosome structure and function" should be more related to each other than to a topic focused on

Key: 4951
Authors: Lim CS;Mian IS;Dernburg AF;Campisi J
Title: C. **elegans(38)** **clk-2,(9)** a **gene(38)** that **limits(9)** **life(9)** **span,(9)** **encodes(38)** a **telomere(9)** **length(9)** **regulator(19)** **similar(38)** to **yeast(9)** **telomere(9)** **binding(13)** **protein(13)** **Tel2p.(9)**
Genes: **clk-2(9)**
Abstract: An **important(38)** **quest(9)** in **modern(9)** **biology(7)** is to **identify(38)** **genes(38)** **involved(9)** in **aging.(9)** Model(7) **organisms(7)** such as the **nematode(38)** **Caenorhabditis(9)** **elegans(38)** are particularly useful in this **regard.(9)** The C. **elegans(38)** **genome(38)** has been **sequenced(38)** [1], and **single(38)** **gene(38)** **mutations(9)** that **extend(9)** **adult(9)** **life(9)** **span(9)** have been **identified(38)** [2]. Among these **longevity-controlling(9)** **loci(34)** are four **apparently(38)** **unrelated(38)** **genes(38)** that **belong(38)** to the **clk(9)** **family(38)** [3-5]. In **mammals,(38)** **telomere(9)** **length(9)** and **structure(38)** can **influence(9)** **cellular,(9)** and **possibly(9)** **organismal,(9)** **aging(9)** [6]. Here, we **show(9)** that **clk-2(9)** **encodes(38)** a **regulator(19)** of **telomere(9)** **length(9)** in C. **elegans.(38)**

Figure 9
The LDA topics most associated with words in the CGC item shown in Figure 1. A word is identified with the topic k given in parenthesis when the document-specific, variational posterior topic probability exceeds a threshold, $\phi_n(z_n = k) > 0.9$. As illustrated by "telomere (9)", identical words within a document are generated by the same topic. Note that only the Title, Genes and Abstract records were concatenated and processed to generate the bag-of-words document used to estimate the LDA.

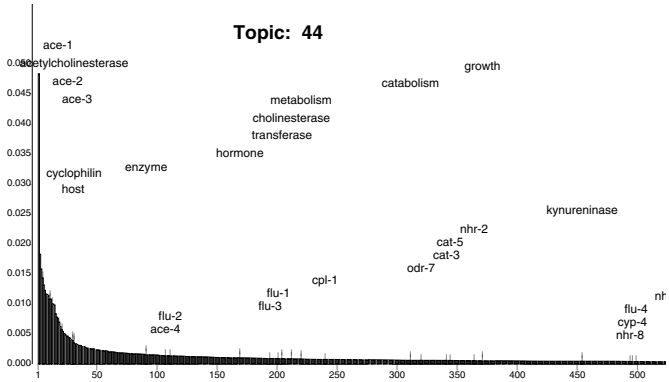
Key: 3416
Authors: Sluder AE;Mathews SW;Hough D;Yin VP;Maina CV
Title: The **nuclear(44) receptor(45) superfamily(38)** has **undergone(38) extensive(38) proliferation(44)** and **diversification(44)** in **nematodes.(44)**
Genes: **daf-12(44) fax-1(44) nhr-1(44) nhr-2(44) nhr-3(44) nhr-4(44) nhr-5(44) nhr-6(44) nhr-7(44) nhr-8(44) nhr-8(44) nhr-10(44) nhr-11(44) nhr-12(44) nhr-13(44) nhr-14(44) nhr-15(44) nhr-16(44) nhr-17(44) nhr-18(44) nhr-20(44) nhr-21(44) nhr-22(44) nhr-23(44) nhr-24(44) nhr-25(44) nhr-28(44) nhr-31(44) nhr-34(44) nhr-35(44) nhr-40(44) nhr-41(44) nhr-42(44) nhr-43(44) nhr-44(44) nhr-45(44) nhr-46(44) nhr-47(44) nhr-48(44) nhr-49(44) nhr-50(44) nhr-51(44) nhr-52(44) nhr-53(44) nhr-54(44) nhr-55(44) nhr-56(44) nhr-57(44) nhr-58(44) nhr-59(44) nhr-60(44) nhr-61(44) nhr-62(44) nhr-63(44) nhr-64(44) nhr-65(44) nhr-66(44) nhr-67(44) odr-7(44) sex-1(44) unc-55(44)**
Abstract: The **nuclear(44) receptor(45)** (NR) **superfamily(38)** is the most **abundant(38) class(44)** of **transcriptional(19) regulators(19) encoded(38)** in the **Caenorhabditis(44) elegans(38) genome,(38)** with >200 **predicted(38) genes(38) revealed(44)** by the **screens(44)** and **analysis(38)** of **genomic(38) sequence(38) reported(44)** here. This is the **largest(44) number(44)** of **NR(44) genes(38)** yet described from a **single(38) species,(44)** although our **analysis(38)** of available **genomic(38) sequence(38)** from the **related(38) nematode(44) Caenorhabditis(44) briggsae(38)** indicates that it also has a **large(44) number.(44) Existing(44) data(38) demonstrate(44) expression(38)** for 25% of the C. **elegans(38) NR(44) sequences.(38) Sequence(38) conservation(38)** and **statistical(44) arguments(44) suggest(38)** that the **majority(44) represent(44) functional(38) genes.(38)** An **analysis(38)** of these **genes(38) based(44)** on the DNA-binding **domain(13) motif(13) revealed(44)** that several **NR(44) classes(44) conserved(38)** in both **vertebrates(38)** and **insects(44)** are also **represented(44)** among the **nematode(44) genes,(38) consistent(44)** with the **existence(38)** of **ancient(44) NR(44) classes(44) shared(44)** among most, and perhaps all, metazoans. None of the **nematode(44) NR(44) sequences,(38)** however, are **distinct(44)** from those currently known in other **phyla,(44)** and **reveal(38)** a **previously(38) unobserved(44) diversity(44)** within the **NR(44) superfamily.(38)** In C. **elegans,(38) extensive(38) proliferation(44)** and **diversification(44)** of **NR(44) sequences(38)** have **occurred(44)** on **chromosome(38)** V, **accounting(44)** for > 50% of the **predicted(38) NR(44) genes.(38)**
Key: 4694
Authors: Sluder AE;Maina CV
Title: **Nuclear(44) receptors(45)** in **nematodes:(44) themes(44)** and variations.
Genes: **daf-12(44) fax-1(44) nhr-2(44) nhr-6(44) nhr-8(44) nhr-23(44) nhr-25(44) nhr-41(44) nhr-48(44) nhr-64(44) nhr-67(44) nhr-69(44) nhr-85(44) nhr-91(44) odr-7(44) sex-1(44) unc-55(44) xol-1(6)**
Abstract: **Large-scale(41) sequencing(41) efforts(41)** are **providing(41)** new **perspectives(41)** on **similarities(38)** and **differences(44)** among species. **Sequences(38) encoding(38) nuclear(44) receptor(45)** (NR) **transcription(19) factors(19) furnish(44)** one **striking(29)** example of this. The three **complete(41)** or nearly **complete(41) metazoan(38) genome(41) sequences(38)** - those of the **nematode(44) Caenorhabditis(38) elegans,(41)** the **fruit(41) fly(38)** (Drosophila melanogaster) and the **human(41)** - **reveal(38) dramatically(44)** different **numbers(44)** of **predicted(38) NR(44) genes:(38)** 270 for the **nematode,(44)** 21 for the **fruit(41) fly(38)** and **similar(38)** to 50 for the **human.(41)** Although some **classes(44)** of **NRs(44) present(38)** in **insects(44)** and **mammals(38)** are also **represented(44)** among the **nematode(44) genes,(38)** most of the C. **elegans(41) NR(44) sequences(38)** are **distinct(38)** from those known in other **phyla.(44) Questions(41)** regarding the **evolution(38)** and **function(38)** of **NR(44) genes(38)** in **nematodes,(44) framed(44)** by the **abundance(44)** and **diversity(44)** of these **genes(38)** in the C. **elegans(41) genome,(41)** are the **focus(41)** of this **article.(41)**
Key: 1885
Authors: Johnson TE;Tedesco PM;Lithgow GJ
Title: **Comparing(29) mutants,(17) selective(7) breeding,(34)** and **transgenics(30)** in the **dissection(7)** of **aging(9) processes(7)** of **Caenorhabditis(9)**
Genes: **age-1(9) fer-15(9) rol-6(39) mnDf63(39) mnDf89(39) mnDf91(34) mnDf92(9)**
Abstract: The **genetic(7) analysis(7)** of **aging(9) processes(7)** has **matured(9)** in the last **ten(17) years(7)** with **reports(34)** that **long-lived(9) strains(34)** of both **fruit(7) flies(7)** and **nematodes(7)** have been **developed.(7)** Several **attempts(7)** to **identify(17) mutants(17)** in the **fruit(7) fly(7)** with **increased(9) longevity(9)** have **failed(30)** and the **reasons(7)** for these **failures(9)** are **analyzed.(17)** A **major(7) problem(7)** in **obligate(9) sexual(34) species,(29)** such as the **fruit(7) fly(7)** is the **presence(30)** of **inbreeding(34) depression(34)** that **makes(9)** the **analysis(7)** of **life-history(34) traits(34)** in **homozygotes(34)** very **difficult.(7)** Nevertheless, several **successful(7) genetic(7) analyses(7)** of **aging(9)** in **Drosophila(7) suggest(9)** that with **careful(9) design,(7) fruitful(39) analysis(7)** of **induced(30) mutants(17) affecting(17) life(9) span(9)** is possible. In the **nematode(7) Caenorhabditis(9) elegans,(7) mutations(17)** in the **age-1(9) gene(17) result(9)** in a **life(9) extension(9)** of some 70%; thus **age-1(9)** clearly **specifies(34)** a **process(7) involved(7)** in **organismic(9) senescence.(9)** This **gene(17) maps(34)** to **chromosome(34)** II, well **separated(34)** from a **locus(34)** (fer-15) which is **responsible(7)** for a **large(7) fertility(34) deficit(9)** in the **original(7) stocks.(34)** There is no **trade-off(29)** between either **rate(9)** of **development(7)** or **fertility(34) versus(30) life(9) span(9)** associated with the **age-1(9) mutation.(17) Transgenic(30) analyses(7) confirm(30)** that the **fertility(34) deficit(9)** can be **corrected(30)** by a **wild-type(17) fer-15(9) transformant(30)** (transgene); however, the **life(9) span(9)** of these **transformed(30) stocks(34)** is **affected(17)** by the **transgenic(30) array(30)** in an **unpredictable(9) fashion.(9)** The **molecular(7) nature(7)** of the **age-1(9) gene(17) remains(7) unknown(9)** and we **continue(7)** in our **efforts(7)** to

Figure 10

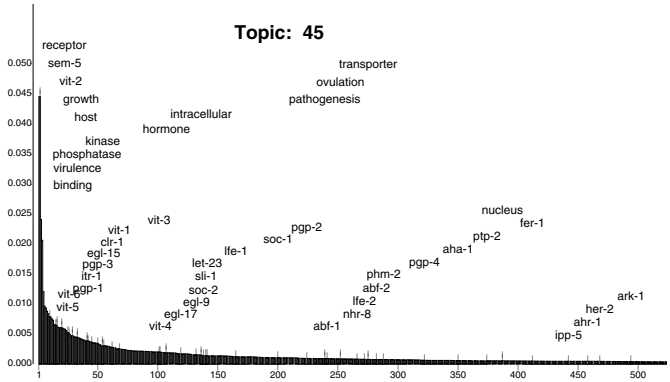
CGC homologs of the *clk-2* item shown in Figure 1. When this *clk-2* item is used as the query document, the three items shown have the largest topic-space pairwise similarity scores, $\mathcal{S}(q, t)$. The documents are depicted in the same format as Figure 9. As illustrated by "elegans(38)" and "elegans(41)" in the first and second top-ranked documents, identical words may be attributed to different topics in different documents.

"locomotion." Thus, a model in which topics were themselves arranged as a hierarchy could prove useful and inspection of the relationship(s) between topics might be informative. The current LDA model represents a model for a particular instantiation of a corpus. It might be interesting to formulate dynamic models of corpora which sought to capture how a collection such as the CGC or MEDLINE changes over time. In addition to theoretical studies of their properties and behavior, such models might be useful not only for biomedical researchers and clinicians, but also policy makers, historians, and sociologists interested in the evolution of biology and its disciplines.

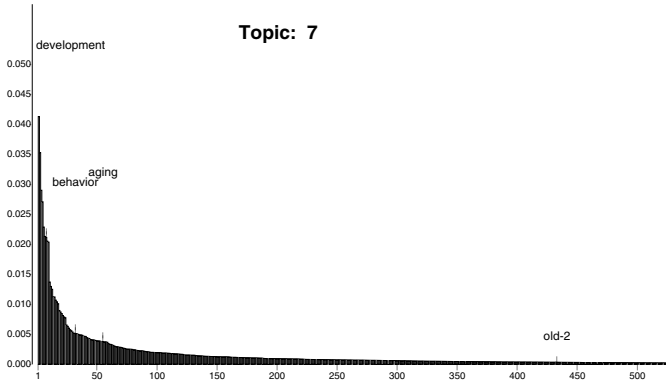
Statistical IR models are applicable not only to "traditional" document collections, but also to genome-scale biological data. In one example of such an LDA model, "documents" could correspond to genes, "topics" to regulatory networks and "words" to motifs. Since LDA does not require mutually exclusive clustering, a given gene can participate in several networks and a given motif can appear in several genes. Merging this type of analysis with that described here would result in evidential support for the functional behavior and role of a gene being derived both from primary biological data and from corpus-based analysis. This capacity to fuse support from disparate sources has been illustrated using a variant of LDA known as "correspondence LDA" and in the context of automated image annotation [23].



GO-based labels: *MF:* ; *CC:* parasitophorous vacuolar membrane, parasitophorous vacuole, host cell plasma membrane; *BP:* drug metabolism **CGC-based labels:** biomass, ache, ace3, web, helminth, cyclophilin, trophic, root, gst, ace, microbial, fungu, nr, plantparasitic, antibiotic, predator, mif, spp, tannin, capture, trap, microcosm, brugia, infective, incognita, parasites, canium, habitat, arthrobotry, supply, ace4, microbivore, nematodetrapping, dependence, flu2, parasitism, helminths, top, condensed, fungivore, ml1, ellagitannin, filarial, greatest, hosts, represented, exploited, productivity, bacterivore, graph, ypd, aphelenchoide, fungal, forms, bacterialfeeding, antifungal, acrobeloide, cycloheximide, prey, lbp, predatory, cpl, wormpd, flu3, microbe, agriculture, flu1, qualitative, malayi, nematophagou, penetran, spore, trophiclevel, omnivore, cpl1, gallo, lph, xylophilu, veterinary, minititin, trapping, predaceou, bacteriophagou, ancylostoma



GO-based labels: *MF:* amino acid binding, calcium ion binding, phospholipase d; *CC:* parasitophorous vacuolar membrane, parasitophorous vacuole, host cell plasma membrane, mitochondrial derivative, apical part of cell, basal part of cell; *BP:* regulation of beta 2 integrin biosynthesis, beta 2 integrin biosynthesis, regulation of fgf receptor signaling pathway, integrin biosynthesis, fgf receptor signaling pathway **CGC-based labels:** vitellogenin, fgf, sh3, vit2, pathogen, sh2, aeruginosa, vit5, vit6, glycoprotein, virulence, fibroblast, ip3, src, pgp1, fgfr, itr1, pgp3, clr1, vpe1, vit1, vpe2, pa14, grb2, cbl, typhimurium, lar, glutamatergic, sem, trisphosphate, pseudomallei, vit3, vit4, sli, insp, soc, oncogene, similarly, vit, sos, ahr, rs, egfr, prolinerich, egl9, pathogenesi, soc2, ligandbinding, burkholderia, fkhr, pgp, clr, cyanide, lfe1, sf1, mucd, liprin, salmonella, fkhrl1, pa, estrogen, grampositive, hydrogen, steroid, drk, soc1, opportunistic, glucocorticoid, neurexin, coil, afx, tractable, leucinerich, broad, pgp2, triphosphate, hostpathogen, vcbl, faecali, igfbp, pathogenicity, pyogene, ip3r, pathogenesis, ccbl, abf1, hced, pathogenic, sms



GO-based labels: *MF:* ; *CC:* ; *BP:* adult feeding behavior, abscission, aging, adult behavior, determination of adult life span **CGC-based labels:** review, detailed, progress, provided, past, offer, focu, chapter, advance, area, systems, brenner, reviewed, example, physiology, opportunity, description, summarize, excellent, phenomena, dissection, genetics, numerou, mechanisms, attention, address, contribution, way, emerged, difficult, briefly, interested, focused, challenge, volume, simplicity, investigator, sparc, sydney, kingdom, remarkable, invertebrates, highlight, elucidation, ideal

Figure 11
The CGC LDA topics most associated with words in the document homologs of the *clk-2* query shown in Figure 10.

Conclusion

LDA is a special case of the family of probabilistic graphical models. This family includes a wide variety of other models that have proved useful in biology such as HMMs, phylogenetic trees, and pedigrees [34]. The graphical model formalism allows such graphical components to be combined into heterogenous, large-scale statistical models that integrate evidence from multiple sources. Doing so would yield a system for facilitating the formulation of ideas that could be interrogated and verified experimentally.

Methods

CGC document corpus

The CGC Bibliography (October 2002 release) was downloaded [35] and included abstracts from the published literature, "worm meetings," and the "Worm Breeder's Gazette." Each CGC item is a series of defined records (Key, Medline, Authors, Title, Citation, Type, Genes, Abstract) and associated free-text (see Figure 1 for an exemplar). In an item, the Genes record associated with an Abstract record was added by a curator and so reflects the personnel interest(s) and background of the individ-

ual, *i.e.*, the list of genes discussed in the abstract was not derived in a systematic, automated manner and according to a fixed scheme and/or philosophy. For each item, the Genes, Title, and Abstract records were concatenated. The resultant text was tokenized by partitioning on the basis of white spaces and punctuation. Variants of the same word were stemmed to produce a single word by removing the suffixes s, ss, ies and sses. A word was discarded if it was in a set of 619 generic stop words that included ii, iii, iv, iv, a, a., yourselves, z, and zero. The ensuing text was not tagged, *i.e.*, words were not annotated in terms of syntax (parts of speech) or semantics (pre-defined class such as "gene name").

A CGC document is the bag of words (vector space representation) obtained after tokenizing, stemming, and removing suffixes from a CGC item. The CGC vocabulary is the non-redundant set of discrete objects produced after application of the preceeding processing steps to all items in the corpus. Words in this vocabulary are unigrams because phrases such as "DNA repair" are considered to be two objects rather than one. The final CGC corpus had M = 5, 225 documents and a vocabulary of V = 28, 971 words. The shortest and longest documents had N = 9 and N = 261 words respectively. It should be noted that words originating from the Genes record were not marked in any way so neither the vocabulary constructed for the corpus nor the bag of words representation of a particular document contained explicit knowledge of, or information about the curation effort.

Latent Dirichlet Allocation (LDA) model

A detailed description of the LDA model and attendant algorithms can be found elsewhere [22,36]. Figure 2 gives a graphical model representation of LDA. Consider a corpus of M documents, $\{\mathbf{w}_1, ..., \mathbf{w}_M\}$, based on a vocabulary of words indexed by [1, ..., V]. A document is a sequence of N words, $\mathbf{w} = [w_1, ..., w_N]$. Each word in a document is represented by a V-dimensional unit-basis vector in which only one component is equal to one, *i.e.*, a word that is the vth word in the vocabulary is described by a vector where $w^v = 1$ and $w^u = 0$ for all $u \neq v$.

LDA generates a document according to the following process,

1. Choose a point θ from a Dirichlet distribution parameterized by α: $\theta \sim$ Dirichlet (α).

2. For each word w_n in turn,

(a) Choose a topic z_n from a multinomial distribution parameterized by θ: $z_n \sim$ Multinomial(θ).

(b) Choose a word w_n from the multinomial distribution associated with the selected topic and parameterized by β_{z_n} : $w_n \sim$ Multinomial (β_{z_n}).

The parameters of a K-iopic LDA are the Dirichlet parameter, α, and the topic-word matrix, β. α is a K-dimensional Dirichlet parameter that determines the distribution over topic proportions, $\alpha = [\alpha_1, ..., \alpha_K]$ where $\alpha_k > 0$. β is a $K \times V$ matrix that determines the likelihood of the vth word in the vocabulary given the kth topic, $\beta_{kv} = p(w^v = 1 \mid z^k = 1)$. The topic-specific word distribution β_k β_{z_n} is the kth row of β. The document-specific topic proportions θ lie in the (K - 1)-dimensional simplex, $\theta = [\theta_1, ..., \theta_k]$ where $\theta_k = p(z^k = 1 \mid \theta)$, $\theta_k > 0$, and $\sum_{k=1}^{K} \theta_k = 1$.

The likelihood of an LDA document is obtained by marginalizing over the latent variables,

$$p(\mathbf{w} \mid \alpha, \beta) = \int p(\theta \mid \alpha) \prod_{n=1}^{N} \left(\sum_{z_n=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta) \right) d\theta. \quad (1)$$

The probability of a corpus is the product of the marginal probabilites of single documents.

Inference and parameter estimation

Estimating the parameters of an LDA from data and calculating the probability of a document both involve inference or computing the posterior distribution of latent variables given a document,

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w} \mid \alpha, \beta)}{p(\mathbf{w} \mid \alpha, \beta)}.$$

The denominator of this expression, the likelihood of a document, cannot be computed exactly because of the coupling between θ and $\mathbf{z}$ (left, Figure 2). Instead, approximate inference methods are required for the LDA model. Such methods include the convexity-based variational approach used here [22], Markov chain Monte Carlo sampling [37] and expectation propagation [38].

Variational inference approximates the posterior by finding a lower bound on the likelihood of a document [22]. The family of graphical models obtained by uncoupling θ and $\mathbf{z}$ (right, Figure 2) is characterized by the following variational distribution,

$$q(\theta, \mathbf{z} \mid \gamma, \phi) = q(\theta \mid \gamma) \prod_{n=1}^{N} q(z_n \mid \phi_n),$$

where the Dirichlet parameter γ and the multinomial parameters $\{\phi_1, ..., \phi_N\}$ are the free variational parameters for document **w**. For the nth word, $\phi_n(z_n = k)$ is the (variational) posterior topic probability for the *kth* topic. Optimal document-specific values can be found by minimizing the Kullback-Leibler (KL) divergence between the variational distribution and the true posterior,

$$(\gamma^*(\mathbf{w}), \phi^*(\mathbf{w})) = \arg\min_{(\gamma,\phi)} KL(q(\theta, \mathbf{z} \mid \gamma(\mathbf{w}), \phi(\mathbf{w})) \mid\mid p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta)).$$

The variational Dirichlet parameter γ^* (**w**) provides a representation of the document in the topic simplex. The variational multinomial parameters $\phi^*(\mathbf{w}) = \{\phi_1^*, ..., \phi_N^*\}$ approximate the true, but intractable distributions $p(z_n|\mathbf{w})$.

Estimating an LDA from data involves finding the Dirichlet parameter α^* and topic-word matrix β^* which maximize the log marginal likelihood of a corpus. A variational Expectation-Maximization procedure results in parameter estimates that are a (possibly local) maximum of a lower bound of the log marginal likelihood. This alternates between maximizing a lower bound with respect to the variational parameters for each document, and maximizing the lower bound with respect to the model parameters.

An LDA-based measure of pairwise document similarity

The transformation of bags of words into bags of topics by LDA provides a means to address the task of searching a corpus to retrieve similar and/or relevant items. Word-space representations of documents (high-dimensional, variable length, vectors of discrete-valued features) are converted into topic-space representations (low-dimensional, fixed length, vectors of real-valued features). In particular, the variational posterior Dirichlet parameter γ^* (**w**) indicates the degree to which each of the K topics is referenced by document **w**.

A new measure of pairwise document similarity was formulated using the Dirichlet probability distribution specified by the document-level LDA parameter γ (Figure 2). Recall that the variational posterior Dirichlet distribution for document d is Dirichlet(γ_d) where $\gamma_d = [\gamma_{d_1}, \ldots, \gamma_{d_K}]$ and γ_{d_k} denotes the extent to which document d refers to the kth topic. A random variable θ drawn from Dirichlet(γ_d) has the probability density

$$p(\theta \mid \gamma_d) = \frac{\Gamma_{k=1}^{K}\left(\sum_{k=1}^{K}\gamma_{d_k}\right)}{\prod_{k=1}^{K}\Gamma\left(\gamma_{d_k}\right)}\prod_{k=1}^{K}\theta_{d_k}^{\gamma_{d_k}-1}$$

where $\theta_{d_k} > 0$, $\sum_{k=1}^{K}\theta_{d_k} = 1$, $\gamma_{d_k} > 1$, and $\Gamma(\cdot)$ is the Gamma function. Given two Dirichlet densities, the KL divergence is given by

$$KL(p(\theta \mid \gamma_i) \mid\mid p(\theta \mid \gamma_j)) = \int p(\theta \mid \gamma_i) \log \frac{p(\theta \mid \gamma_i)}{p(\theta \mid \gamma_j)} d\theta.$$

The similarity between two documents i and j can be quantified by computing the KL divergence between their corresponding Dirichlet distributions as follows

$$\begin{aligned} KL(\gamma_i \mid\mid \gamma_j) = &\sum_{k=1}^{K}(\log\Gamma(\gamma_{i_k}) - \log\Gamma(\gamma_{j_k})) \\ &-\log\Gamma\left(\sum_{k=1}^{K}\gamma_{i_k}\right) + \log\Gamma\left(\sum_{k=1}^{K}\gamma_{j_k}\right) \\ &+\sum_{k=1}^{K}(\gamma_{i_k} - \gamma_{j_k})\left(\Psi(\gamma_{i_k}) - \Psi\left(\sum_{k=1}^{K}\gamma_{i_k}\right)\right) \end{aligned} \quad (2)$$

where γ_i and γ_j are the parameters for the two documents. $\psi(\cdot)$ is the digamma function and arises when taking expectations of log θ.

Let $\mathcal{S}(q, t)$ be the topic-space pairwise similarity score between a query document q and a target document t. Here, this score is defined as the symmetrized KL divergence between the variational posterior Dirichlet distributions,

$$\mathcal{S}(q,t) = \frac{1}{2}KL(\gamma_q \mid\mid \gamma_t) + \frac{1}{2}KL(\gamma_t \mid\mid \gamma_q), \quad (3)$$

where the component KL terms are computed using Equation 2.

An alternative definition of the topic-space pairwise similarity score is the Jensen-Shannon divergence

$$\mathcal{S}(q,t) = \frac{KL(\gamma_q \mid\mid \gamma_x) + KL(\gamma_t \mid\mid \gamma_x)}{2}$$

where $\gamma_x = (\gamma_q + \gamma_t)/2$ so that γ_x^k is the average degree to which documents q and t refer to topic k.

Given a database of D documents, t_1, ..., t_D, the task of retrieving documents related to a query q was addressed by computing the score between q and each document in

the collection, $\mathcal{S}(q, t_1)$, ..., $\mathcal{S}(q, t_D)$ (*cf.* searching a protein or nucleic acid sequence database to find homologous sequences). For simplicity, a "homolog" of a document q is a CGC document t_d that has a high topic-space pairwise similarity score $\mathcal{S}(q, t_d)$.

Mixture of unigrams model and unigram model

Figure 3 shows graphical model representations of two existing models of text. Like the graphical model representation of the LDA model (left, Figure 2), the open circles represent parameters for the mixture of unigrams model (mixing weights θ, word distribution β_k)and unigram model (word distribution β).

The unigram model contains no latent variables and each word in every document is assumed to have been drawn from the same multinomial distribution. Denoting this word distribution as β, the likelihood of a unigram document is

$$p(\mathbf{w} \mid \beta) = \prod_{n=1}^{N} p(w_n \mid \beta). \qquad (4)$$

The mixture of unigrams model [39] assumes that each document is generated by first choosing one of K topics, and then drawing words independently, conditioned on that topic. As in LDA, the K multinomial distributions represent an underlying semantic structure in the corpus but a mixture of unigrams document is a manifestation of only one of these topics. Denoting the mixing weights as θ and the word distributions as β, the likelihood of a mixture of unigrams document is

$$p(\mathbf{w} \mid \theta, \beta) = \sum_{z=1}^{K} p(z \mid \theta) \left(\prod_{n=1}^{N} p(w_n \mid z, \beta) \right) \qquad (5)$$

The LDA model builds upon the mixture of unigrams in that documents are able to manifest multiple topics. Overall, the estimated LDA word distributions are better reflections of the underlying topics in a corpus, particularly in light of heterogenous documents.

Assessment of statistical models of text

Perplexity: generalization performance

The performance of a statistical model on unseen data was evaluated by computing the perplexity of a test set of J documents not used to estimate the model,

$$\textit{perplexity}(\mathbf{w}_1, \ldots, \mathbf{w}_J) = \exp\left\{ -\frac{\sum_{j=1}^{J} \log p(\mathbf{w}_j)}{\sum_{j=1}^{J} N_j} \right\}, \qquad (6)$$

where N_j is the number of words in test document $\mathbf{w}_j$. The perplexity is equivalent to the inverse of the geometric mean per-word likelihood and a lower score indicates better generalization performance.

The three different bag of words models described above were assessed by determining their respective perplexity on the same set of test documents. The likelihood of a test document $p(\mathbf{w}_j)$ was computed using Equation 1 for an LDA model (left, Figure 2), Equation 4 for a unigram model (left, Figure 3), and Equation 5 for a mixture of unigrams model (right, Figure 3). A model was trained using 90% of the CGC corpus (4,700 documents) and the remaining 10% (J = 525) used to compute perplexity. A single unigram model was estimated and evaluated whereas LDA and mixture of unigram models with varying numbers of latent topics were estimated and evaluated (K = 5,10, 20, 50,100).

Precision/recall (PR) curves and F1 measure: retrieval capability

The ability of a statistical model to retrieve a set of related documents was evaluated using a language modeling approach [20,40]. Let N_D be the number of "relevant" documents in a collection of D documents, documents that are related according to some criterion. Let N_P be the number of these relevant documents that are present in a list of P of these documents ($P \leq D$). Precision, $\mathcal{P}$, is the fraction of documents in the list that are relevant and is defined as N_P/P. Recall, $\mathcal{R}$, is the fraction of relevant documents in the list and is defined as N_P/N_D ($\mathcal{R}$ = 1 when the list of documents and collection of documents are identical, $N_P = N_D$). The $F1$ measure is the harmonic mean of these two statistics,

$$F1 = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}} = \frac{2}{1/\mathcal{P}+1/\mathcal{R}}$$

Given a ranking of the documents in the collection, D lists of documents are produced by selecting the top P-ranked documents where P = 1, ... ,D. For the dth list, the precision, $\mathcal{P}_d$, recall, $\mathcal{R}_d$, and F1 measure, $F1_d$ are computed. The average of the D Fl measures for the collection, F_1, ..., F_D, is calculated.

Three different models were assessed by determining their respective retrieval capability on the same set of relevant documents present in the collection of D = 5, 225 CGC documents (an LDA model, a mixture of unigrams model, and a model which ranks documents randomly). The common subject matter of the relevant documents was genes implicated in extending or shortening the life span of *C. elegans*. This set of N_D = 842 aging-related documents was created by identifying CGC items in which the Genes

record (Figure 1) refers to one or more of the genes listed in Table 1 that are known to modify life span.

A K = 50 topic LDA model was trained using all D CGC documents. Given this model, the symmetrized KL divergence between an aging-related document posterior (γ_q) and every CGC document posterior (γ_t) was computed using Equations 2 and 3. These D topic-space pairwise similarity scores, $\mathcal{S}(q, t_1)$, ..., $\mathcal{S}(q, t_D)$, were used to rank the CGC documents from most (highest score) to least similar. This process was repeated for each relevant document in turn, q_1, ..., q_{N_D}, resulting in N_D rankings of the CGC documents. Let P_i be the P top-ranked CGC documents for query q_i. The number of known aging-related documents in this list was determined and used to calculate precision and recall. The average precision (average recall) is the mean of the precision (recall) value computed for each list P_1,..., P_{N_D}. A series of such average precision and average recall values were computed by varying the number of top-ranked documents used to create lists, P = 1, ..., D. A PR curve was constructed from these D pairs of average precision and average recall values. The average F1 measure is the mean of the F1 measure computed for each of the N_D relevant documents.

A K = 50 topic mixture of unigrams model was trained using all D CGC documents. Given this model, the symmetrized KL divergence between an aging-related document $\mathbf{w}_q$ and a CGC document $\mathbf{w}_t$ was computed using

$$\mathcal{M}(\mathbf{w}_q, \mathbf{w}_t) = \frac{1}{2} KL(\mathbf{w}_q \,||\, \mathbf{w}_t) + \frac{1}{2} KL(\mathbf{w}_t \,||\, \mathbf{w}_q). \quad (7)$$

In the mixture model, the divergence is between posteriors of the single latent topic rather than the vector of latent topic proportions as is the case with LDA. The component KL term between documents $\mathbf{w}_i$ and $\mathbf{w}_j$ is calculated using

$$KL(\mathbf{w}_i \,||\, \mathbf{w}_j) = \sum_{z=1}^{K} p(z \mid \mathbf{w}_i) log \frac{p(z \mid \mathbf{w}_i)}{p(z \mid \mathbf{w}_j)}. \quad (8)$$

where $p(z|\mathbf{w}_i)$ and $p(z|\mathbf{w}_j)$are the posterior probabilities of class z given documents $\mathbf{w}_i$ and $\mathbf{w}_j$ respectively. The mixture model pairwise similarity scores, $\mathcal{M}(\mathbf{w}_{q'}, \mathbf{w}_{t_1})$,..., $\mathcal{M}(\mathbf{w}_{q'}, \mathbf{w}_{t_D})$ were used to rank the CGC documents from most (highest score) to least similar and the procedure repeated for each of the N_D relevant documents. The resultant N_D rankings were used to construct a PR curve and calculate average F1 as described above for the LDA model.

A random model was created by randomly ordering the D CGC documents and repeating this process N_D times. These N_D "rankings" of the D documents in the collection were used to construct a PR curve and calculate averate F1 as described above for the LDA model.

Authors' contributions

DMB, MIJ, and ISM conceived and designed the study, analyzed the results, and wrote the manuscript. DMB formulated algorithms, wrote the software, and performed the experiments. KF wrote the software needed for the Gene Ontology-based analysis of LDA topics. All authors approved the manuscript.

Additional material

Additional File 1

Results for each of the LDA topics specified by a 50-topic model estimated from a corpus of 5,225 documents and a 28,971 word vocabulary.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-7-250-S1.pdf]

Acknowledgements

This work was supported by the National Science Foundation, Office of Naval Research, National Institute on Aging, National Institute of Environmental Health Sciences, U.S. Department of Energy (OBER) and California Breast Cancer Research Program. DMB was additionally supported by a Fellowship from the Microsoft Corporation.

References

1. **Entrez Gene** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene]
2. **Online Mendelian Inheritance in Man (OMIM)** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM]
3. **Gene Ontology (GO)** [http://www.geneontology.org/]
4. **BLAST** [http://www.ncbi.nlm.nih.gov/BLAST/]
5. **Ensembl** [http://www.ensembl.org]
6. **UCSC Genome Browser** [http://genome.ucsc.edu]
7. **Gene Expression Omnibus (GEO)** [http://www.ncbi.nlm.nih.gov/geo/]
8. MacCallum R, Kelley R, Sternberg M: **SAWTED: Structure Assignment With Text Description – Enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons.** *Bioinformatics* 2000, **16:**125-129.
9. Jenssen T, Laegreid A, Komorowski J, Hovig E: **A literature network of human genes for high-throughput analysis of gene expression.** *Nature Genetics* 2001, **28:**21-28.
10. Raychaudhuri S, Chang J, Imam F, Altman R: **The computational analysis of scientific literature to define and recognize gene expression clusters.** *Nucleic Acids Research* 2003, **31:**4553-4560.
11. Shatkay H, Feldman R: **Mining the biomedical literature in the genomic era: an overview.** *J Comput Biol* 2003, **10:**821-855.
12. Hirschman L, Park J, Tsuji J, Wong L, Wu C: **Accomplishments and challenges in literature data mining for biology.** *Bioinformatics* 2002, **18:**1553-1561.
13. Yandell M, Majoros W: **Genomics and natural language processing.** *Nature Reviews Genetics* 2002, **3:**601-610.

14. Manning C, Schütze H: *Foundations of Statistical Natural Language Processing* Cambridge, MA: MIT Press; 1999.
15. **BioNLP** [http://www.bionlp.org]
16. **Textpresso** [http://www.textpresso.org/]
17. **Telemakus** [http://www.telemakus.net/]
18. Libbus B, Kilicoglu H, Rindflesch T, Mork J, Aronson A: **Using Natural Language Processing, LocusLink and the Gene Ontology to Compare OMIM to MEDLINE.** In *BioLink 2004: Linking Biological Literature, Ontologies and Databases* Association for Computational Linguistics; 2004:69-76.
19. Korbel J, Doerks T, Jensen L, Perez-Iratxeta C, Kaczanowski S, Hooper S, Andrade M, Bork P: **Systematic association of genes to phenotypes by genome and literature mining.** *PLoS Biol* 2005, **3(5):**e134.
20. Baeza-Yates R, Ribeiro-Neto B: *Modern Information Retrieval* New York: ACM Press; 1999.
21. Alter O, Brown P, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci* 2000, **97:**10101-10106.
22. Blei DM, Ng AY, Jordan MI: **Latent Dirichlet Allocation.** *Journal of Machine Learning Research* 2003, **3:**993-1022.
23. Blei D, Jordan M: **Modeling annotated data.** In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* ACM Press; 2003:127-134.
24. Griffiths T, Steyvers M: **Finding scientific topics.** *Proc Natl Acad Sci* 2004, **101:**5228-5235.
25. **Wormbase** [http://www.wormbase.org]
26. Rothman J: **Aging: from radiant youth to an abrupt end.** *Current Biology* 2002, **12:**R239-R241.
27. Kurz D, Hong Y, Trivier E, Huang H, Decary S, Hong Z, Luscher T, Erusalimsky J: **Fibroblast Growth Factor-2, But Not Vascular Endothelial Growth Factor, Upregulates Telomerase Activity in Human Endothelial Cells.** *Arterioscler Thromb Vasc Biol* 2003, **23:**748-754.
28. Bissell M, Radisky D: **Putting tumours in context.** *Nat Rev Cancer* 2001, **1:**46-54.
29. Askree S, Yehuda T, Smolikov S, Gurevich R, Hawk J, Coker C, Krauskopf A, Kupiec M, McEachern M: **A genome-wide screen for *Saccharomyces cerevisiae* deletion mutants that affect telomere length.** *Proc Natl Acad Sci* 2004, **101:**8658-8663.
30. Rog O, Smolikov S, Krauskopf A, Kupiec M: **The yeast VPS genes affect telomere length regulation.** *Current Genetics* 2005, **47:**18-28.
31. Durbin R, Eddy S, Krogh A, Mitchison G: *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* Cambridge: Cambridge University Press; 1998.
32. Bateman A, Birney E, Durbin R, Eddy S, Howe K, Sonnhammer E: **The Pfam protein families database.** *Nucleic Acids Research* 2000, **28:**263-266.
33. Blei D, Griffiths T, Jordan M, Tenenbaum J: **Hierarchical topic models and the nested Chinese restaurant process.** In *Neural Information Processing Systems Volume 16*. MIT Press, Cambridge MA; 2003.
34. Jordan M: **Graphical models.** *Statistical Science* 2004, **19:**140-155.
35. ***Caenorhabiditis* Genetic Center Bibliography** [http://elegans.swmed.edu/wli/cgcbib]
36. **C implementation of LDA** [http://www.cs.princeton.edu/~blei/lda-c]
37. Griffiths T, Steyvers M: **A probabilistic approach to semantic representation.** *Proceedings of the 24th Annual Conference of the Cognitive Science Society* 2002.
38. Minka T, Lafferty J: **Expectation-propagation for the generative aspect model.** *Uncertainty in Artificial Intelligence (UAI)* 2002.
39. Nigam K, McCallum A, Thrun S, Mitchell T: **Text classification from labeled and unlabeled documents using EM.** *Machine Learning* 2000, **39:**103-134.
40. Ponte J, Croft B: **A Language Modeling Approach to Information Retrieval.** *ACM SIGIR 1998* 1998:275-281.
41. Moler E, Chow M, Mian I: **Analysis of molecular profile data using generative and discriminative methods.** *Physiological Genomics* 2000, **4:**109-126.
42. Moler E, Radisky D, Mian I: **Integrating naïve Bayes models and external knowledge to examine copper and iron homeostasis in *Saccharomyces cerevisiae*.** *Physiological Genomics* 2000, **4:**127-135.
43. Bhattacharjee A, Richards W, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E, Lander E, Wong W, Johnson B, Golub T, Sugarbaker D, Meyerson M: **Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.** *Proc Natl Acad Sci* 2001, **98:**13790-13795.
44. Teh YW, Jordan MI, Beal MJ, Blei DM: **Hierarchical Dirichlet processes.** *JAMA (in press)* .