A Bayesian Translational Framework for Knowledge Propagation, Discovery, and Integration Under Specific Contexts

Michelle Deng[1,*], Amin Zollanvari[1,2,*], Gil Alterovitz[1,2,3]

[1] Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Boston, MA;[2] Center for Biomedical Informatics, Harvard Medical School, Boston, MA; [3] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. * These co-authors contributed equally to this work.

Abstract

The immense corpus of biomedical literature existing today poses challenges in information search and integration. Many links between pieces of knowledge occur or are significant only under certain contexts—rather than under the entire corpus. This study proposes using networks of ontology concepts, linked based on their co-occurrences in annotations of abstracts of biomedical literature and descriptions of experiments, to draw conclusions based on context-specific queries and to better integrate existing knowledge. In particular, a Bayesian network framework is constructed to allow for the linking of related terms from two biomedical ontologies under the queried context concept. Edges in such a Bayesian network allow associations between biomedical concepts to be quantified and inference to be made about the existence of some concepts given prior information about others. This approach could potentially be a powerful inferential tool for context-specific queries, applicable to ontologies in other fields as well.

1 Introduction and Objective

The millions of published works of biomedical literature cover an enormous array of knowledge. Over 21 million articles are indexed in PubMed alone, and around 700,000 new articles are added yearly[1]. Additionally, data from millions of experiments are archived in diverse databases. The large size of today's body of biomedical knowledge and swiftness with which new information is being added present challenges in organization and navigation. The rise of such a large amount of information in recent years is changing the nature of biological knowledge from a descriptive practice to a more data-driven one, and finding specific information through manual search is growing increasingly difficult.

Biomedical ontologies can potentially be used to address these challenges. Tremendous efforts have been made to create diverse ontologies that together include all biomedical concepts. The National Center of Biomedical Ontology (NCBO) BioPortal[2,3] provides over 250 such ontologies with over 5 million concepts[6] to researchers. Moreover, researchers use ontology terms to annotate experimental data and works of literature. Hence, an automated, efficient framework that navigates and integrates the information embedded in these ontological links would be a powerful research tool that utilizes an immense range of biomedical knowledge. However, ontologies are usually developed in a silo, and the separateness of ontologies has so far hindered the practical application of ontological organization. Hence, a crucial question remains unanswered: is it possible to automatically and efficiently use biomedical ontologies to infer new knowledge?

This work presents such an automated framework that integrates biomedical ontologies and infers knowledge from abstracts of literature and descriptions of experimental data in response to a user-defined query. In particular, this framework infers information particular to a given *context*, or situation. Context-specificity is useful because researchers often have questions relevant to specific situations, and the same biological concepts may be linked in some contexts but not in others. For example, two traits might not generally be observed together, but in the context of a specific genetic condition, they may coexist frequently. The proposed framework identifies these types of linkages.

2 Mapping Ontologies: An Overview

A logical first step is to integrate the disparate biomedical ontologies. We seek a reliable framework for mapping ontological relationships that (1) considers diverse types of relationships between terms, (2) accounts for uncertainty in ontology integration, (3) is scalable to the size of biomedical ontologies, and (4) is able to be tailored to specific contexts. So far, no such framework has been developed.

The Unified Medical Language System (UMLS)[4] has integrated over 2 million names for approximately 900,000 biological concepts. However, mappings of UMLS concepts were manually curated, so there remain inconsistencies and errors in the mappings, and it is difficult for mappings to keep pace with the rate at which knowledge is expanding[5]. Many non-manual methodologies exist for ontological integration, including semi-automatic methods such as PROMPT[7] and GLUE[8], and automatic methods such as IF-MAP[9], ANCHOR-PROMPT[10], and MAFRA[11]. In Chua et al.[12], more than 30 ontology mapping methods are surveyed and categorized into 7 categories. However, almost all proposed methods are not publicly available or are not scalable to the size of biomedical ontologies[13].

Two recent methods, Association Rule Ontology Matching Approach (AROMA)[14] and Lexical OWL Ontology Matcher (LOOM)[13], are publicly available and easily scalable. These two methods differ significantly. LOOM is used for discovering equivalence correspondences between concepts, is based on lexical matching, and does not require text corpora to work. In contrast, AROMA is used for inferring subsumption relationships between concepts, is based on a statistical measure known as implication intensity, and requires additional text corpora. Though these methods are steps forward, they do not consider the inevitable uncertainty of ontology mapping.

Some ontology-mapping studies do consider uncertainty by incorporating probabilistic uncertainty into their description logic by using Bayesian networks[15-19]. For example, a framework called OMEN[16] creates Bayesian networks of ontologies by drawing initial probabilities from *a priori* knowledge and then using a set of meta-rules to determine conditional probabilities between nodes. The conditional probabilities represent influences induced by nodes on their children. Two other algorithms, MSBN[17] and AEBN[18], create pairwise correspondences between semantically identical concepts and propagate information through these correspondences between two ontology-specific Bayesian networks. The algorithm BayesOWL[15] uses a process similar to those of MSBN and AEBN but is more comprehensive: it links similar concepts as well as identical concepts by defining the similarity of concepts probabilistically by their joint distribution. More methods for probabilistic modeling of uncertainty in linking ontologies can be found in Lukasiewicz[20]

3 Context-Specific Ontology Mapping and Probabilistic Inference: A Novel Technique

The method presented in this paper is distinct from the aforementioned ontology-mapping methods in its use of a context-sensitive algorithm. In prior work, a context-specific mapping algorithm based on the Bayes factor[21] was developed. This study adapts and applies that mapping method to construct the backbones of context-centered Bayesian networks for inference about biomedical relationships. This context-specific mapping approach has three main advantages: (1) mappings created are specific to the question under investigation, so unrelated concepts are pruned; (2) inference is less prone to noise generated from considering many unrelated concepts and can be more accurate; and (3) pruning many irrelevant concepts allows the inference algorithm to be scaled to the large size of most biomedical ontologies.

Once the Bayesian backbone is constructed, probabilistic inference on the framework accounts for conditional uncertainties in biological connections in the given context and gives more nuanced conclusions. The prior study[21] focused primarily on gathering the literature base and developing the Bayes factor to conduct univariate linkage analysis between terms; here, the Bayes factor is used as a tool to consider multivariate relationships and in a high-dimensional Bayesian network, leading to more nuanced and meaningful results.

4 Materials and Methods

The proposed framework constructs and analyzes networks based on knowledge embedded in ontological annotations of descriptions of experimental data and abstracts of published literature. After obtaining an annotated knowledge database, the framework comprises three main stages: (1) defining the query, (2) constructing a Bayesian graph based on that query, and (3) using that graph to perform probabilistic inference.

4.1 The Annotated Knowledge Base

Table 1: *Sources of records in the knowledge base.* Records were compiled in 2009. For all sources except PubMed, every existing record was included in the knowledge base. For efficiency, 100,000 PubMed records were randomly selected from the 16,000,000 existing at the time. A B-tree index was created on the records for searching.

Source Database	Records
Adverse Event Reporting System[14]	774,606
Array Express[15]	9281
BioSiteMaps[16]	1013
caNanoLab[17]	444
Conserved Domain Databases[18]	34,735
Clinical Trials Database[19]	75,828
Drug Bank[20]	4774
Database of Phenotypes and Genotypes[21]	75,828
Gene Expression Omnibus[22]	15,968
Stanford Microarray Database[23]	16,148
Published articles in PubMed[24]	100,000
Drug Bank[20]	4774

We prepared an indexed B-tree for searching the knowledge base that comprised annotated records from eleven corpora available from NCBO Bioportal in 2009 (Table 1). Then, 220 ontologies were obtained from the NCBO BioPortal; for caching sufficient statistics when searching through the literature, the dictionary of all available ontology concepts (4,153,358 terms) was obtained. More details on preparation of this B-tree and the ontology data are provided in Kshitij et al.[21]

4.2 Queries

In this work, a *query* consists of a concept of interest (the context under which linkages are identified) and two ontologies (containing the terms between which linkages are drawn). One ontology is designated the source ontology; the other is designated the destination ontology. Users define elements of their queries based on their applications. For example, a researcher interested in obesity-related phenotypes and genes might choose "obesity" for the context and Human Phenotype Ontology and Gene Ontology for the two ontologies.

4.3 Constructing the Network

Determining the Network Structure

Based on the query, a tree-augmented naïve (TAN) Bayesian network is constructed, where each node is a random variable that represents the state a specific concept takes in an annotation (either "exists" or "does not exist"). Nodes corresponding to concepts from the source ontology are the parents of nodes corresponding to concepts from the destination ontology. The root node corresponds to the context concept and is a parent of all nodes in the network. The TAN structure is adopted because its requirement that the root node is a parent of every other node parallels the way the context term is present when every ontological connection is identified. The structure is appropriate for context-specific inference.

The time complexity of learning a TAN structure from data using a maximally weighted spanning tree algorithm[25] is $O(n^2N)$, where n is the number of features (the number of concepts in both the source and the destination ontologies), and N is the number of samples[26]. In this study, the data is a large collection of literature annotated by ontologies. However, the large size of biomedical ontologies renders the use of the original TAN learning structure[25] infeasible. Hence, this framework uses a Bayes factor[23] (*BF*) to identify linkages between any source concept S and any destination concept D under a given context concept C[7]. The higher the Bayes factor, the larger the magnitude of association between two random variables. Therefore, we prune many weak linkages between S and D under C and map only S and D concepts that share a mutual *BF* greater than a threshold value.

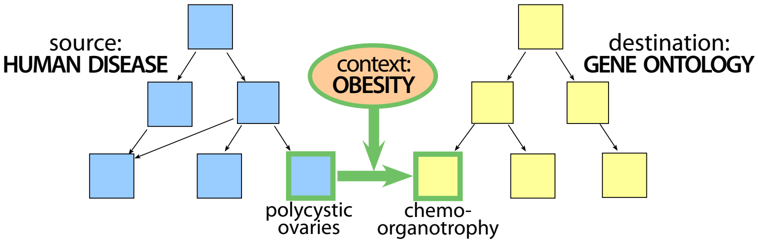

Figure 1: *The context-specific ontology-mapping process*. This diagram shows the source term "polycystic ovaries" being mapped to the destination term "chemoorganotrophy" under the context "obesity." Bold green arrows represent associations between concepts. In general, a source term and a destination term are mapped if and only if they are significantly associated with one another in annotations containing the context concept.

In order to calculate the *BF* between any S and D considering the context C, we first create a 2-by-2 contingency table. Each element of this table is determined from the frequencies of co-occurrences of S and D in literature that contain C. Let n be the number of documents, its subscript (S, D, or C) be the type of concept being counted, and the superscript (+ or –) be the state of the concept, where a plus sign (+) signifies "exists" and a minus sign (–) signifies "does not exist." The contingency table contains n^{++}, n^{+-}, n^{-+}, and n^{--}. Counts are obtained through full-text searches of the knowledge database (Table 1) and are used to calculate *BF* using the procedure described in Albert[27]. However, *BF* is not calculated for every pair of S and D: the hierarchical structure of ontologies allows a more efficient depth-first branch-and-bound algorithm[28] to be used to traverse the two ontologies.

After all significantly co-occurring pairs of S and D under the context C are linked, in accordance to the TAN structure, every concept in the network is linked to the context concept as well. The same destination concept may appear several times, each time linked to a different source concept, because of the TAN requirement that nodes have no more than one non-root parent. The different instances of the same destination concept are not considered as one node because keeping them separate drastically facilitates probabilistic inference.

Associating Edges with Conditional Probabilities

The network must next be associated with probabilities. For each concept in the net, a table containing the conditional probabilities that each of its states ("exists" or "does not exist") is true is determined for all combinations of states its parent nodes can take. The conditional probability values are derived from the counts of different combinations of the states of the concept in question and its parents in annotations. For example, $P(C^{+})=n_{C^{+}}/(n_{C^{+}}+n_{C^{-}})$ is one context probability value, $P(S^{+}|C^{+})=n_{S^{+}C^{+}}/(n_{S^{+}C^{+}}+n_{S^{-}C^{+}})$ is one source probability value, and $P(D^{+}|S^{+}C^{+})=n_{D^{+}S^{+}C^{+}}/(n_{D^{+}S^{+}C^{+}}+n_{D^{-}S^{+}C^{+}})$ is one destination probability value. Queries are performed, and counts are collected in the same way as when calculating the Bayes factor to build the network structure. Based on transitive closure of concepts in ontologies, we used the same depth-first branch-and-bound procedure described in Kshitij et al.[21] to prune the ontologies and cache the statistics. This pruning makes the Bayesian network construction efficient enough and scalable to the size of biomedical ontologies.

The final product is thus a three-tiered TAN Bayesian network with the context term at the root, source ontology terms as intermediates, and destination ontology terms as the leaves, related to one another by conditional probabilities based on the frequencies of their co-occurrence in annotations of literature and of experimental data.

The power of these networks comes from Bayesian inference. Because nodes are linked by probabilities, given the prior probability distribution of the root nodes, predictions can be made about the states of any of the other nodes. In this study, Pearl's message-passing algorithm in trees[25] is implemented so that state information about one or more nodes can propagate along the graph edges and influence the probabilities of the states of other nodes. For example, if certain biological concepts are known to be affected, expressed, or active, the nodes corresponding to those concepts are set to true (*P*(exists) = 1). The tree is then updated to reflect this new knowledge, and *P*(exists) values of all other nodes change accordingly.

4.4 Probabilistic Inference on the Network

Identifying Inter-Concept Linkages Using Belief Propagation

The power of these networks lies in Bayesian inference. Because nodes are linked by probabilities, predictions can be made about the states of any of the other nodes given the prior probability distribution of the root nodes. Pearl's message-passing algorithm[29] is implemented so that state information about one or more nodes can propagate along the graph edges and influence the probabilities of the states of other nodes. For example, if certain biological concepts are known to be affected, expressed, or active, the nodes corresponding to those concepts are set to true (*P*(exists) = 1). The tree is then updated to reflect this new knowledge, and *P*(exists) values of all other nodes change accordingly.

One application of the constructed networks and the proposed inference algorithm is the identification of source or destination concepts in the network that are related to the context *C*. To measure the relatedness of a term *T* to the context, we associate it with a likelihood ratio *L*:

$$L=\frac{P(T\ \text{exists}|C\ \text{exists})}{P(T\ \text{exists}|C\ \text{does not exist})} \tag{1}$$

To calculate *L*, state of the context node is set to "exists," and the states of all other nodes are left unknown. Beliefs are then propagated, and $P(T\ \text{exists}|C\ \text{exists})$ is found for each node. The context node is then set to "does not exist," and the other nodes are still left with unknown states. Again, beliefs are propagated, and $P(T\ \text{exists}|C\ \text{does not exist})$ is found for each node. *L* is the ratio of those two probabilities. It measures how much more likely it is that *T* is true when *C* is true than when *C* is false, not simply how likely it is that the two terms coexist. Hence, a general term such as "disease/disorder" would not score a high *L* because there would be little difference in the probability that it exists whether or not the context is true. That is, the terms with the highest *L* are most likely to be related specifically to the context and are therefore terms of interest.

A *p*-value can also be found for each *T*-*C* link. First, the Bayes factor that *T* and *C* are associated is determined again in the manner described in Section 4.2, except this time the contingency table contains $n_{T^{+}C^{+}}$, $n_{T^{+}C^{-}}$, $n_{T^{-}C^{+}}$, and $n_{T^{-}C^{-}}$. Using that *BF*, an upper bound for the *p*-value can be determined as follows[36], where $p < 1/e$:

$$-p\ln p<\frac{1}{BF\cdot e}$$

This study examines the terms with the highest *L* as the "most related" and then separately associates them with p-values for ease of understanding. This choice was made because *L*, which is based on probabilistic propagation over the network, considers all terms in a high-dimensional joint distribution, whereas using Bayes factor or p-value as the final mapping is essentially a univariate, deterministic linkage from source to context. Without using *L*, the benefits of considering intricate, multivariate biomedical relationships represented in the network would be lost.

Extending the Network: Gene Inference

Inference using the Bayesian framework is not limited to identification of connections between ontological concepts. For example, the framework can be used to identify the genes most relevant to a queried context. To do so, networks are built with Gene Ontology (GO) as one of the two queried ontologies, and GO concepts in the network are linked with relevant genes based on gene set information from MSigDb[46]. Because links between genes and GO concepts are deterministic, this additional gene level is not actually part of the probabilistic inference framework, and genes do not correspond to network nodes. Therefore, the inference procedure for finding genes relevant to a given context cannot rely on belief propagation. Instead, relatedness of genes to the context are determined based on network structure alone. For each gene, a one-sided Fisher's exact test is used to determine whether there is a significant difference between the proportion of GO terms in the network (which are presumably related to the context) that are associated with that gene and the proportion of GO terms outside the network (which are conversely not strongly related to the context) that are associated with the gene. The Benjamini-Hochberg method was used to control for the false discovery rate from multiple hypotheses[47].

5 Results

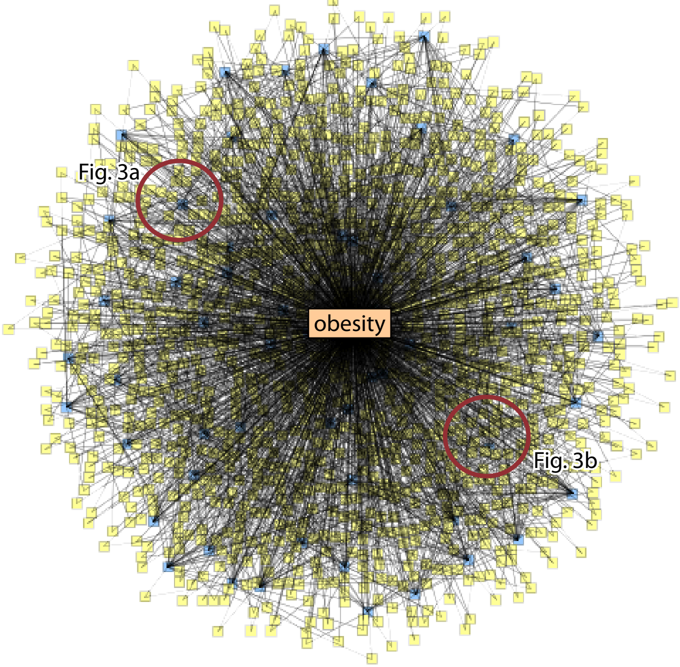

Figure 2: *TAN Bayesian network constructed for context "obesity."* Yellow nodes represent GO (destination) terms; light blue nodes represent DOID (source) terms, and the orange node represents the context. Displayed are 4401 out of the 447374 total edges in the network; the full network contains 240 unique nodes from DOID and 8218 unique nodes from GO. Magnified views of the circled subnets are provided in Fig. 3.

In this work, we use the proposed learning network structure and inference procedures to identify diseases and genes related to specific pathologies of interest. The context concept is set to be the pathology, the source ontology is set to be Human Disease (DOID), and the destination ontology is set to be Gene Ontology (GO). Both ontologies are available through NCBO BioPortal[2,3]. That way, all biomedical relationships represented in the Bayesian network are specific to the context or pathology of interest. Since, as explained in Section 4, inference is done over this context-specific network, all identified relationships are tailored for the context.

Networks were constructed and analyzed for diverse context pathologies, including several cancers, substance abuse disorders, obesity and heart disease, and HIV/AIDS. Similar patterns were observed in the results of all contexts; results from contexts "alcoholism" and "obesity" are discussed further in this paper. The network built using context "obesity" is shown in Figure 2. In the rest of Section 5, italicized body text represents terms or genes identified by the algorithm as associated with the context.

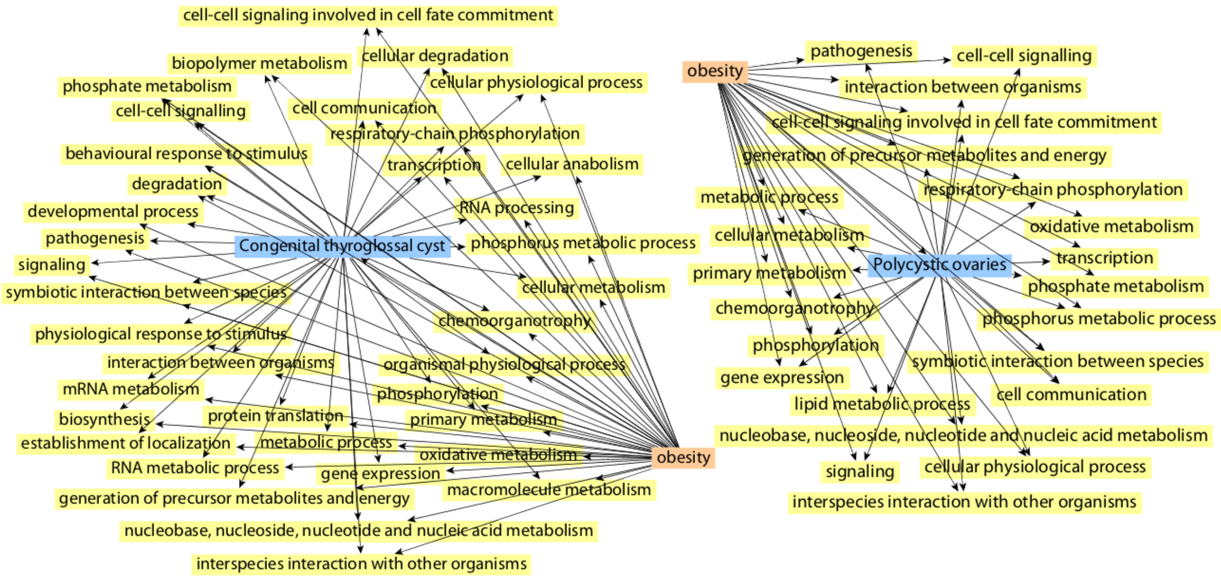
cell-cell signaling involved in cell fate commitment
biopolymer metabolism
cellular degradation
phosphate metabolism
cellular physiological process
cell-cell signalling
cell communication
respiratory-chain phosphorylation
behavioural response to stimulus
transcription
cellular anabolism
degradation
developmental process
RNA processing
pathogenesis
Congenital thyroglossal cyst
phosphorus metabolic process
signaling
symbiotic interaction between species
cellular metabolism
chemoorganotrophy
physiological response to stimulus
interaction between organisms
organismal physiological process
mRNA metabolism
phosphorylation
biosynthesis
protein translation
primary metabolism
establishment of localization
metabolic process
oxidative metabolism
obesity
RNA metabolic process
gene expression
macromolecule metabolism
generation of precursor metabolites and energy
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
interspecies interaction with other organisms
obesity
pathogenesis
cell-cell signalling
interaction between organisms
cell-cell signaling involved in cell fate commitment
generation of precursor metabolites and energy
metabolic process
respiratory-chain phosphorylation
oxidative metabolism
cellular metabolism
transcription
primary metabolism
Polycystic ovaries
phosphate metabolism
chemoorganotrophy
phosphorus metabolic process
phosphorylation
symbiotic interaction between species
gene expression
lipid metabolic process
cell communication
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
signaling
cellular physiological process
interspecies interaction with other organisms

Figure 3: *Magnified views of circled subnets in Fig. 2*. Nodes are color-coded as in Fig. 2. **a)** This subnet is centered about the DOID (source) concept "congenital thyroglossal cyst," a parent of numerous GO (destination) concepts. **b)** This subnet has the DOID (source) concept "congenital thyroglossal cyst." True to the TAN structure, the context node "obesity" is a parent of all other nodes as well.

5.1 Alcoholism

Disease Term	***L***	***BF***	***p*-value**
Drug abuse	340.755	12.949	4.187×10^{-4}
Alcohol-related disorder NOS	294.652	12.460	5.718×10^{-3}
Substance-related disorder	185.522	11.309	6.450×10^{-3}
Disease of environmental origin	74.763	8.707	8.962×10^{-3}
Environmentally induced disease	74.763	8.707	8.962×10^{-3}
Addiction	68.806	10.934	6.726×10^{-3}
Schizophrenia	24.199	5.157	1.768×10^{-2}
Alzheimer's dementia	16.957	3.400	3.121×10^{-2}
Organic mental disorder of unknown etiology	6.543	6.341	1.345×10^{-2}
Tauopathies	5.114	6.007	1.445×10^{-2}

Table 2. The ten Disease Ontology (source) terms identified as most strongly linked to the context "alcoholism."

The disease concepts with the strongest links to the context "alcoholism" using (1) as a measure of link strength are indeed closely biologically related to alcoholism (Table 2). Alcoholism is a *substance-related disorder,* is a form of *addiction*, and would be associated with *alcohol-related disorders NOS*. Alcohol consumption is known to interfere with the nervous system, leading to impaired perception, coordination, memory, and judgment, all possible components of *organic mental disorder of unknown etiology*[38,39]. Psychotic disorders such as *schizophrenia* occur more frequently in alcoholics than in nonalcoholics[40,41], and alcohol consumption can lead to *tauopathies*, diseases involving aggregation of abnormal tau protein in the brain[42], such as *Alzheimer's dementia*[43]. Moreover, environmental factors such as socioeconomic status or education quality play major roles in the development of alcoholism, a *disease of environmental origin* or *environmentally induced disease*[44], and there exists a high comorbidity between *drug abuse* problems and alcoholism[45].

The gene inference procedure (Section 4.4) identified many promising genes as significant. For example, a number of genes had already been found by other studies to be associated with the context of alcoholism, including *PTGDS* ($P < 10^{-15}$), the gene with the lowest *p*-value; *MIF* ($P < 10^{-13}$); *BRCA1* ($P < 10^{-13}$); *IL4* ($P < 10^{-7}$); and the three types of peroxisome proliferator-activated receptor genes (PPARs), *PPARA* ($P < 10^{-15}$), *PPARD* ($P < 10^{-15}$), and *PPARG* ($P < 10^{-5}$). *PTGDS* codes for prostaglandin D2 synthase, which is negatively correlated with alcohol intake[48]. In liver tissues affected by alcoholic liver disease, serum levels of macrophage migration inhibitory factor, coded by *MIF,* are elevated[49], while alcohol inhibits *IL4*, which controls B-cell proliferation and immunoglobulin class switching[50,51]. Alcohol consumption is associated with heightened incidence of breast cancer, and ethanol down-regulates *BRCA1*, the second most likely gene, of which mutations are closely linked to breast cancer[52,53]. Both *PPARA* and *PPARD* are downregulated by ethanol, *PPARD* agonists alleviate alcohol-induced liver damage, and *PPARG* activation may suppress addictive drinking behaviors[54]. Other significant genes, such as the transcription factor gene *TCF7* ($P = 6.027 \times 10^{-10}$), have not yet been linked in a molecular biological study to ethanol; however, they have been found in other bioinformatics studies to be significantly associated with ethanol or with alcohol withdrawal[55] and therefore are encouraging targets for future biological studies of alcohol dependence.

6.2 Obesity

When the same procedure was conducted with the context "obesity," the algorithm just as successfully identified diseases relevant to the context concept (Table 3). *Obesity, unspecified* is a synonym of the concept itself; *morbid obesity*, defined as weighing 45 kg or more above the ideal weight or having a BMI of at least 40, is a subset of the context[56]. *Polyphagia*, an *eating disorder* characterized by excessive consumption of food, can cause weight gain and lead to obesity. Alcohol intake is another potential cause of obesity[57], and both alcohol and obesity are associated with fatty liver disease[58] (*alcoholic liver damage, alcohol induced liver disorder*). Obesity increases the risk of *cholelithiasis*, the development of gallstones, especially during the weight loss process[59] and is highly associated with polycystic ovary disease (*ovarian dysfunction*, *ovarian non-neoplastic disease*), with around 30% of individuals with polycystic ovary disease being obese[60]. There exists a genetic disorder, Ayazi syndrome, characterized by obesity, *choroideremia*, and congenital deafness[61].

Similar to the case of the context "alcoholism," the proposed method identified as significant a promising mix of already-corroborated and potentially-related genes for the context "obesity." For instance, *TGFB1* ($P < 10^{-7}$), a tissue growth factor that regulates proliferation, migration, and differentiation of diverse cells, is linked to abdominal obesity and insulin and glucose imbalance[62]. *PPARD* ($P < 10^{-7}$) activates other genes that direct fatty acid catabolism and thermogenesis; underexpression of *PPARD* results in obesity[63], while *PPARD* agonists mimic exercise and make promising targets for treatment of metabolic syndromes[63-65]. *UBB* ($P < 10^{-6}$) is one of several genes that codes for ubiquitin, a protein-recycling regulator involved in lipid metabolism and whose levels are inversely associated with BMI[66]. Indeed, mice lacking the *UBB* gene exhibit adult-onset obesity[67]. *CARTPT* ($P < 10^{-6}$) encodes hypothalamic satiety factors[66], the dysregulation of which may lead to overeating, and *FADS1* ($P < 10^{-10}$), which codes for fatty acid desaturase, is related to lipid metabolism and the plasma triacylglycerol response[69]; both genes easily might relate to obesity. One interesting find was *YWHAH* ($P < 10^{-9}$). Polymorphisms of *YWHAH* are associated with schizophrenia, and antipsychotic drugs[70], including schizophrenia medications, are known to induce obesity[71]. Perhaps, *YWHAH* is a missing link in knowledge that this method has identified.

Disease Term	***L***	***BF***	**p-value**
Obesity, unspecified	12225.085	630.816	5.999×10^{-4}
Polyphagia	11458.037	20.524	3.102×10^{-3}
Morbid obesity	5790.621	11.878	6.067×10^{-3}
Alcoholic liver damage, unspecified	1416.854	7.704	1.047×10^{-2}
Eating disorder, unspecified	163.686	16.905	3.929×10^{-3}
Cholelithiasis	123.205	6.733	1.246×10^{-2}
Choroideremia	123.205	6.733	1.246×10^{-2}
Alcohol induced liver disorder	113.348	6.733	1.246×10^{-2}
Ovarian dysfunction	93.180	15.432	4.392×10^{-3}
Ovarian non-neoplastic disease	93.180	15.432	4.392×10^{-3}

Table 3. The ten Disease Ontology (source) terms identified as most strongly linked to the context "obesity."

6 Discussion and Future Work

Our technique can be seen as the first "automatic" probabilistic inference algorithm that uses large biomedical ontologies in conjunction with the vast corpus of existing biomedical literature and experimental data to address specific queries. Of the many probabilistic Bayesian frameworks proposed so far, only this one uses context-specific formulae to map concepts and to calculate conditional probabilities and specializes on a context-specific Bayesian network structure for inference. Therefore, inference using this technique is customized for the researcher's interests than inference using previous methods.

In this particular work, the proposed framework was effectively used to identify disease concepts and genes related to a context pathology of interest using existing knowledge embedded in literature and in ontologies. Identified disease concepts were invariably closely related to the context. Many of the genes the method identified likewise were known to be associated with the context pathology. Of the remaining genes, many had functions that could logically link them to the context concept or had been identified by other bioinformatics studies as differentially expressed in individuals exhibiting the context pathology. Such genes are promising and interesting because they may constitute new links that augment existing knowledge. The disease concepts and genes identified here may seem to be new information for a researcher with a specific query but no prior information.

All inferences are drawn from the annotated knowledge base that the framework uses as data. Therefore, it is critical that the annotation methods and the selection of data and literature included are comprehensive and representative. We must assume that our knowledge base satisfies the previous condition. Nonetheless, this work advances our ability to generate inferences from such bases. Because literature, experimental data, and ontologies continually evolve, the database must be up-to-date to comprehensively use the prior biomedical knowledge available. We intend to fully automate the data-preparation process in the future and integrate it with the inference framework presented here.

The proposed algorithm can be enhanced to improve its utility as an inferential tool. In this work, a query consists of a context concept of interest and two ontologies of terms from which connections are drawn. In future studies, we intend to extend the algorithm to be able to handle more complex queries. Additionally, future work can examine the predictive power of the framework in identifying drug-disease, drug-pathway, and pathway-disease relationships.

7 Acknowledgements

This work was supported by NIH grants 5R21DA025168-02 (G. Alterovitz), 1R01HG004836-01 (G. Alterovitz), and 4R00LM009826-03 (G. Alterovitz)

References

1. MEDLINE® Citation Counts by Year of Publication. Available at http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html (08/10/2011).
2. Musen MA, Shah NH, Noy NF, Dai BY, Dorf M, Griffith N, Buntrok J, Jonquet C, Montegut MJ, Rubin DL: BioPortal: ontologies and data resources with the click of a mouse. AMIA Annu Symp Proc 2008:1223-1224.
3. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG et al: BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res 2009, 37(Web Server issue):W170-173.
4. Bodenreider O: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004, 32(Database issue):D267-270.
5. Morrey CP, Geller J, Halper M, Perl Y: The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS. J Biomed Inform 2009, 42(3):468-489.
6. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011, 39(Web Server issue):W541-545.
7. Noy NF, Musen M: PROMPT: algorithm and tool for automated ontology merging and alignment. Proceedings of the 17th National Conference on Artificial Intelligence 2000.
8. Doan A, Madhavan J, Domingos P, Halevy A: Learning to map between ontologies on the semantic web. Proceedings of the 11th International World Wide Web Conference 2002.
9. Kalfoglou Y, Schorlemmer M: Ontology Mapping: the State of the Art. The Knowledge Engineering Review 2003, 18(1).
10. Noy NF, Musen MA: Anchor-PROMPT: Using non-local context for semantic matching. Proceedings of the Workshop on Ontologies and Information Sharing at (IJCAI 2001), 2001:63-70.
11. Maedche A, Motik B, Silva N, Volz R: MAFRA - A MApping FRAmework for Distributed Ontologies. In: Proceedings of the EKAW 2002.
12. Chua WWK, Goh AES: Techniques for discovering correspondences between ontologies. Int J Web and Grid Services, 2010, 3.
13. Ghazvinian A, Noy NF, Musen MA: Creating Mappings For Ontologies in Biomedicine: Simple Methods Work. 009 AMIA Annual Symposium, 2009.
14. David J, Guillet F, Briand H: Association rule ontology matching approach. Int J on Semantic Web and Information Systems 2007, 3(2):27-49.
15. Ding Z, Peng Y, Pan R: BayesOWL: Uncertainty modeling in Semantic Web ontologies. In: Soft Computing in Ontologies and Semantic Web, volume 204 of Studies in Fuzziness and Soft Computing. Edited by Ma Z; 2006.
16. Mitra P, Noy NF, Jaiswal A: OMEN: A probabilistic ontology mapping tool In: In Proceedings ISWC. 2005: 537-547.

17. Pan R, Ding Z, Yu Y, Peng Y: A Bayesian network approach to ontology mapping. In: In Proc ISWC. vol. 3729; 2005: 563-577.
18. Valtorta M, Kim Y, Vomlel J: Soft Evidential Update for Probabilistic Multiagent Systems. International Journal Approximate Reasoning 2002, 29(1):71-106.
19. Xiang Y: Probabilistic Reasoning in Multiagent Systems: A Graphical Models Approach.: Cambridge University Press; 2002.
20. Lukasiewicz T: Probabilistic description logics for the semantic web. In: INFSYS Research Report; 2007: 1-24.
21. Kshitij M, Katzin D, Zollanvari A, Noy N. F. Ramoni M, Alterovitz G: Context-Specific Ontology Integration: A Bayesian Approach, accepted for publication in AMIA 2012 proceedings.
22. Friedman N, Geiger D, Goldszmidt M: Bayesian Network Classifiers. Machine Learning 1997, 29:131-163.
23. Chow CK, Liu CN: Approximating Discrete Probability Distributions with Dependence Trees. IEEE Transactions on Information Theory 1968, 14:462-467.
24. Albert J: Bayesian Computation with R, 2nd edn. Baltimore, MD: Springer; 2009.
25. Zhang W: Depth-first branch-and-bound versus local search: A case study. In: Proceedings of the 17th National Conference on Artificial Intelligence 2000:930-936.
26. Pearl J: Reverend Bayes on inference engines: A distributed hierarchical approach. In: Proceedings of the National Conference on Artificial Intelligence 1982:133-136.
27. Ross SD, Reynolds MW: Use of the FDA spontaneous adverse event reporting system (SAERS), or why your MedWatch reports really do matter. Journal of Clinical Ontology 2004, 22(14S).
28. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M et al: ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005, 33(Database issue):D553-555.
29. Marenco L, Wang R, Shepherd GM, Miller PL: The NIF DISCO Framework: facilitating automated integration of neuroscience content on the web. Neuroinformatics 2010, 8(2):101-112.
30. Maojo V, Martin-Sanchez F, Kulikowski C, Rodriguez-Paton A, Fritts M: Nanoinformatics and DNA-based computing: catalyzing nanomedicine. Pediatr Res 2010, 67(5):481-489.
31. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M et al: CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 2009, 37(Database issue):D205-210.
32. Mi M: Clinical Trials Database: Linking Patients to Medical Research. Journal of Consumer Health On the Internet 2005, 9(3):59-67.
33. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008, 36(Database issue):D901-906.
34. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L et al: The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007, 39(10):1181-1186.
35. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA et al: NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 2009, 37(Database issue):D885-890.
36. Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TB, Wymore F, Zachariah ZK, Sherlock G, Ball CA: Implementation of GenePattern within the Stanford Microarray Database. Nucleic Acids Res 2009, 37(Database issue):D898-901.
37. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U: AliBaba: PubMed as a graph. Bioinformatics 2006, 22(19):2444-2445.
38. Sellke T, Bayarri MJ, Berger JO: Calibration of p Values for Testing Precise Null Hypotheses. The American Statistician 2001, 55(1):62-71.
39. Adams LR, Parsons OA, Culbertson LJ, Nixon SJ: Neuropsychology for clinical practice: Etiology, assessment, and treatment of common neurological disorders; 1996.
40. Ammendola A, Tata MR, Aurilio C, Ciccone G, Gemini D, Ammendola E, Ugolini G, Argenzio F: Peripheral neuropathy in chronic alcoholism: a retrospective cross-sectional study in 76 subjects. Alcohol Alcohol 2001, 36(3):271-275.
41. Degenhardt L, Hall W, Lynskey M: Alcohol, cannabis and tobacco use among Australians: a comparison of their associations with other drug use and use disorders, affective and anxiety disorders, and psychosis. Addiction 2001, 96(11):1603-1614.
42. Regier DA, Farmer ME, Rae DS, Locke BZ, Keith SJ, Judd LL, Goodwin FK: Comorbidity of mental disorders with alcohol and other drug abuse. Results from the Epidemiologic Catchment Area (ECA) Study. JAMA 1990, 264(19):2511-2518.
43. Iqbal K, Alonso Adel C, Chen S, Chohan MO, El-Akkad E, Gong CX, Khatoon S, Li B, Liu F, Rahman A et al: Tau pathology in Alzheimer disease and other tauopathies. Biochim Biophys Acta 2005, 1739(2-3):198-210.
44. Ruitenberg A, van Swieten JC, Witteman JC, Mehta KM, van Duijn CM, Hofman A, Breteler MM: Alcohol consumption and risk of dementia: the Rotterdam Study. Lancet 2002, 359(9303):281-286.
45. Enoch MA: Genetic and environmental influences on the development of alcoholism: resilience vs. risk. Ann N Y Acad Sci 2006, 1094:193-201.
46. Stinson FS, Grant BF, Dawson DA, Ruan WJ, Huang B, Saha T: Comorbidity between DSM-IV alcohol and specific drug use disorders in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Drug Alcohol Depend 2005, 80(1):105-116.

47. Benjamini Y, Hochberg Y: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 1995, 57(1):289-300.
48. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci 2005, 102(43):15545-15550.
49. Wallenius V, Elias E, Bergstrom GM, Zetterberg H, Behre CJ: The lipocalins retinol-binding protein-4, lipocalin-2 and lipocalin-type prostaglandin D2-synthase correlate with markers of inflammatory activity, alcohol intake and blood lipids, but not with insulin sensitivity in metabolically healthy 58-year-old Swedish men. Exp Clin Endocrinol Diabetes 2011, 119(2):75-80.
50. Kumagi T, Akbar F, Horiike N, Onji M: Increased serum levels of macrophage migration inhibitory factor in alcoholic liver diseases and their expression in liver tissues. Clin Biochem 2001, 34(3):189-193.
51. Guerre-Millo M, Gervois P, Raspe E, Madsen L, Poulain P, Derudas B, Herbert JM, Winegar DA, Willson TM, Fruchart JC et al: Peroxisome proliferator-activated receptor alpha activators improve insulin sensitivity and reduce adiposity. J Biol Chem 2000, 275(22):16638-16642.
52. Aldo-Benson M, Pratt L, Hardwick J: Alcohol can inhibit effect of IL-4 on activated murine B cells. Immunol Res 1992, 11(2):117-124.
53. Dennis J, Krewski D, Cote FS, Fafard E, Little J, Ghadirian P: Breast Cancer Risk in Relation to Alcohol Consumption and BRCA Gene Mutations - A Case-Only Study of Gene-Environment Interaction. Breast J 2011.
54. Dennis J, Ghadirian P, Little J, Lubinski J, Gronwald J, Kim-Sing C, Foulkes W, Moller P, Lynch HT, Neuhausen SL et al: Alcohol consumption and the risk of breast cancer among BRCA1 and BRCA2 mutation carriers. Breast 2010, 19(6):479-483.
55. Venkata NG, Aung CS, Cabot PJ, Monteith GR, Roberts-Thomson SJ: PPARalpha and PPARbeta are differentially affected by ethanol and the ethanol metabolite acetaldehyde in the MCF-7 breast cancer cell line. Toxicol Sci 2008, 102(1):120-128.
56. Uddin RK, Singh SM: Ethanol-responsive genes: identification of transcription factors and their role in metabolomics. Pharmacogenomics J 2007, 7(1):38-47.
57. Brolin RE: Bariatric surgery and long-term control of morbid obesity. JAMA 2002, 288(22):2793-2796.
58. Colditz GA, Giovannucci E, Rimm EB, Stampfer MJ, Rosner B, Speizer FE, Gordis E, Willett WC: Alcohol intake in relation to diet and obesity in women and men. Am J Clin Nutr 1991, 54(1):49-55.
59. Mantena SK, King AL, Andringa KK, Eccleston HB, Bailey SM: Mitochondrial dysfunction and oxidative stress in the pathogenesis of alcohol- and obesity-induced fatty liver diseases. Free Radic Biol Med 2008, 44(7):1259-1272.
60. Shiffman ML, Sugerman HJ, Kellum JM, Brewer WH, Moore EW: Gallstone formation after rapid weight loss: a prospective study in patients undergoing gastric bypass surgery for treatment of morbid obesity. Am J Gastroenterol 1991, 86(8):1000-1005.
61. Dietz WH: Health consequences of obesity in youth: childhood predictors of adult disease. Pediatrics 1998 101(3):518-525.
62. Ayazi S: Choroideremia, obesity, and congenital deafness. Am J Opthalmol 1981, 92(1):63-9.
63. Rosmond R, Chagnon M, Bouchard C, Bjorntorp P: Increased abdominal obesity, insulin and glucose levels in nondiabetic subjects with a T29C polymorphism of the transforming growth factor-beta1 gene. Horm Res 2003, 59(4):191-194.
64. Evans RM, Barish GD, Wang YX: PPARs and the complex journey to obesity. Nat Med 2004, 10(4):355-361.
65. Luquet S, Lopez-Soriano J, Holst D, Gaudel C, Jehl-Pietri C, Fredenrich A, Grimaldi PA: Roles of peroxisome proliferator-activated receptor delta (PPARdelta) in the control of fatty acid catabolism. A new target for the treatment of metabolic syndrome. Biochimie 2004, 86(11):833-837.
66. Narkar VA, Downes M, Yu RT, Embler E, Wang YX, Banayo E, Mihaylova MM, Nelson MC, Zou Y, Juguilon H et al: AMPK and PPARdelta agonists are exercise mimetics. Cell 2008, 134(3):405-415.
67. Chang TL, Chang CJ, Lee WY, Lin MN, Huang YW, Fan K: The roles of ubiquitin and 26S proteasome in human obesity. Metabolism 2009, 58(11):1643-1648.
68. Ryu KY, Garza JC, Lu XY, Barsh GS, Kopito RR: Hypothalamic neurodegeneration and adult-onset obesity in mice lacking the Ubb polyubiquitin gene. Proc Natl Acad Sci U S A 2008, 105(10):4016-4021.
69. Altarejos JY, Goebel N, Conkright MD, Inoue H, Xie J, Arias CM, Sawchenko PE, Montminy M: The Creb1 coactivator Crtc1 is required for energy balance and fertility. Nat Med 2008, 14(10):1112-1117.
70. Mangravite LM, Dawson K, Davis RR, Gregg JP, Krauss RM: Fatty acid desaturase regulation in adipose tissue by dietary composition is independent of weight loss and is correlated with the plasma triacylglycerol response. Am J Clin Nutr 2007, 86(3):759-767.
71. Toyooka K, Muratake T, Tanaka T, Igarashi S, Watanabe H, Takeuchi H, Hayashi S, Maeda M, Takahashi M, Tsuji S, Kumanishi T, Takahashi Y: 14-3-3 protein η chain gene (YWHAH) polymorphism and its genetic association with schizophrenia. Am J Med Genet 1999, 88(2):164-167.
72. Silverstone T, Smith G, Goodall E: Prevalence of obesity in patients receiving depot antipsychotics. Br J Psychiatry 1988, 153:214-217.