PhenoGO: Assigning Phenotypic Context to Gene Ontology Annotations with Natural Language Processing

Yves Lussier, Tara Borlawsky, Daniel Rappaport, Yang Liu, and Carol Friedman

Pacific Symposium on Biocomputing 11:64-75(2006)

PHENOGO: ASSIGNING PHENOTYPIC CONTEXT TO GENE ONTOLOGY ANNOTATIONS WITH NATURAL LANGUAGE PROCESSING

YVES LUSSIER*,[1,2], TARA BORLAWSKY[§,1], DANIEL RAPPAPORT[§,1], YANG LIU[1], CAROL FRIEDMAN*,[1]

1- Department of Biomedical Informatics, Columbia Center for Systems Biology, 2- Department of Medicine ; Columbia University, New York, NY 10032

Natural language processing (NLP) is a high throughput technology because it can process vast quantities of text within a reasonable time period. It has the potential to substantially facilitate biomedical research by extracting, linking, and organizing massive amounts of information that occur in biomedical journal articles as well as in textual fields of biological databases. Until recently, much of the work in biological NLP and text mining has revolved around recognizing the occurrence of biomolecular entities in articles, and in extracting particular relationships among the entities. Now, researchers have recognized a need to link the extracted information to ontologies or knowledge bases, which is a more difficult task. One such knowledge base is Gene Ontology annotations (GOA), which significantly increases semantic computations over the function, cellular components and processes of genes. For multicellular organisms, these annotations can be refined with phenotypic context, such as the cell type, tissue, and organ because establishing phenotypic contexts in which a gene is expressed is a crucial step for understanding the development and the molecular underpinning of the pathophysiology of diseases. In this paper, we propose a system, PhenoGO, which automatically augments annotations in GOA with additional context. PhenoGO utilizes an existing NLP system, called BioMedLEE, an existing knowledge-based phenotype organizer system (PhenOS) in conjunction with MeSH indexing and established biomedical ontologies. More specifically, PhenoGO adds phenotypic contextual information to existing associations between gene products and GO terms as specified in GOA. The system also maps the context to identifiers that are associated with different biomedical ontologies, including the UMLS, Cell Ontology, Mouse Anatomy, NCBI taxonomy, GO, and Mammalian Phenotype Ontology. In addition, PhenoGO was evaluated for coding of anatomical and cellular information and assigning the coded phenotypes to the correct GOA; results obtained show that PhenoGO has a precision of 91% and recall of 92%, demonstrating that the PhenoGO NLP system can accurately encode a large number of anatomical and cellular ontologies to GO annotations. The PhenoGO Database may be accessed at the following URL: http://www.phenoGO.org

1 Introduction, Related Work and Background

In recent years, there has been a growing interest in automatic methods that annotate biomedical journals. Several methods that use Medline Abstracts in order to annotate genes to *Gene Ontology* (**GO**) terms[1] have been proposed and have yielded up to 10% to 20% recall and 61-99% precision[2,3]. However, to our knowledge, no method is available to automatically process text in order to map contextual pheno-

* Corresponding authors that have contributed equally to the work
§ These authors have contributed equally to the work

types to Gene Ontology Annotations. Establishing ***phenotypic contexts*** in which a gene is expressed is a crucial step for understanding the molecular underpinning of the pathophysiology of diseases. Since complete genomes of multicellular organisms are increasingly annotated in GO, phenotypic context annotations of these genes could serve as a basis for large scale comparative analyses of gene phenotype interactions (phenomics). For example, the specific cell type(s) in which a gene is expressed are very useful to establish the functional molecular networks of differentiated cells (e.g. "CD4+ T Lymphocytes", but not "CD8+", are responsible for murine "interleukin-2-deficient" colitis resembling ulcerative colitis in humans[4] [MGI: 96548 Il2 interleukin 2, GO:0005134 interleukin-2 receptor binding]). More particularly, bioinformatics methods in systems biology are based on the analysis of datasets relating multiple scales of biology together. In this paper we describe an automated system, PhenoGO, which combines NLP and knowledge-based methods to infer the anatomical and cellular context of existing associations between gene products and GO terms as specified in *GOA*[5]. GOA databases comprise gene-GO associations according to a reference (usually a *Pubmed identification*: **PMID**) and the curation process in GOA has a precision of 91%-100% and a recall of 72%[2]. In addition, we performed an evaluation of PhenoGO over the *Mouse Genome Database* (**MGI**) annotations of GO, and report on the results, which show high precision and recall.

1.1 Related Work

A key step for understanding the phenomes is to provide phenotype and genotype information. While GO provides some phenotypic information, other "orthogonal" ontologies have been developed such as *Cell Ontology* (**CO**)[6], the Adult Mouse Anatomy (**MA**)[7] and the *Mammalian Phenotype Ontolog*y (**MP**)[8]. In addition, more traditional ontologies and terminologies can be filtered to yield other complementary phenotypes, such as the *Unified Medical Language System* (**UMLS**)[9] and the *NCBI Taxonomy*[10].

Natural Language Processing (NLP). Since 1998 there has been an increasing amount of language processing research that involves extraction and mining of biomolecular information in journal articles. Some systems recognize and/or identify biomolecular entities, some detect relations among biomolecular entities, and some discover new knowledge by piecing together information from heterogeneous resources[11]. Krauthammer and Nenadic provide a review of entity recognition systems[12], Cohen and Hersh, and Hirschman and colleagues each provide an overview of relation extraction and text mining systems[13,14]. Until recently biological language processing systems generally extracted terms and

relations, but did not map them to concepts in an established ontology. In the biological domain, it has recently been recognized that to achieve interoperability and improved comprehension, it is critical for text processing systems to map extracted information to ontological concepts. A number of researchers have developed systems mapping genes to GO codes[2,3,14,15,16]. Work in the medical domain involving the mapping of text to UMLS concepts has also been explored. For example, Aronson developed MetaMap, which consists of a mixture of statistics and linguistic methods[17], Nadkarni and colleagues[18] use a string matching approach, and Friedman and colleagues[19]use an *NLP system* called MedLEE. MedLEE differs from the other NLP coding systems in that the codes are shown with modifier relations so that concepts may be associated with temporal, negation, uncertainty, degree, and descriptive information, which affect the underlying meaning and are critical for accurate retrieval of subsequent medical applications. The PhenoGO system discussed in this paper utilizes an adaptation of the MedLEE engine, as discussed in the methods. While bioinformatics techniques have been developed to infer phenotypes from biological databases[20] (e.g. microarray experiments), to our knowledge, none have used NLP techniques over the literature to extract relations that associate anatomical or cellular phenotypes to genes and GO terms together.

Knowledge Management, MeSH Indexing and NLP. The *Medical Subject Headings* terminology (**MeSH**) is the National Library of Medicine controlled vocabulary thesaurus[21] and covers all biomedical concepts classes, including phenotypes. To index the main concepts in the Medline paper, MeSH headings are created manually by experts who read the entire Medline paper and then assign MeSH headers to relevant PMIDs. Of relevance to the proposed methods, MeSH terms have been mapped to other terminologies and organized in a semantic network in the UMLS. Recently, phenomic systems have been developed by Bodenreider and Lussier to relate phenotypes and genes relying exclusively on computational terminology methods[22]. The PhenoGO system reuses the knowledge of MeSH indexes and GOA to infer the phenotypic context of genes mapped to GO terms.

1.2 Background

BioMedLEE NLP System. The NLP component of PhenoGO utilizes an existing NLP system, called BioMedLEE, which is under development by the Friedman language processing group. BioMedLEE extracts and encodes genotype-phenotype relations from information in text. An early version is described in Chen and colleagues[23], but differs substantially from the current one in that it extracted phenotypic information only, and did not map textual terms to codes. The BioMedLEE system is based on an adaptation of the MedLEE system[19], which

```
<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 1. XML output of BioMedLEE for "*Wnt5A regulates proliferation of progenitor cells*."

extracts and encodes clinical information in patient reports. An important feature of MedLEE/BioMedLEE for PhenoGO is the flexible infrastructure for **mapping textual terms to codes** and is described in more detail by Friedman elsewhere[19].

A detailed description of the BioMedLEE system is being submitted as a separate publication. In this paper, we summarize the components critical to PhenoGO: 1) the first one, prepares the articles for processing by extracting relevant textual sections, and by handling parenthetical expressions, 2) **the entity tagging component** performs semantic tagging of certain entities, such as biomolecular entities 3) the next component identifies section and sentence boundaries, and performs lexical lookup using a lexicon that specifies the semantic and syntactic categories of terms that were not previously tagged. Many of the lexical entries were automatically generated using GO, MA, MP, cell ontology, and the UMLS, 4) the parser structures the sentence according to grammar rules which specify the relations among the concepts. A parse of the complete sentence is attempted first, and then, if unsuccessful, large segments are attempted in an effort to maximize the capture of relations, 5) the last stage performs encoding using a coding table and the structured output from the previous parsing stage to find the most specific codes, and to generate XML output. Codes in the table are represented as triples consisting of a prefix that identifies the specific ontology, an identifier within the ontology, and the name of the concept (e.g. GO:00507899^regulation of biological process). ***Figure 1*** illustrates a simplified form of the output that was generated by processing *Wnt5A regulates proliferation of progenitor cells.* The molecular function **genefunc** with a value attribute **regulation** is the main output structure; it has a code **GO:0050789** followed by the name of the concept. The function is associated with two arguments where one has a cellular modifier. One argument is a **process,** which has the value **proliferation**, but does not have a code. **Process** is the target of **regulation,** and thus, it has an **arg** tag with a value **target**. The **cell** tag with the value **progenitor cell** and **code UMLS:C0038250** modifies **process**.The other argument, the agent of **regulation**, is a gene whose code is **MGI:98958.** What is significant about the above output is that the MGI-GO-cell triplet is contained within one structure **genefunc**, signifying that BioMedLEE found a relation between the 3. In other relevant sentences, BioMedLEE may find a relation between the phenotype and only one of the MGI-GO pair, or may just find the context without any relation to MGI or GO.

Phenotype Organizer System (***PhenOS***) *Knowledge Management and Computational Terminology System*. As the focus is on NLP, and the knowledge management, and computational terminology methods have been published earlier in related papers, we provided a summary of PhenOS. *PhenOS* is a system under development by the Lussier research group with purpose of bridging the gap between heterogeneous biomedical terminologies. The system produces a directed acyclic graph from the UMLS, provides lexico-semantic and model theoretic methods that automatically map an ontology to another one independently of the UMLS and organizes and structures phenotypes across heterogeneous datasets [22,24,25]. Specific methods of PhenOS used in the current study were the integration of phenomic knowledge structures via structured terminology.[25,24]

2 Methods

The method for assigning phenotypic context to GOA was implemented as a system called PhenoGO. An overview of the overall components is illustrated in ***Figure 2.*** First, Medline abstracts and their gene-GO annotations are identified from GOA, and then obtained. Two distinct and independent processes extract phenotypic context from the abstracts. The design is such that PhenoGO can function with either or both of the processes. ***Component 1a*** consists of the **BioMedLEE NLP** system, which processes the title and abstract and generates structured output, which in this study identifies genes, GO codes, coded phenotypes (UMLS, CO, MA, MP), along with gene-GO-context relations. ***Process 1b*** simply obtains relevant phenotypic MeSH headings from the Medline abstracts. ***Component 2***, the PhenOS knowledge management system completes the contextual assignment of contextual phenotypes to GOA gene-GO terms pairs related to the same PMIDs.

2.1 Phenotypic Context Determination and Encoding Components

Process 1a- Determining Context using NLP (BioMedLEE). Abstracts are parsed by the NLP system BioMedLEE (which was not adapted for PhenoGO) according to 2 different abstract sections: (i) titles and (ii) body. BioMedLEE can extract about 50 distinct semantic types from biological text as well as generate codes from multiple ontologies, but this study focuses by design on the following 6 coding systems, 4 types of entities, and their associations: 1) MGI:genes, 2) GO: terms, 3) Cells coded in CO or UMLS and 4) anatomies above the cellular level coded in MA, MP, and UMLS. Two training sets of 50 PMIDs were selected, parsed by BioMedLEE and analyzed thoroughly. Consequently, ten UMLS encodings observed ambiguous in

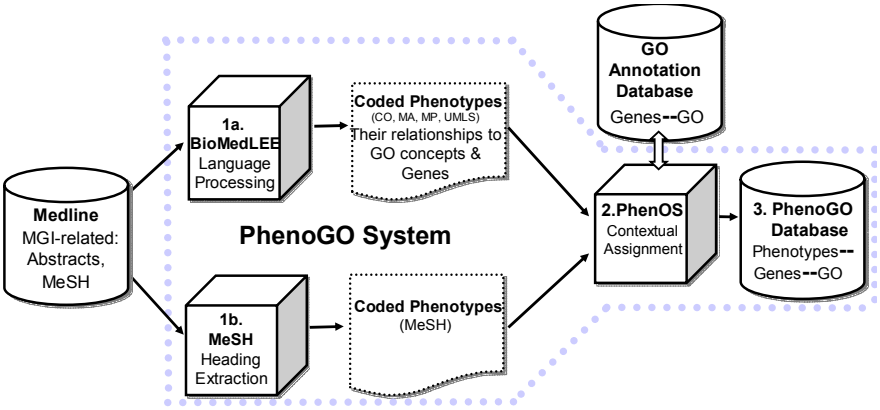

Figure 2: ***Diagram of the PhenoGO System showing Software and Database Components Involved in the Assignment of Phenotypic Context to Gene Ontology Annotations***.

the training set or meaningless for our aims are filtered out (e.g. *back*, *helix, tissue*).

Process 1b- Determining Context using Curated Knowledge (MeSH and PhenOS). MeSH terms considered useless in PhenoGO were filtered out during training (e.g. "cell", etc.), while the MeSH headings subsumed by the following *concepts of the UMLS semantic networks* (**TUI**) were selected: Anatomical Structure, Embryonic Structures, Body Part, Organ, Tissue, Body Substance, and Systems.

2.2 Phenotype Organizer System (PhenOS) Contextual Assignment

*The PhenOS system (**Component 2**, **Figure 2**)* performs the contextual assignment, and uses a different process depending on whether the coded phenotype came from processes ***1a*** *or* ***1b*** (***Figure 2***). First, component 2 obtains the gene-GO pair that was associated with the specific article in the GOA database. If the coded phenotype originated as a MeSH header that was selected by PhenOS, it is assigned as a contextual phenotype augmenting the gene-GO pair. *Our hypothesis is that when the MeSH header lists a contextual phenotype, it is relevant not only to the article but also to the gene-GO pair associated with the Medline article in GOA*.

If the coded phenotype was generated by the NLP system, a more complex procedure is followed. The assumption in this case is that context may be mentioned incidentally and picked up by the NLP system, but may not be related to the gene-GO pair. In order to achieve high precision, *we hypothesize that if there is a relation between the codes in a gene-GO-phenotype triple and the gene and GO codes match the corresponding ones in GOA, then the phenotype is highly likely to be related to the matched pair, and thus is likely to be the correct context.* The other extreme is that if the context does not match any gene or GO code in the pair asso-

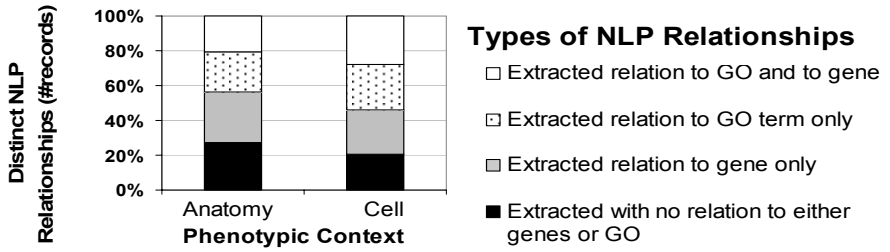

Figure 3. Descriptive Statistics of the NLP Component: Coded Phenotypes according to the Method 1a. The analysis was performed on data captured downstream of the the BioMedLEE NLP component (*1a, Fig. 2*) and upstream (before) the PhenOS component (*2, Fig. 2*) that uses the types of NLP relations provided by the NLP to match the cellular and anatomical phenotypes with specific GO annotations.

ciated with the article, it is more likely that the phenotype may not be related to the pair. Thus, to help analyze the type of relations based on NLP processing that affect performance of PhenoGO, we record the types of relations and matches (i.e. GO-gene-phenotype, gene-phenotype, phenotype-GO, no match). If a GO code is extracted that does not match the database GO code, it may also be because the NLP system obtained a more specific code than the one in GOA. PhenOS is then used to determine whether an ancestor of the extracted GO code matches the code in GOA; in that case this GO code is considered matched. For example, codes are considered to be matching if the extracted code corresponds to 'negative regulation of biological process' and the curated code to 'regulation of biological process'.

2.3 Evaluation

Medline Abstracts, Sampling and Gold Standard. We have focused the experiment on 3,705 PMIDs of the GOA of the Mouse Genome Database (MGI), which contains 2,327 distinct GO terms, 4,269 distinct MGI genes and 12,220 GO-gene pairs. Random samples of PMIDs were selected for creating the *Gold Standard* (**GS**) as described below. For evaluating recall, a sample of 50 PMIDs was randomly selected from the 3,705 MGI PMIDs. In this sample, each occurrence of anatomical and cellular phenotypes as well as their relevance to their respective MGI GO annotations were manually curated by one curator and confirmed by another. Samples for calculating precision were randomly selected from the set of coded results applicable to our study for evaluating, respectively, the cellular and the anatomical phenotypic contexts assigned by PhenoGO. The precision GS comprised a total of 575 curated phenotypes associated with genes and GO terms.

Accuracy Measures. We measured the precision and recall of the assignment of coded anatomies and cell types based on the individual and combined NLP and the

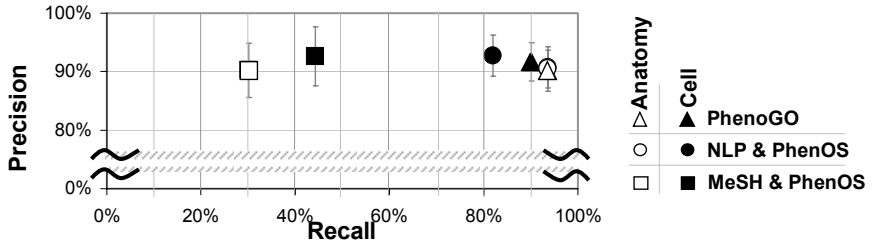

Figure 4. Precision and recall of the assignment of coded phenotypes to gene-GO term pairs by PhenoGO and by its NLP and MeSH components separately (*refer to Fig.2, components 1a & 1b*). Confidence intervals are based on the binomial distribution.

knowledge components of PhenoGO. Recall was calculated as the ratio of the number of distinct [GO-MGI gene-phenotype] triplets that were identified by the mapping method (***Figure 2***) that matched those in the GS, divided by the total number of triplets in the GS, (TP)/(TP+FN). Precision was measured as the same numerator as recall divided by the total number of triplets predicted by the mapping method (NLP&PhenOS, MeSH&PhenOS, PhenoGO), TP/(TP+FP). Thus, in this evaluation, in order to count a true positive score, the PhenoGO system must accurately encode a phenotype and also relate it with its gene-GO term pair.

3 Results

Overall, PhenoGO provided phenotypic context for 96% of the 3,705 PubMedIDs. The NLP coded a larger percentage of phenotypes than the knowledge-based method, and the joint NLP-KB method was significantly better for the cellular annotations demonstrating that the combined NLP-KB method yielded 50% more cellular annotations (***Figure 4***). The NLP process provided several times more annotations than the knowledge-based one, though the difference was more important in anatomies above the cell level than for cellular anatomies (***Figure 4***). As shown in ***Figure 3,*** the majority of the NLP annotations were related to the genes only (e.g. context and gene in same NLP structure with no GO), followed by relationships to GO only; a smaller number were not related to GO or genes (i.e. context not in same structure as GO or gene), and finally an even smaller number were related to both genes and GO terms (example of this relation is shown in ***Figure 1***). By design, the MeSH headings have no mapping to gene or GO before they are received by the PhenOS component (***Figure 2***), thus there is only one type of contextual assignment method for MeSH terms because the phenotypic context is always assumed to be related to the gene and GO pair as specified in GOA. In order to provide a summary of the scales of anatomies mapped by the system, the following are the counts

of distinct types of concepts mapped according to their ontologies: MA:345, MP: 305, CO:148, MeSH: 460 (Embryo:32, Organs and body parts:240, Tissues:42, Systems: 22, Cells:124), UMLS: 1,259 (Embryo:97,Organs and body parts: 786, Tissues: 100, Systems:37, Cells:239). The PhenoGO precision and recall for cells and anatomies combined are 91% and 92% respectively (the accuracy measures the combined coding to the ontology and the assignment of the correct gene-GO pair with the phenotype). Details of the results are in ***Figure 4*** showing that the "cell" recall of the PhenoGO system is substantially better than that of the NLP, while the precision remains unchanged.

4 Discussion

PhenoGO performed well in accuracy scores. Precision, which is the most important metric, was over 90% for both the MeSH and the NLP methods. Of note, the NLP component, which had access to only the title and abstract, was not significantly different from the MeSH component which is based on manual curation of the complete paper by experts who focus on main concepts of the article. In addition, the NLP system did not have the expert knowledge of curators, and therefore could not discern whether or not the context was incidentally mentioned or significant to the paper. It is quite interesting that only 1 error in precision was caused by an incorrect association of an anatomical location with a gene-GO pair. Thus, based on our evaluation, use of NLP to augment the GO database with phenotypic information is very promising. Moreover, recall of the NLP component was much higher than that of the MeSH component, possibly because curators do not focus on indexing context. It is also striking that the results for the NLP component are associated with performance in coding and in assigning these codes to the right GOA, which are much more difficult tasks than extraction alone, and thus, the NLP performance in extraction precision is likely to be higher. BioMedLEE was not trained for this particular task, and it is likely that further revisions will enable it to perform better.

An analysis of the BioMedLEE errors was performed, and the most frequent types of errors are shown in ***Table 1***. Not surprisingly, word sense ambiguity was the most frequent cause of error in precision. Often a gene name was also a phenotypic entity, and the incorrect sense was chosen. For example, *Notch* is a gene and also an anatomical part according to the UMLS. Another cause of error in precision was due to use of the synonym lists associated with the ontologies because they often list incomplete terms as synonyms of more complete terms, causing coding and word errors. For example *band* is listed as a synonym of *band form neutrophil* in cell ontology, but it occurred as a different sense in an article. It appears that on

Table 1 - Most frequent types of errors in precision and recall are shown along with examples.

	Reason	Example
Precision	Ambiguity	*Notch* was interpreted as **anatomy** but is **gene** in *defects of notch pathway*
	Ontology	*Band* incorrectly mapped to *band form neutrophil* when it occurred in *50 k-Da bands* since *band* is listed as a synonym of *band form neutrophil* in CO
	Term recognition	*Finger* was interpreted as **anatomy** instead of part of term *zinc finger protein*
	Incorrect relation	*Skeletal defect* was associated with gene *Lfng* in article instead of *delta-like 3*
Recal	Mapping to ontology	*Lymphoid & adipose cell* not mapped to *lymphoid tissue* & *adipocyte*
	Lexicon	*Epithelia*, *epididymus* not defined in lexicon
	Term recognition	*Mast cell* not captured as **anatomy** since it is part of term *mast cell tryptase*
	Ontology	*Precursor* was incorrectly listed as synonym of *blood precursor* in UMLS

tologies make assumptions, which may be applicable within a particular domain, but not across broader domains. Other causes of precision errors occurred when a biomolecular term was not recognized, but part of the term was. Thus, *finger* was assigned as context when it occurred in the phrase *zinc finger protein*. The most frequent causes of error in recall were due to failures in coding and to terms that were not recognized because they were missing from the lexicon. Coding errors were typically caused when a synonym of a term was not listed in any of the ontologies. Thus, *lymphoid* (where the word *tissue* was omitted) could not be associated with a code. Lexical omissions typically occurred when a rare variant form of a term occurred in an article.

PhenOS Contextual Assignment Evaluation . We have validated the hypothesis that the curated phenotypes found in a MeSH terms pertained to the whole article, thus to every GO annotation of that article. Indeed, the evaluation measures both the validity of the phenotypic encoding and that of its assigned contextual relationship to a gene-GO term pair. Additionally, the precision of PhenoGO assignments based on NLP are 91% and 93% for the anatomies and the cell types respectively (**Fig. 4)**. It is likely that some association relations confer higher quality to the phenotypic context extracted from the abstract by the NLP. For example, when both gene and GO term associations are found related to a phenotype, we predict that the average precision would be higher than that of no associations at all.

There were *some limitations to this study*. Two students who had a background in biology were used to sample the abstracts and results, and create the gold standard. Some of the evaluation required reading the entire documents or the abstracts, both of which are time-consuming tasks, and therefore a limited number of samples were used in the evaluation. An additional limitation is that the study was performed using articles selected from GOA also focused on the mouse. Results of the NLP component may be different for other organisms.

Significance of the Integrated PhenoGO System. An integrated system that combines existing knowledge with NLP coded information has many advantages. First, through GOA, knowledge of the precise gene-GO pair and model organism

associated with articles that were annotated is known, providing a fairly accurate way to resolve the identity of an ambiguous gene for contextual assignment. Although the name of an ambiguous gene is associated with more than one identifier, if one of the identifiers matches the one found in the article, it is highly likely to be the correct one. Another very significant advantage is that PhenoGO can be scaled up very quickly and can be deployed to automatically create a database of substantial size. It is scalable in several dimensions. Using the NLP component, it is possible to process huge volumes of journal articles as well as textual database fields where there are no MeSH codes. However, the NLP system must be trained for phenotypic and genotypic information associated with the organism that corresponds to the text. Using the MeSH component it is possible to rapidly determine contextual information across organisms, and thus to augment GOA with context.

Future Work. We are revising PhenoGO to increase performance by improving the BioMedLEE NLP system as well as the PhenOS contextual assignment method and we are mapping to additional types of phenotypes which are beyond this study, such as morphologies and diseases. We are currently using PhenoGO to process every PMID associated with GOA for Mus musculus and Homo sapiens, and we will perform additional evaluations of the results as we extend to every species in GOA. Our ultimate objective is to provide an accurate and regularly update open source PhenoGO database of phenotypic and contextual annotations for the biological and informatics communities.

5 Conclusion

We developed and evaluated an automatic NLP system, PhenoGO, for augmenting gene product and GO associations using MeSH headings, the UMLS hierarchy, and an existing NLP system that maps terms in text to identifiers in multiple biological ontologies. Results demonstrated that the PhenoGO NLP system encodes anatomical and cellular ontologies to GO annotations with high recall and precision. The system is scalable in a number of dimensions: (i) it enables high throughput, (ii) it encodes in multiple ontologies and terminologies (GO, CO, MA, MP, UMLS, MeSH), (iii) it provides different modifiers for the encoded phenotypes, (iv) it uses external knowledge bases when they exist to increase accuracy (e.g. MeSH), (v) it can provide other types of context, such as diseases. In addition, the system can also map to other species and encode textual data from biological databases (results not shown). Of significance, the hypothesis that MeSH anatomical knowledge generally applies to every GO annotation assigned to the same PMID has been confirmed, which is likely to allow for rapid scalability across GOA of every species. In sum-

mary, the PhenoGO database is expected, upon completion, to provide valuable high throughput resources for biological in silico experiments as one can investigate in high throughput the differential expression of gene and their GO annotations across different cell types, organs, systems, etc. For example, this tool could be used to investigate the role of genes in the cellular differentiation of complex organisms using GOA with their phenotypic contexts.

Acknowledgments

This study was supported in part by NIH/NLM grants 1K22 LM008308-01(YL) and R01 LM007659(CF). The authors thank Lyudmila Shagina and Jianrong Li for their respective contribution in the development of BioMedLEE and PhenOS.

References

1.Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
2.Camon EB, Barell DG, Dimmer EC,et al. *BMC Bioinform* **6** S1(2005).
3.Raychaudhuri S,Chang J,Sutphin P,Altman RB.*Genome Res.* **12**(1),203-14(2002).
4. Simpson SJ, et al. *Eur J Immunol.* Sep;**25**(9):2618-25 (1995).
5.Camon E, Magrane M, Barrell D, et al. *Nucleic Acids Res.* **1**;32:D262-6 (2004).
6.Bard J, Rhee SY, Ashburner M. *Genome Biol.* **6**(2):R21.(2005).
7.Hayamizu TF, Mangan M, Corradi JP, et al. *Genome Biol.* **6**(3):R29 (2005).
8.Smith CL, Goldsmith CA, Eppig JT. *Genome Biol.* **6**(1):R7 (2005).
9.Lindberg C. *J Am Med Rec Assoc*.**61**(5):40-2 (1990).
10.Wheeler DL, Chappey C, Lash AE, et al. *Nucleic Acids Res* **1**;28(1):10-4(2000).
11.Tiffin N, Kelso JF, Powell AR, et al. *Nucleic Acids Res*. **33**(5):1544-52(2005).
12.Krauthammer M, Nenadic G. *J Biomed Inform.* **37**(6):512-26(2004).
13.Cohen AM, Hersh WR. *Brief Bioinform* **6**(1):57-71 (2005)
14.Hirschman L, Yeh A, Blaschke C, Valencia A. *BMC Bioinform* **6** S1(2005)
15.Perez AJ,Perez-Iratxeta C,Bork P, et al.*Bioinformatics* **20**(13),2084-91(2004).
16. Koike A, Niwa J, Takagi T. : *Bioinformatics.* Apr 1;21(7):1227-36(2005).
17.Aronson AR. *Proc AMIA Symp* 17-21 (2001).
18.Nadkarni P, Chen R, Brandt C. *J Am Med Inform Assoc*. **8**(1):80-91(2001).
19.Friedman C, Shagina L, Lussier Y, Hripcsak G. *JAMIA* **11**(5):392-402 (2004).
20.King OD, Lee JC, Dudley AM, et al. *Bioinform* **19** Suppl 1:i183-9 (2003).
21.Rogers FB. Medical subject headings. *Bull Med Libr Assoc.* **51**, 114-6 (1963).
22.Cantor M, Sarkar,Bodenreider O,Lussier YA.*Pac Symp Biocomp.*103-14(2005).
23.Chen L, Friedman C *Medinfo* **11**(Pt 2):758-62 (2004).
24.Lussier YA, Li J. *Pac Symp Biocomput* 202-13(2004).
25.Sarkar I, Cantor M, Lussier YA. *Pac Symp Biocomput* 439-50(2003).