Evaluation of an Ontology-anchored Natural Language-based Approach for Asserting Multi-scale Biomolecular Networks for Systems Medicine

Tara B. Borlawsky, MA[1]; Jianrong Li[2]; Lyudmila Shagina[3]; Matthew G. Crowson[2]; Yang Liu, PhD[2]; Carol Friedman, PhD[3,*]; Yves A. Lussier, MD[2,*]

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH
[2]Center for Biomedical Informatics, Dept. of Medicine, The University of Chicago, IL
[3]Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

The ability to adequately and efficiently integrate unstructured, heterogeneous datasets, which are incumbent to systems biology and medicine, is one of the primary limitations to their comprehensive analysis. Natural language processing (NLP) and biomedical ontologies are automated methods for capturing, standardizing and integrating information across diverse sources, including narrative text. We have utilized the BioMedLEE NLP system to extract and encode, using standard ontologies (e.g., Cell Type Ontology, Mammalian Phenotype, Gene Ontology), biomolecular mechanisms and clinical phenotypes from the scientific literature. We subsequently applied semantic processing techniques to the structured BioMedLEE output to determine the relationships between these biomolecular and clinical phenotype concepts. We conducted an evaluation that shows an average precision and recall of BioMedLEE with respect to annotating phrases comprised of cell type, anatomy/disease, and gene/protein concepts were 86% and 78%, respectively. The precision of the asserted phenotype-molecular relationships was 75%.

Introduction

Unlike the traditional, reductionist practice of clinical medicine that independently examines individual components, systems medicine approaches focus on the dynamic interactions among multiple factors that affect complex diseases, such as diabetes, coronary artery disease and cancers[1]. The increasing availability of powerful high-throughput technologies, computational tools and integrated knowledge bases, has made it possible to establish new links between genes, biologic functions and human diseases, providing the hallmarks of systems medicine, including signatures of pathology biology, and links to clinical research and drug discovery[2]. Holistic systems biology methodologies promise to provide the foundation for such prospective medicine through the construction of integrated biomolecular networks[3]. However, one of the primary limitations to such an approach is the availability of integration methodologies for combining diverse types of data and generating knowledge bases that are precise and detailed enough to derive testable hypotheses across different scales of biology[2]. *To address this gap in knowledge, we have evaluated a natural language and semantic processing-based approach for generating integrated biomolecular and phenotypic data sets from existing published literature and biomedical ontologies.*

Background

Biomedical Ontologies

One key to the emergence of systems medicine will be the ability to harness the vast amounts of biomolecular and phenotypic data produced by high-throughput technologies and advanced measurement techniques[3,4]. Community efforts for the integrative annotation of such data sets include combining automated computation with human-supervised curation, the use of quality indices, text mining tools, biological ontologies and the semantic web[2]. Biomedical ontologies, in particular, provide a means for structuring this information such that it is computationally tractable and comparable across resources. One such effort is that of the Open Biomedical Ontologies (OBO) Foundry, which has the broad-based goal of "creating a suite of orthogonal interoperable reference ontologies in the biomedical domain"[5]. Examples of biomolecular and phenotype ontologies available via the OBO Foundry include: Cell Type Ontology (CO), Mammalian Phenotype (MP), Adult Mouse Anatomy (MA), Gene Ontology (GO), and National Center for Biotechnology Information (NCBI) taxonomy.

PhenoGO

Natural language processing (NLP) tools and semantic reasoning techniques can help to increase the availability of annotated resources and address the current gap in integrative translational knowledge necessary for the fields of systems biology and medicine[2]. The PhenoGO system utilizes an existing natural language processing (NLP) system, called BioMedLEE[6], and a knowledge-based phenotype organizer system (PhenOS) in conjunction with

* Corresponding authors

MeSH indexing and established biomedical ontologies, including the Unified Medical Language System (UMLS) and those comprising the OBO Foundry, to add contextual phenotypic information to existing associations between gene products and GO terms as specified in the GO Annotations (GOA)[7]. A feasibility assessment, focused on the extraction of phenotypic information from the scientific literature related to the mouse model, using an early version of BioMedLEE demonstrated 64.0% precision and 77.1% recall respectively[6]. A previous evaluation of the PhenoGO system, conducted in the context of the Mouse Genome Database, resulted in precision of 91% and recall of 92%, with respect to coding anatomical and cellular concepts and assigning the coded phenotypes to the correct GOA[7]. The PhenoGO database has recently been updated to include eleven of the species defined in the NCBI taxonomy, including *Homo sapiens*[8].

This manuscript expands the previous evaluations of BioMedLEE[6] and PhenoGO[7] to assess the feasibility of applying an NLP and semantic processing approach to the construction of a high-quality network comprised of integrated biomolecular and phenotypic data for *Homo sapiens*.

System Design

In the following sections, we describe a method for extracting, encoding and associating phenotypic and biomolecular concepts found in PubMed abstracts (Figure 1).

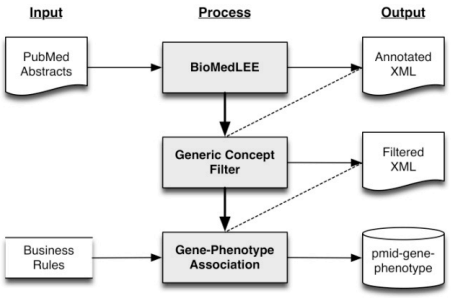

Figure 1. Overview of NLP and semantic methods for relating genes and phenotypes.

Natural Language Processing

The BioMedLEE system has been developed to automate the integration of phenotypic data contained within biomedical literature and genomic databases using NLP[6]. Specifically, it extracts and represents a comprehensive set of phenotypes using ontological codes and molecular mechanisms, as well as their specific relationships as expressed in the natural language of the literature. BioMedLEE utilizes six ontologies to encode the extracted concepts: Cell Type Ontology, Mammalian Phenotype, Mouse Anatomy, Gene Ontology (GO), NCBI Taxonomy, and the Unified Medical Language System (UMLS). In addition, BioMedLEE assigns a semantic type (e.g., body location, clinical finding, gene, cellular component) to each extracted concept. BioMedLEE was utilized to process a corpus of cancer-related (determined using a heuristic algorithm) PubMed abstracts, from 1990-2007, corresponding to *Homo sapiens* (MeSH: Humans).

Determining Gene-Phenotype Associations

After the parsed terms were extracted from the XML output of BioMedLEE, a filter was applied to remove overly generic terms (e.g., *cell*, *activity*, *structure*, *disease*, etc.). We employed similar methods to those used to construct the PhenoGO database[7] to associate biomolecular and clinical phenotypic concepts from the BioMedLEE XML output. That is, utilizing the phenotypic and genetic concepts identified by BioMedLEE, and the human Gene Ontology Annotations (GOA), the following rules were applied to determine gene-phenotype relationships:

1. Gene extracted by BioMedLEE from a given abstract must match a PubMed-gene pair found in GOA, based upon the abstract's PubMed ID (pmid) and GO code assigned by BioMedLEE.
2. Phenotype must be in the same BioMedLEE relationship (XML block) as the gene (Figure 2).
3. Phenotype must have a semantic tag (assigned by BioMedLEE) corresponding to a body function, body location, cellular location, clinical finding or problem.

```
<genefunc v = "regulation" code = "GO:0050789^regulation of biological process">
<process v = "proliferation"><arg v = "target"></arg>
<cell v = "progenitor cell" code = "UMLS:C0038250^stem cell"></cell></process>
<gene_gproduct v = "MGI:98958^Wnt5a"><arg v = "agent"></arg></gene_gproduct> </genefunc>
```

Figure 2. XML output of BioMedLEE for "Wnt5A regulates proliferation of progenitor cells."

Evaluation Methods

Precision Evaluation

We selected four random samples from the corpus of pmid-gene-phenotype associations in order to assess precision with respect to cell types, anatomy, diseases and genes/proteins. These associations were assessed based upon the following features:

1. *Semantic parsing and classification*: The BioMedLEE result was compared with the natural language phrase contents to assess the accuracy of the assigned semantic type (e.g. if the parsed term is "arm" from the phrase "broken arm", is it associated with the semantic type "bodyloc"?).
2. *Ontological annotation*: If (1) evaluated to a true positive (i.e., the phenotype was parsed and classified correctly by BioMedLEE), we assessed whether or not the ontology code assigned by BioMedLEE was correct for the parsed term (e.g. for the phrase "increased heart

rate" a code for "hypertension" would be considered to be a true positive).

3. *Gene-phenotype relationship*: If (1) evaluated to a true positive, we assessed whether or not the gene was correctly associated to the parsed phenotype by examining the context of the entire sentence from which the concepts were extracted.

Recall Evaluation

We evaluated four semantic types: *genes/proteins*, *cell/cell line*, *anatomy*, and *disease* (also includes clinical problem such as diagnosis or symptom). Initially, we chose four independent sets of 50 sentences at random (one for each semantic type). The relevant PubMed abstracts (i.e., from the human data set) were subsequently re-sampled to retrieve additional sentences until the subject matter expert (SME) evaluator noted 50 concepts per semantic type. During this process, no sentences were chosen more than once per data set, and no abstracts were sampled from more than once per data set. For those concepts that were part of a complex compositional phrase, the entire phrase was utilized for the evaluation. To assess how well BioMedLEE parsed the concept or compositional phrase that was identified by an SME, a score of one assigned to most relevant assessment category (see below), and a score of zero was assigned otherwise:

- *Correct – exact (true positive – TP)*: exact match;
- *Correct – different (TP)*: concept may have been assigned a semantic type that is not incorrect, but different from that being evaluated;
- *Partial (false positive – FP)*: some component of the concept or compositional phrase was missing or incorrect, but part of the concept or compositional phrase was parsed correctly;
- *Incorrect (FP)*: a parsed structure was output by BioMedLEE, but was completely incorrect;
- *None (false negative – FN)*: concept was not parsed by BioMedLEE, and a suitable UMLS concept exists
- *No code (true negative – TN)*: concept was not parsed by BioMedLEE, but no appropriate UMLS concept exists

For the coding evaluation we assessed the codes that were assigned to an SME-identified concept or compositional phrase by BioMedLEE. We evaluated either the most specific assigned code(s), or all component codes for those concepts and compositional phrases for which the entire phrase was not captured in a single code. Each code associated with a particular phrase was evaluated independently, but the overall score for each phrase totaled one (e.g., if a phrase had two UMLS codes assigned to it, each assessment would have a value of 0.5). The following assessment categories were utilized:

- *Correct – exact (TP)*: exact match;
- *Correct – partial (FP):* at least one assigned code was only partially correct with respect to the complete concept term or compositional phrase;
- *Partial (FP):* at least one assigned code was only partially incorrect with respect to the complete concept term or compositional phrase;
- *Incorrect (FP)*: at least one assigned code was completely incorrect with respect to the concept term or compositional phrase;
- *None (FN):* no code was assigned to the complete concept or compositional phrase, but a correct UMLS code exists;
- *No code (TN):* no code was assigned to the complete concept or compositional phrase, and no correct UMLS code exists

Results

BioMedLEE's grammar contains 810 rules, and its lexicon contains 830,058 lexical entries and 502,965 distinct targets. We utilized BioMedLEE to process a corpus of 11,407 PubMed abstracts. The XML output of BioMedLEE was comprised of 759,026 unique textual terms coded in 451,547 distinct concepts.

Determining Gene-Phenotype Associations

After applying a script to remove all overly generic concepts from the BioMedLEE XML output, the resulting data set was comprised of over 200 million annotated phenotypes (non-distinct). In addition, over 100,000 pmid-gene-phenotype associations were asserted using the structured relationships in the BioMedLEE output and human Gene Ontology Annotations (GOA). This data set has now been incorporated into the PhenoGO database, which can be accessed at: http://www.phenogo.org/.

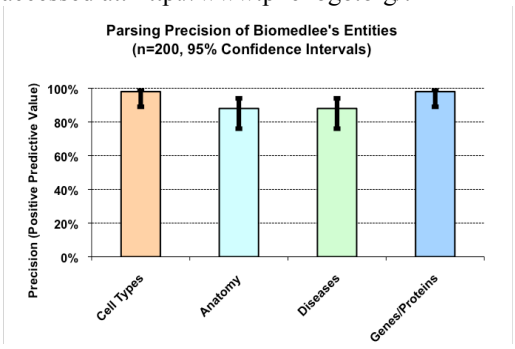

Figure 3. Summary of parsing precision results.

Evaluation

For the precision evaluation, we selected four random samples comprised of 50 pmid-gene-phenotype associations per semantic class (cellular anatomy, supracellular anatomy, finding/morphology/disease, gene/protein). The precisions with respect to parsing

by BioMedLEE across these semantic types averaged 93±0.07%, and are summarized in Error! Reference source not found.. For the coding evaluation, the precisions averaged 77% (maximum: 94% [gene/protein]; minimum: 50% [supracellular anatomy]) across semantic types. Recall averaged 85.91% and 70.65% for parsing and coding, respectively. The recall results are summarized in Table 1 and Figure 4. The overall coding precision for the pmid-gene-phenotype relationships was 75%.

Table 1. Examples from recall (coding) evaluation

Metric	Input Sentence (focus **concept** and phrase)	BioMedLEE Code
Correct (E)	PTP2C is widely expressed in … heart, **brain**, and skeletal muscle.	MA:0000168 [brain] UMLS:C0006104 [brain]
Correct (P)	Inhibition of CXCR4-dependent **HIV-1 infection** ...	UMLS:C0021311 [infection] GeneID:7852 [CXCR4] (missing code for HIV-1)
Partial	… suggest … interfering with the **CD28** costimulatory pathway may …	GeneID:317783 [CELIAC3] GeneID:940 [CD28] (CD28 not alias of CELIAC3)
Incorrect	… mRNAs induced in **BL cells** have been cloned ….	UMLS:C0009013 [clone cells] (missing UMLS:C0006413 [Burkitt Lymphoma])
None	… identified … as the gene responsible for **macular corneal dystrophy**.	Missed UMLS:C0024439 [Macular corneal dystrophy]
No code	… region within the candidate locus for **lethal neonatal metabolic syndrome** …	No exact UMLS code

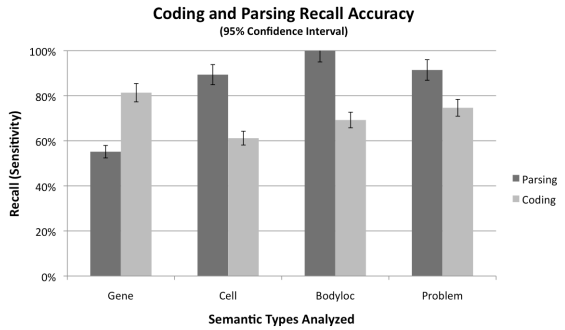

Figure 4. Summary of recall evaluation results.

Discussion

The domains of systems biology and medicine inherently involve the integration of multiple diverse data sets, including the vast amounts of knowledge buried within the biomedical literature. However, the burden of manual annotation and the ability to perform such integration with the precision and detail necessary for hypothesis generation are some of the primary limitations to these approaches. To-date, there have been several techniques utilized to extract and correlate genes and phenotypes based upon the information contained in this literature, including supervised learning[9] and text mining[10]. In addition, Hunter et al.[11] and Sam et al.[12] have demonstrated the use of approaches that integrate NLP techniques with biomedical ontologies to predict and discover protein-protein interaction networks. *However, to the best of our knowledge novel natural language processing (NLP) engines, such as BioMedLEE, have not previously been applied to automatically extract, encode with standard ontologies, and associate biomolecular and phenotypic information for the generation of integrated reference knowledge sets for use in the systems biology and medicine domains.*

Payne, Embi and Sen[13] define a translational research informatics framework for the design and execution of informatics-enabled studies that aim to integrate, analyze and disseminate large-scale, heterogeneous biomedical datasets, such as those prevalent in systems biology. The relatively high precision and recall evaluation metrics associated with the use of BioMedLEE for annotating and encoding biomedical literature, and associating genes/proteins and phenotypes indicate that this methodology could be utilized to generate integrated data sets that are sufficiently precise and timely for generating testable hypotheses within such a translational research framework. Though others have predicted systems medicine properties (e.g., protein-protein interactions) from the literature by mining co-occurrences[14], such methodologies are not as precise as NLP-based approaches[15]. However, the BioMedLEE-derived ontology-anchored networks allow for the generation of relationships between data types that are more analogous to the mining of semi-structured or structured datasets, such as the work reported by Hansen, et al[16].

A prototypical biological problem that can serve as an exemplary application for an integrated biomolecular and phenotype network, such as that generated using the methods described in this manuscript, is that of adaptive therapy planning for chronic lymphocytic leukemia (CLL). Recent publications[17] have demonstrated a paucity of empirically validated biomolecular markers that correlate with treatment outcome in CLL. Further, the same literature demonstrates a lack of systematic approaches to the design and execution of studies to elucidate such linkages. By leveraging the network generated in this study, it would be possible to systematically evaluate novel genotype-phenotype relationships in CLL based upon comprehensive literature-based knowledge sources in order to design such studies and ultimately generate evidence capable of supporting adaptive therapy planning.

Though our results are promising, our study did have several limitations, including: (1) the developer of the gene-phenotype association algorithm also conducted the performance evaluations; (2) the resulting integrated gene-phenotype data set is only as accurate and timely as the ontologies utilized at the foundation

of the NLP engine; and (3) when calculating the accuracy metrics for BioMedLEE, all partially correct assessments were considered to be false positives, thus deflating the reported values. Our future work involves pipelining the methodology described in the manuscript with existing knowledge and hypothesis discovery tools (e.g., PGSchema, PhenoGO, and protein interaction networks) to enable scalable and comprehensive integration of the annotated biomedical literature, independent research databases and existing genomic knowledge sets. We are currently completing the processing of PubMed abstracts from 1865-2009, and the update of the public PhenoGO database.

Conclusions

Though the discovery of novel linkages among genes, biologic functions and human diseases is foundational for the domains of systems biology and medicine, one of the primary limitations to such an approach is the availability of methodologies for adequately integrating the inherently heterogeneous datasets. We have developed and evaluated a natural language and semantic processing-based approach that utilizes the BioMedLEE NLP engine for extracting and encoding biomolecular and phenotypic concepts from existing published literature and biomedical ontologies, and a subset of the PhenoGO contextual assignment algorithms to determine relationships among these concepts. The relatively high precisions and recalls resulting from the subsequent evaluation in the domain of *Homo sapiens* indicate that our methodology has promise for the generation of integrated biomolecular and phenotypic knowledge sets that are precise enough to discover testable systems biology hypotheses.

Acknowledgements. This work was supported in part by the NIH/NLM/NCI National Center for Multiscale Analyses of Genomic and Cellular Networks (MAGNET, 1U54CA121852), NIH/NCRR CTSA (1U54 RR023560-01A1), R0LM17659 and R01LM008635 from the NLM, The Cancer Research Foundation and UCCRC.

References

1. Ahn, A.C., et al., *The clinical applications of a systems approach.* PLoS Med, 2006. **3**(7): p. e209.
2. Auffray, C., Z. Chen, and L. Hood, *Systems medicine: the future of medical genomics and healthcare.* Genome Med, 2009. **1**(1): p. 2.
3. Price, N.D., et al., *Systems Biology and the Emergence of Systems Medicine, in Genomic and Personalized Medicine*, H.F. Willard and G.S. Ginsburg, Editors. 2009, Elsevier. p. 74-86.
4. Joyce, A.R. and B.O. Palsson, *The model organism as a system: integrating 'omics' data sets.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 198-210.
5. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
6. Chen, L. and C. Friedman, *Extracting phenotypic information from the literature via natural language processing.* Stud Health Technol Inform, 2004. **107**(Pt 2): p. 758-62.
7. Lussier, Y., et al., *PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing.* Pac Symp Biocomput, 2006: p. 64-75.
8. Sam, L.T., et al., *PhenoGO: an integrated resource for the multiscale mining of clinical and biological data.* BMC Bioinformatics, 2009. **10 Suppl 2**: p. S8.
9. Korbel, J.O., et al., *Systematic association of genes to phenotypes by genome and literature mining.* PLoS Biol, 2005. **3**(5): p. e134.
10. Leitner, F., R. Hoffmann, and A. Valencia, *Biological Knowledge Extraction: A Case Study of iHOP and Other Language Processing Systems*, in *Bioinformatics for Systems Biology*. 2009, Humana Press. p. 413-433.
11. Hunter, L., et al., *OpenDMAP: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression.* BMC Bioinformatics, 2008. **9**: p. 78.
12. Sam, L., et al., *Discovery of protein interaction networks shared by diseases.* Pac Symp Biocomput, 2007: p. 76-87.
13. Payne, P.R., P.J. Embi, and C.K. Sen, *Translational Informatics: Enabling High Throughput Research Paradigms.* Physiol Genomics, 2009.
14. von Mering, C., et al., *STRING: known and predicted protein-protein associations, integrated and transferred across organisms.* Nucleic Acids Res, 2005. **33**(Database issue): p. D433-7.
15. Jensen, L.J., J. Saric, and P. Bork, *Literature mining for the biologist: from information retrieval to biological discovery.* Nat Rev Genet, 2006. **7**(2): p. 119-29.
16. Hansen, N.T., S. Brunak, and R.B. Altman, *Generating genome-scale candidate gene lists for pharmacogenomics.* Clin Pharmacol Ther, 2009. **86**(2): p. 183-9.
17. Grever, M.R., et al., *Comprehensive assessment of genetic and molecular features predicting outcome in patients with chronic lymphocytic leukemia: results from the US Intergroup Phase III Trial E2997.* J Clin Oncol, 2007. **25**(7): p. 799.