RECOGNITION AND NORMALIZATION OF TERMINOLOGY FROM LARGE BIOMEDICAL ONTOLOGIES AND THEIR APPLICATION FOR PHARMACOGENE AND PROTEIN FUNCTION PREDICTION

by

CHRISTOPHER STANLEY FUNK

B.S., Baylor University, 2009

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Computational Bioscience Program
2015

This thesis for the Doctor of Philosophy degree by

Christopher Funk

has been approved for the

Computational Bioscience Program

by

Kevin B. Cohen, Chair

Lawrence E. Hunter, Advisor

Karin M. Verspoor

Asa Ben-Hur

Joan Hooper

Date 4/29/2015

Funk, Christopher Stanley (Ph.D., Computational Bioscience)

Recognition and Normalization of Terminology From Large Biomedical Ontologies and their Application for Pharmacogene and Protein Function Prediction

Thesis directed by Professor Lawrence E. Hunter

ABSTRACT

In recent years many high-throughput techniques have enabled the study and production of large amounts of biomedical data; as a result, the biomedical literature is growing at an exponential rate and shows no sign of slowing. With this comes a significant increase in knowledge that remains unreachable for many applications due to their reliance on manually curated database or software that is unable to scale. This dissertation focuses on benchmarking and improving performance of scalable biomedical concept recognition systems along with exploring the utility of text-mined features for biomedical discovery.

The initial focus of my dissertation is on the task of concept recognition from free text – identifying semantic concepts from well utilized and community supported open biomedical ontologies. Identifying these concepts and grounding text to known ontological identifiers allows any application to also leverage the vast knowledge associated and curated to the concept. I establish current baselines for recognition through a rigorous evaluation and full parameter exploration of three concept recognition systems on eight biomedical ontologies using the CRAFT corpus as gold-standard. Additionally, I create synonym expansion rules that show improved performance, specifically recall, of Gene Ontology concepts.

The later chapters focus on the application of text-mining features obtained from large literature collections for biomedical discovery. The two specific problems presented are the prediction of pharmacogenes and automated protein function prediction. Information contained in the literature has only begun to be harnessed and incorporated into machine learning systems to make biomedical predictions, so I explore two widely open questions:

1. What information should be mined from the literature?

2. How should it be combined with other of data, both literature and sequenced-based?

I demonstrate that improving the ability to recognize concepts from the Gene Ontology produces more informative functional predictions and illustrate that not only can literature features be helpful for making prediction but offer the ability to aid in validation.

Other contributions of this dissertation include publicly available software, a fast user friendly concept recognition and evaluation pipeline along with the hand-crafted compositional rules for increasing recognition of Gene Ontology concepts.

The form and content of this abstract are approved. I recommend its publication.

Approved: Lawrence E. Hunter

To my loving wife, Charis, and my ever supportive parents, Ralph and Linda...

ACKNOWLEDGMENT

My family, friends, and colleagues have been an incredible support as I've been working the past years. I would first like to thank my loving wife, Charis. She has been very supportive during these past years. She has been understanding on the evenings that I must work along with the time away for conferences. I would also like to thank my parents, Ralph and Linda Funk, both educators who have stressed the importance of education growing up. They have provided support daily with their prayers and encouraging words; from the beginning they have encouraged me to pursue what I love.

I would like to thank all the professors in the Computational Bioscience Program (CPBS) for taking a chance on a student with only a bachelors degree and only marginal research experience. I also want to thank my entire thesis committee: Larry Hunter, Kevin Cohen, Karin Verspoor, Asa Ben-Hur, and Joan Hooper. It has been wonderful working with them all the past five years; I've learned a lot from watching them run their labs, apply for grants, and am encouraged through observing their enthusiasm for research. I would like to specifically thank my thesis advisor Larry Hunter. With his continual stressing of the importance of networking, he has pushed an introvert out of his comfort zone. Next I would like to thank my committee chair Kevin Cohen. He has kept me on track during my time here and always has his door open for any conversations – from work to sharing embarrassing memories. I would like to thank Karin Verspoor for agreeing to take on a student that was rough around the edges and for continuing to mentor me despite the 15 hour time difference. My entire committee has shaped my research and this dissertation; their tough feedback and questions have made me a better scientist and prepared me for my future career.

I would next like to thank the current and past students within CPBS: Bill Baumgartner, David Knox, Alex Poole, Michael Hinterberg, Meg Pirrung, Mike Bada, Kevin Livingston, Ben Garcia, Charlotte Siska, and Gargi Datta. I have enjoyed collaborating They have been always been willing to listen to ideas and provide honest feedback. Not only have they been work colleagues, but they have been friends; I look forward to keeping up and seeing the amazing places we all end up. I would also like to thank students Betsy Siewert and Daniel

Dvorkin for saving me countless weeks worth of work time by proving a LaTeX template for this dissertation.

I would like to thank the program administration Kathy Thomas, Elizabeth Wethington, and Liz Pruett. Their knowledge and experience of the inner workings of the graduate school has made that part. I would also like to thank Dave Farrell for his help with any technical difficulties that came my way.

In the following sections, I acknowledge my collaborators and co-authors on the specific projects presented in this dissertation.

Chapter II I would like to thank William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, Kevin Cohen, Lawrence E Hunter, and Karin Verspoor, my co-authors for this original work. I give special thanks to both Bill Baumgartner and Christophe Roeder for their hours of time spent teaching and helping me to learn UIMA. I would like to acknowledge Willie Rodgers from the National Library of Medicine for his help with MetaMap. I appreciate the time spent by the three anonymous reviewers who read and helped to improve this work.

Chapter III I would like to thank Kevin Cohen, Lawrence E Hunter, and Karin Verspoor, my co-authors for this original work. The comments from these co-authors has pushed me beyond the original scope of the work, but has greatly improved this work. I also would like to thank Mike Bada, Judy Blake, and Chris Mungall for their input and discussions about the synonym generation rules.

Chapter IV I would like to thank Kevin Cohen and Lawrence E Hunter, my co-authors for this original work. I appreciate the committee organizers for the "text-mining for biomedical discovery" session at PSB (not to mention the wonderfully chosen conference location) and the time spent by the three anonymous reviewers who read and helped to improve this work.

Chapter V I would like to thank Karin Verspoor, Asa Ben-Hur, Artem Sokolov, Kiley Graim, and Indika Kahanda, who we all instrumental in accomplishing the work in this chapter. I thank the organizers for the two CAFA community challenges, for providing the motivation, data, and man hours spent evaluating many predictions. I appreciate the time spent by the three anonymous reviewers who read and helped to improve all papers combined for this work.

Chapter VI I would like to thank Karin Verspoor, Asa Ben-Hur, and Indika Kahanda, my co-authors for this work. I appreciate the time spent by the two anonymous reviewers who read and helped to improve this work.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER I

INTRODUCTION

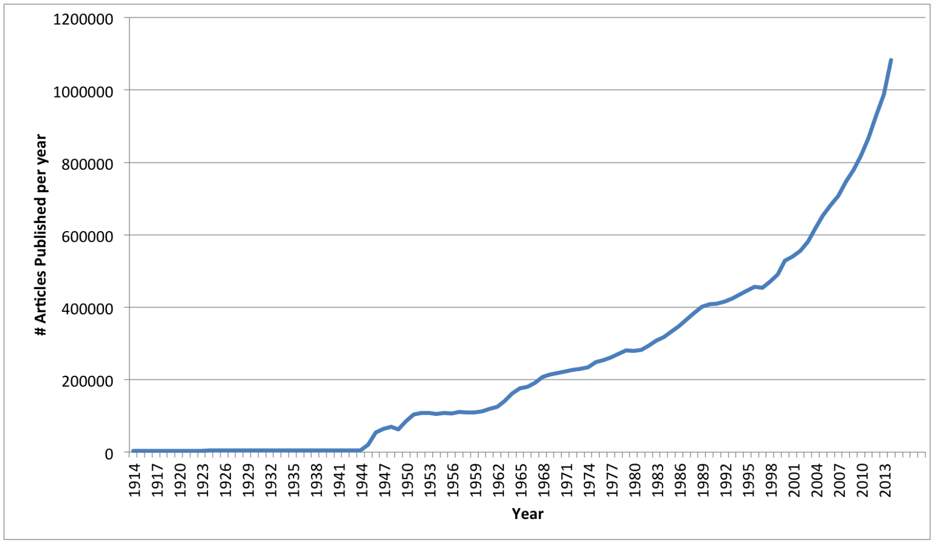


Figure 1.1: Publications published per year for the last 100 years. Plotting the number of articles published each year from 1914-2014; the last couple of yeas has shown an explosion in the number of publications per year.

Currently, there are over 24 million published biomedical articles indexed in PubMed. The biomedical literature is growing at an exponential rate and shows no sign of slowing down (Figure 1.1). With this explosion in publications comes a great increase in knowledge. Unfortunately, most of the knowledge is trapped within the original publication due to the article being located behind a paywall and remain unaccessible to many readers or it could take years for information contained to be curated into databases used by the majority of researchers; this work is focused with the latter. It has been well documented that with this explosion in literature, manual curation cannot keep up (Baumgartner et al., 2007b). Additionally, Howe *et al* state the following about manual curation: “Extracting, tagging with controlled vocabulary and representing data from literature are some of the most important and time-consuming tasks”(Howe et al., 2008). As computational biologists working with biomedical text mining, we shoulder the burden of extracting useful and relevant information from the overbearing amount of literature and translating that biomedical knowledge into advances in health and understanding of complex biological systems.

The main theme that runs through this dissertation is concept normalization, tagging of the biomedical literature with semantic annotations from community supported ontologies. This enables the linking of text to other vast knowledge sources and is an important step in many other more complex tasks, such as relation extraction or integration with semantic web applications (Spasic et al., 2005). Additionally, I explore the effectiveness of features derived from these semantic annotations mined from large literature collections in conjunction with machine learning methods for biomedical prediction and validation – both by themselves and in combination with complimentary biological data.

In this chapter, I introduce concepts related to the work presented in this dissertation. For brevity, I cover topics and relevant work that is not discussed further within the corresponding chapters. Many of the topics covered here are discussed with more depth elsewhere; I provide references to this work that has proven helpful during my dissertation work. I conclude with a description of what can be found in each of the following chapters of my dissertation.

1.1 Biomedical ontologies

Ontologies have grown to be one of the great enabling technologies of modern computational biology, particularly in areas like model organism database curation, where they have facilitated large-scale linking of genomic data across organisms, but also in fields like analysis of high-throughput data (Khatri and Draghici, 2005) and protein function prediction (Krallinger et al., 2005; Sokolov et al., 2013b). Ontologies have also played an important role in the development of natural language processing systems in the biomedical domain, which can use ontologies both as terminological resources and as resources that provide important semantic constraints on biological entities and events (Hunter et al., 2008). Ontologies provide such systems with a target conceptual representation that abstracts over variations within the text. This conceptual representation of the content of documents in turn enables development of sophisticated information retrieval tools that organize documents based on categories of information in the documents (Muller et al., 2004; Doms and Schroeder, 2005; Van Landeghem et al., 2012).

There are over 420 biomedical ontologies contained within the National Center for Biomedical Ontologies (NCBO) (Noy et al., 2009) containing a multitude of different concepts. The community has come together to create, develop, and establish relationships between ontologies in the Open Biomedical Ontologies (OBO) for scientific advancement http://www.obofoundry.org. To help to understand the information contained within an ontology, an entry of a concept in OBO format from the Cell Ontology is shown below (Figure 1.2). Each concept has a unique identifier and term name along with manually curated text defining exactly what the concept represents. Alternative ways to refer to terms are expressed as synonyms; there are many types of synonyms that can be specified with different levels of relatedness to the concept (exact, broad, narrow, and related). An ontology can contain a hierarchy among its terms which are expressed in the "is_a" or "part_of" entry. Other types of relationships between concepts within the same ontology are expressed as a "relationship". Some ontological concepts containing links to other ontologies, these are referred to as "cross-products" and are mainly used for reasoning tasks.

```
id: GO:0006900
name: membrane budding
namespace: biological\_process
def: ``The evagination of a membrane resulting in formation of a vesicle.''
synonym: ``membrane evagination'' EXACT
synonym: ``nonselective vesicle assembly'' RELATED
synonym: ``vesicle biosynthesis'' EXACT
synonym: ``vesicle formation'' EXACT
is\_a: GO:0016044 ! membrane organization and biosynthesis
relationship: part\_of GO:0016192 ! vesicle-mediated transport
```

Figure 1.2: Example ontology entry for the concept "membrane budding".

1.2 Biomedical text mining

Biomedical text mining is focused on developing automatic methods aimed at extracting relevant and important data from the biomedical literature and performing further downstream tasks with the data. There are many specific tasks that fall under this umbrella term, such as named entity recognition (Nadeau and Sekine, 2007; Leaman et al., 2008; Tanabe and Wilbur, 2002), concept recognition (Baumgartner Jr et al., 2007; Jonquet et al., 2009; Funk et al., 2014a; Aronson, 2001), text classification/summarization (Reeve et al., 2007;

Donaldson et al., 2003), synonym and abbreviation extraction (Chang and Schutze, 2006; Schwartz and Hearst, 2003; Yu et al., 2006), relationship extraction (Fundel et al., 2007; Ananiadou et al., 2010; Björne et al., 2010; Liu et al., 2011; Kim et al., 2011), question answering (Athenikos and Han, 2010; Zweigenbaum, 2003; Tsatsaronis et al., 2012), hypothesis generation (Stegmann and Grohmann, 2003; Weeber et al., 2003; Srinivasan, 2004). Specialized solutions must be developed for the biomedical domain because methods trained and created on traditional English text, such as newswire, do not transfer well, due to the highly specialized terminology and complex events and relationships that are expressed (e.g. the interaction of a chemical and domain on a protein localized to a cell compartment type during a specific cell cycle phase within the context of a disease). For a broader view of the field there are many wonderful reviews, some more updated than other, presented in (Cohen and Hersh, 2005; Zweigenbaum et al., 2007; Ananiadou and McNaught, 2006; Rodriguez-Esteban, 2009; Aggarwal and Zhai, 2012; Simpson and Demner-Fushman, 2012; Cohen and Demner-Fushman, 2014).

The following sections describes the task of biomedical concept normalization and the corpora that have been developed and annotated by the community to train and evaluate current systems.

1.2.1 Concept recognition/normalization

Concept recognition or normalization[1] is a subtask of biomedical text mining that is concerned with associating specific spans of text with semantic concepts from the set of biomedical ontologies. The task is also known by other names, such as, concept normalization, named entity resolution, named entity normalization, etc. Concept normalization imparts computer readable semantic meaning into unstructured text and by doing so, provides the ability to extract information through knowledge-directed methods (Spasic et al., 2005). It also enables easier representation of what is contained within the literature.

Most methods for concept normalization are dictionary based. A detailed description of current ontology driven and domain specific tools is presented in Section 2.2.1. Machine

[1] During the entire dissertation I refer to the task of tagging spans of text with a biomedical ontology identifier as both normalization and recognition; these terms are interchangeable.

learning methods are used throughout the named entity recognition task, but because of lack of enough training data, do not perform well on the concept normalization task. Combination of both can be seen in the tool *Neji*, which is a machine learning framework for identification of entity classes followed by normalization through dictionary lookup (Campos et al., 2013). It is unknown if the combination performs better than dictionary lookup alone.

1.2.1.1 BioCreative

The Critical Assessment for Information Extraction in Biology (BioCreative, BC) hosts community challenges aimed at evaluating current state-of-the-art methods for many information extraction tasks; these challenges aim to push and advance the field. There are two main competitions that incorporate concept normalization of Gene Ontology concepts as part of the whole evaluation. There are corpora associated with each of these challenges, but they are incomparable with the defined task of concept normalization presented above. Systems were evaluated on their ability to identify concepts that appear with a span of text, but not required to provide exact spans, only supporting sentences/paragraphs. Some of the well performing methods could be utilized and evaluated on a gold-standard corpus, but most solve a different problem than the focus on my dissertation

BioCreative I – task 2

Task 2 of the first BioCreative was focused on manual curation of genes function and consisted of three subtasks (Blaschke et al., 2005): 1) identification of annotation relevant text passages, 2) assignment of GO terms to gene products, and 3) selection of relevant papers. The corpora consisted of 1,015 full text articles (803 training, 113 and 99 for two sets of testing) but contained no gold-standard. For evaluation GOA curators manually checked each submission. There were three main types of methods used: 1) those using pattern matching of words making up GO concepts along with sentential context (Couto et al., 2005; Ehrler et al., 2005; Verspoor et al., 2005; Krallinger et al., 2005), 2) those using machine learning or statistical techniques (Rice et al., 2005b; Ray and Craven, 2005) and, 3) high precision phrasal matching (Chiang and Yu, 2004). Overall, the performance was lacking for all subtasks as the best performing system only had recall of ∼7% (78/1227) on task 2.2.

BioCreative IV – Gene Ontology task

BC IV also had a task focused on manual curation of gene function and consists of two different tasks (Mao et al., 2014): A) retrieving GO evidence for relevant genes and B) predicting GO terms for relevant genes. Task B is closest to the defined task of concept normalization of GO terms. The input for both tasks is a document-gene pair and the expected output for task 1 is the evidence sentence and for task 2 is the relevant GO concepts related to the genes from the paper. The BC4GO corpus consists of 200 full text articles (100 training, 50 development, 50 testing) with fully annotated evidence sentences by model organism database biocurators.

Overall, the best performing systems for each task utilize simple methods. The best performing system (F-measure 0.270) for returning sentences that support a protein function given a gene-document (task A) pair uses distant supervision from GeneRIFs along with simple simple features extracted from the literature (bag-of-words, bigrams, section/topic features, binary presence of genes) to train a model to recognize sentences containing a gene mention (Zhu et al., 2014). They then utilize dictionary lookup to identify which protein is mentioned. The best performing system (F-measure 0.134) for identifying specific GO concepts given a gene-document pair (task B) uses a knowledgebase indexed with curated abstracts, *GOCat*(Gobeill et al., 2013b), to assign the most similar GO classes, using *k*-Nearest Neighbors, to the input text. They used post-processing filter to only submit the top 10 ranked similar GO classes. It is not surprising that this performed so well, as it does not rely on the GO terms to appear exactly in the input text (Gobeill et al., 2013a).

While performance on this assessment was much higher than the original from BC1, with best performing team reaching recall of 10-30%, they still note that mining GO terms from literature is challenging due to the fact that there are many GO concepts (40,000+) and that GO terms are designed for unifying gene function rather than text mining and are rarely found verbatim in the article (only 1/3 were found exactly represented). Additionally, through this evaluation it is again noted that gold standard corpus data for building machine learning approaches is still lacking with only 1,311 GO terms represented within these 200 articles.

1.2.1.2 Corpora

There are multiple corpora manually annotated for the task of concept normalization. They all vary in size, scope, and type of concepts identified. The GENIA corpus is annotated with its own ontology and has been useful for named entity recognition tasks. The two BioCreative, mentioned above, tasks are concerned with finding mentions of Gene Ontology concepts and relating those to proteins. CRAFT is annotated with many different biomedical ontologies and best resembles the semantic annotations desired by the concept normalization task.

GENIA

The GENIA corpus is commonly used for named entity recognition tasks (Kim et al., 2003). It is a very limited domain corpus consisting of 2,000 abstracts with the MeSH terms *Human*, *Blood cell*, and *Transcription factors*. It is annotated with its own ontology of 36 biomedical entities (Kim et al., 2003).

CRAFT

The Colorado Richly Annotated Full Text Corpus (CRAFT) consists of 97 full-text documents selected from the PubMed Central Open Access subset. Each document in the collection serves as evidence for at least one mouse functional annotation. The "public release" consists of 21,000 sentences from 67 articles and is used in Chapters II and III of this dissertation. There are over 100,000 concept annotations from eight different biomedical ontologies (Gene Ontology, Sequence Ontology, NCBI Taxonomy, Entrez Gene, Sequence Ontology, Protein Ontology, ChEBI, Cell Ontology) in this public subset. What differentiates this corpus from the previous discussed is that each annotation specifies the identifier of the concept from the respective ontology along with the beginning and end points of the text span(s) of the annotation. Other corpora link concepts, some only entity classes, to specific sentences, paragraphs, or abstracts.

1.3 A brief into to automated protein function prediction

Experimentally determining the function of a protein is very time consuming and expensive. Computational approaches can help biologists gain insight into what functions novel proteins perform. Function can be hard to define, as it tends to act as an umbrella term for any type of activity a protein can perform. We take the generalized idea presented in Rost *et al* (Rost et al., 2003) that *function is everything that happens to or through a protein.* The community standard assigns the function of proteins using Gene Ontology concepts (The Gene Ontology Consortium, 2000). The three branches of GO specify different functions in which proteins are involved. Molecular Function (MF) describes activities that occur at the molecular level, such as binding and catalytic activities. Biological Process (BP) represents series of events accomplished by one or more ordered MF, such as signal transduction or alpha-glucose transport. Cellular Component (CC) represents the components of a cell, such as nucleus or ribosome. CC can be useful for function prediction because shared subcellular localization can be taken as evidence of shared function.

There have been many different computational methods used to automatically predict function. The most commonly used method is transfer of function based upon homology (Loewenstein et al., 2009) using database search with a tool such as BLAST. The rationale for homology-based transfer is that if two sequences are similar then they have evolved from a common ancestor and share a similar function. Proteins that have common functions tend to share sequence motifs such as functional domains; these motifs are important and if seen are very indicative of having a certain function. Because the structure of proteins is more important and more conserved than their sequence (Illergård et al., 2009), examining the structure of a protein can identify folds that are indicative of a specific function. Gene expression data is able to identify genes that are somehow related; individual functions cannot be assigned, but a general pathway could be.

Machine learning techniques are able to combine many different sources of heterogeneous data. The classification task being addressed is whether a protein should be associated with a given GO term. Support vector machines (SVMs) (Ben-Hur et al., 2008) are one way to address this problem. The first SVMs for function prediction built individual binary classifiers to classify proteins either as associated or not associated with a given GO term

(Pavlidis et al., 2002). The problem with constructing individual binary classifiers for each GO term comes from the hierarchical nature of the Gene Ontology; a protein can be associated with a term but not associated with its parent, which violates the hierarchy. More recent research have used that information in their predictions (Mostafavi and Morris, 2009). GOstruct (Sokolov et al., 2013b) is a state-of-the-art SVM specifically designed for automated function prediction; it represents the entire GO hierarchy as a binary vector. Doing so enables prediction of all GO terms at once. The GOstruct framework will be used throughout this dissertation in Chapters V and VI.

There have been two main critical assessments to evaluate the progress of automated function prediction, Mousefunc (Peņa-Castillo et al., 2008) and Critical Assessment of Function Annotations (CAFA) http://biofunctionprediction.org. The goal of Mousefunc was to provide GO predictions for a set of genes in *M. Musculus*. Labeled training data was provided (gene expression, PPI, protein domain information, and phylogenetic profiles) with removed gene IDs to prevent supplementing the training data with other data. After much evaluation, the main takeaway is we are unable to predict function well (at a recall of 20%, they achieved 41% precision). CAFA is the most recent critical assessment. No training data provided, only sequences with protein IDs; teams were free to use whatever data they desired. The goal was to predict functions for sets of proteins from different organisms. The main takeaway from CAFA is that there is still room for improvement.

During the CAFA competitions there have only been two teams to utilize text-mined features for function prediction, they are described further within Chapters V and VI.

1.4 Dissertation goals

I now briefly describe the content, hypotheses tested, and main conclusions of each dissertation chapter.

1.4.1 Chapter II

Chapter II presents a large-scale comparative evaluation of concept normalization in the biomedical domain. The end goal was to establish baselines for normalization of concepts from eight biomedical ontologies utilizing the Colorado Richly Annotated Full Text (CRAFT) corpus. For most software packages, using the default parameters is common practice; one hypothesis tested is that ontologies would perform better under different parameters. I performed full parameter exploration of three different dictionary lookup systems (NCBO Annotator, MetaMap, and ConceptMapper) and reach the conclusion that for most ontologies the defaults are not optimal. In fact, optimal parameters are different based upon ontological characteristics. With this knowledge, I provide recommendations for best parameter settings for use with any ontology noting specific characteristics.

Additional topics such as tuning parameters for precision or recall, exploring interacting parameters, and tool ease of use are discussed. One thing that will become very evident throughout the chapter is the importance of morphological variation for matching text to a dictionary; this idea is further explored and used in Chapter III. Another important result of the work in this chapter is a user friendly concept normalization and evaluation pipeline. This pipeline has been utilized in all subsequent chapters along with various projects not presented in this thesis. It is made public and freely available and presented for the advancement of the concept normalization field.

1.4.2 Chapter III

Having established baselines for concept normalization and noting the poor performance on the complex Molecular Function and Biological Process branches of the Gene Ontology, Chapter III focuses on improving performance, specifically recall, of concepts from the Gene Ontology. To achieve this goal, I manually created natural language synonym generation rules that take into account the known compositional nature of GO concepts. First, concepts are recursively decomposed into their smallest composite concepts, then synonyms are generated for these concepts through derivational rules, finally as synonyms are compositionally combined, syntactic variation is introduced. Applying the rules generates ∼1.5 million new synonyms for over two-thirds of all concepts in GO. The hypothesis that overgeneration will not hinder performance due to incorrect synonyms not being found in text was tested and confirmed.

Both intrinsic and extrinsic evaluations were performed to estimate impact of the generated synonyms on normalization of concepts. The CRAFT corpus was used for intrinsic evaluation and an increase in F-measure performance of ∼0.2 was seen. A large collection of one million full text documents was used for extrinsic evaluation. Manual validation and error analysis of random representative samples reveals that synonyms generated through the rules have reasonable accuracy (0.82) while the accuracy over all concepts is higher (0.88). The synonyms generated help increase recall of GO concepts by bridging the gap between representing in ontology and expression in natural language. Compositional rules are not only useful for GO, I discuss and provide examples of how similar rules could generalize to other biomedical ontologies.

1.4.3 Chapter IV

Chapter IV marks a shift in the thesis; moving from a focus on the task of concept normalization to the application of concept normalization for biomedical discovery. The first application discussed is the ability of text-mined GO concepts to predict disease related or pharmacogenes on a genome-wide scale; genes where a variant could affect drug response or be implicated within a disease. The hypothesis that there is a common set of functions that known pharmacogenes share is tested and confirmed; the common set of enriched functions shared by known pharmacogenes is analyzed. Using this as a basis for classification, I explore multiple machine learning algorithms with combinations of curated GO functions, text-mined GO terms, and surface linguistic features (bigrams and collocations). Using an SVM implementation and text-mined GO terms and bigrams as features the classifier was able to distinguish known pharmacogenes from a background set with an F-measure performance of 0.86 and AUC of 0.860 on 5-fold cross validation. Using only information mined from the literature, our classifier was able to predict 114 yet uncurated pharmacogenes.

The top 10 predicted pharmacogenes, ranked by similarity to the known pharmacogenes, were manually examined with respect to other datasources. When the work was originally done, none of the predicted genes had clinical annotations within PharmGKB, but a few had specific variants associated with disease from OMIM. As time passes, curated annotations within databases accrue and are able to serve as new knowledge and validate predictions made by past experiments. In the ∼2 years since the original evaluation in this chapter was performed, 6 of the top 10 predicted genes now have at least one clinical variant within PharmGKB and a few have new disease variants.

1.4.4 Chapter V

Chapter V continues the application of concept normalization for biomedical discovery; this chapter outlines my work on the automated function prediction task. We introduce the idea of a co-mention – co-occurrence of two entities within a predefined span of text. These are the primary literature features utilized for function prediction. I explore different ways to combine co-mention feature sets. They are combined with sequence- and network-based features within the GOstruct framework, a state-of-the-art support vector machine framework designed for prediction of Gene Ontology terms.

I begin by discussing our work pertaining to the first and second Critical Assessment of Functional Annotation (CAFA), a community challenge for automated function prediction. Each competition had slightly different goals and evaluation criteria. I highlight the differences and discuss the impact on both protein and GO concept normalization and literature collection selection. For both systems designed for a CAFA challenge we performed external experimentation and validation of predictions. I present two different evaluations of literature features vs. sequence- and network-based features and the combination. I find that literature features alone approach the performance of commonly used sequence-based features but the combination of both produces the best predictions.

1.4.5 Chapter VI

The dissertation concludes with an in-depth study on the impact of literature features on the function prediction task. This is the culmination of my work on concept normalization (presented in Chapters II and III) and function prediction (presented in Chapter V). I specifically focus on normalization of GO concepts. Two sets of co-mentions are mined from the literature with differing Gene Ontology dictionaries: 1) using only official Gene Ontology information and 2) using the compositional Gene Ontology synonym generation rules presented in Chapter III. I reach the conclusion that increasing the ability to recognize GO concepts from the biomedical text leads to more informative function predictions. Additionally, simple bag-of-words features are explored and produce surprisingly good performance, but I argue the extra work required to recognize co-mentions is valuable because it offers the ability to easily verify predictions based upon the specific literature context of co-mentions. To aid in manual analysis of co-mentions I developed a “medium-throughput” co-mention curation pipeline that has the possibility of speeding up the process of protein function curation.

CHAPTER II

LARGE SCALE EVALUATION OF CONCEPT RECOGNITION[2]

2.1 Introduction

All chapters of my dissertation incorporate concept recognition to some extent. Before any other work can be presented, it is important to explore how accurately systems can recognize concepts from text. As described in Section 1.2.1, there have been very few linguistic oriented concept recognition evaluations – where the concept is grounded to not only an ontological identifier but also a specific span of text. This chapter evaluates three dictionary based concept recognition systems and performs full parameter exploration.

This evaluation is important for my dissertation through establishing baselines of performance for later comparison. It also aids in the determination of which the concept recognition system is used and provides insights into correct parameter values to set for best performance depending on ontological characteristics. This work has received lots attention and is important for the field of natural language processing in that it is a linguistically oriented and rigorous evaluation like none performed before for biomedical concept recognition.

2.2 Background

Ontologies have grown to be one of the great enabling technologies of modern bioinformatics, particularly in areas like model organism database curation, where they have facilitated large-scale linking of genomic data across organisms, but also in fields like analysis of high-throughput data (Khatri and Draghici, 2005) and protein function prediction (Krallinger et al., 2005; Sokolov et al., 2013b). Ontologies have also played an important role in the development of natural language processing systems in the biomedical domain, which can use ontologies both as terminological resources and as resources that provide important semantic constraints on biological entities and events (Hunter et al., 2008). Ontologies provide such systems with a target conceptual representation that abstracts over

[2] The work presented in this chapter is republished with permission from: *Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters* (BMC bioinformatics 15.1 (2014): 59).

variations in the surface realization of terms. This conceptual representation of the content of documents in turn enables development of sophisticated information retrieval tools that organize documents based on categories of information in the documents (Muller et al., 2004; Doms and Schroeder, 2005; Van Landeghem et al., 2012).

Finally, ontologies themselves can benefit from concept recognition in text. Yao *et al* (Yao et al., 2011) propose new ontology quality metrics that are based on the goodness of fit of an ontology with a domain-relevant corpus. They note that a limitation of their approach is the dependency on tools that establish linkages between ontology concepts and their textual representations.

However, a general approach to recognition of terms from any ontology in text remains a very open research problem. While there exist sophisticated named entity recognition tools that address specific categories of terms, such as genes or gene products (Settles, 2005), protein mutations (Caporaso et al., 2007), or diseases (Jimeno, 2008; Leaman et al., 2008), these tools require targeted training material and cannot generically be applied to recognize arbitrary terms from large, fine-grained vocabularies (Doms and Schroeder, 2005). Furthermore, as Brewster *et al* (Brewster et al., 2004) point out, there is often a disconnect between what is captured in an ontology and what can be expected to be explicitly stated in text. This is particularly true for relations among concepts, but it is also the case that concepts themselves can be expressed in text with a huge amount of variability and potentially ambiguity and underspecification (Verspoor et al., 2003; Cohen et al., 2008).

The work reported in this chapter to advance the state of the art in recognizing terms from ontologies with a wide variety of differences in both the structure and content of the ontologies and in the surface characteristics of terms associated with concepts in the ontology. We evaluate a number of hypotheses related to the general task of finding references to concepts from widely varying ontologies in text. These include the following:

- Not all concept recognition systems perform equally on natural language texts.
- The best concept recognition system varies from ontology to ontology.
- Parameter settings for a concept recognition system can be optimized to improve performance on a given ontology.

- Linguistic analysis, in particular morphological analysis, affects the performance of concept recognition systems.

To test these hypotheses, we apply a variety of dictionary-based tools for recognizing concepts in text to a corpus in which nearly all of the concepts from a variety of ontologies have been manually annotated. We perform an exhaustive exploration of the parameter spaces for each of these tools and report the performance of thousands of combinations of parameter settings. We experiment with the addition of tools for linguistic analysis, in particular morphological analysis. Along with reporting quantitative results, we give the results of manual error analysis for each combination of concept recognition system and ontology.

The gold standard used is the Colorado Richly Annotated Full-Text (CRAFT) Corpus (Verspoor et al., 2012; Bada et al., 2012). The full CRAFT corpus consists of 97 completely annotated biomedical journal articles, while the "public release" set, which consists of 67 documents, was used for this evaluation. CRAFT includes over 100,000 concept annotations from eight different biomedical ontologies. Without CRAFT, this large-scale evaluation of concept annotation would not have been possible, due to lack of corpora annotated with a large number of concepts from multiple ontologies.

2.2.1 Related work

A number of tools and strategies have been proposed for concept annotation in text. These include both tools that are generally applicable to a wide range of terminology resources, and strategies that have been designed specifically for one or a few terminologies. The two most widely used generic tools are the National Library of Medicine's MetaMap (Aronson, 2001) and NBCO's Open Biomedical Annotator (NCBO Annotator)(Shah et al., 2009), based on a tool from the University of Michigan called MGREP. Other tools, including Whatizit (Rebholz-Schuhmann, 2008), KnowledgeMap (Denny et al., 2003, 2005), CONANN (Reeve and Han, 2007), IndexFinder (Zou et al., 2003; Chu Wesley W, 2007), Terminizer (Hancock et al., 2009), and Peregrine (Schuemie et al., 2007; Kang et al., 2012) have been created but are not publicly available or appear not to be in widespread use. We therefore focus our analysis in this work on the NCBO Annotator and MetaMap. In addition, we include ConceptMapper (Sandbox, 2009; Tanenblatt et al., 2010), a tool that was not specifically developed for biomedical term recognition but rather for flexible look up of terms from a dictionary or controlled vocabulary.

The tools MGREP and MetaMap have been directly compared on several term recognition tasks (Shah et al., 2009; Stewart et al., 2012). These studies indicate that MGREP outperforms MetaMap in terms of precision of matching. Both studies also note that MetaMap returns many more annotations than MGREP. Recall is not calculated in either study because the document collections used as input were not fully annotated. By using a completely annotated corpus such as CRAFT, we are able to generate not only precision but recall, which gives a complete picture of the performance of the system.

The Gene Ontology (The Gene Ontology Consortium, 2000) has been the target of several customized methods that take advantage of the specific structure and characteristics of that ontology to facilitate recognition of its constituent terms in text (Krallinger et al., 2005; Verspoor et al., 2005; Ray and Craven, 2005; Couto et al., 2005; Koike et al., 2005). In this work, we will not specifically compare these methods to the more generic tools identified above, as they are not applicable to the full range of ontologies that are reflected in the CRAFT annotations.

The CRAFT corpus has been utilized previously in the context of evaluating the recognition of specific categories of terms. Verspoor *et al.* (Verspoor et al., 2012) provide a detailed assessment of named entity recognition tool performance for recognition of genes and gene products. As with the work mentioned in the previous paragraph, these are specialized tools with a more targeted approach than we explore in this work, typically requiring substantial amounts of training material tailored to the specific named entity category. We do not repeat those experiments here as they are not relevant to the general problem of recognition of terms from large controlled vocabularies.

2.2.2 A note on "concepts"

We are aware of the controversies associated with the use of the word "concept" with respect to biomedical ontologies, but the content of this work is not affected by the conflicting positions on this issue; we use the word to refer to the tuple of namespace, identifier, term(s), definition, synonym(s), and metadata that make up an entry in an ontology.

2.3 Methods

2.3.1 Corpus

We used version 1.0, released October 19, 2012, of the Colorado Richly Annotated Full Text Corpus (CRAFT) data set (Verspoor et al., 2012; Bada et al., 2012). The full corpus consists of 97 full-text documents selected from the PubMed Central Open Access subset. Each document in the collection serves as evidence for at least one mouse functional annotation. For this work we used the "public release" subsection, which consists of 21,000 sentences from 67 articles. There are over 100,000 concept annotations from eight different biomedical ontologies in this public subset. Each annotation specifies the identifier of the concept from the respective ontology along with the beginning and end points of the text span(s) of the annotation.

To fully understand the results presented, it is important to understand how CRAFT was annotated (Bada et al., 2012). Here we present three guidelines. First, the text associated with each annotation in CRAFT must be semantically equivalent to the term from the ontology with which it is annotated. In other words, the text, in its context, has the same meaning as the concept used to annotate it. Second, annotations are made to a specific ontology and not to a domain; that is, annotations are created only for concepts explicitly represented in the given ontology and not to concepts that "should" be in the ontology but are not explicitly represented. For example, if the ontology contains a concept representing vesicles, but nothing more specific, a mention of "microvesicles" would not be annotated: Even though it is a type of vesicle, it is not annotated because microvesicles are not explicitly represented in the ontology and annotating this text with the more general vesicle concept would not be semantically equivalent, i.e., information would be lost. Third, only text directly corresponding to a concept is tagged; for example, if the text "mutant vesicles" is seen,"vesicles" is tagged by itself (i.e. without "mutant") with the vesicle concept. Because only the most specific concept is annotated, there are no subsuming annotations; that is, given an annotation of a text span with a particular concept, no annotations are made within this text span(s) with a more general concept even if they appear in the term. For an example from the Cell Type Ontology, given the text “mesenchymal cell”, this phrase is annotated with “CL:0000134 - mesenchymal cell” but the nested “cell” is not additionally annotated with “CL:0000000 - cell”, as the latter is an ancestor of the former and therefore redundant. There are very specific guidelines as to what text is included in an annotation set out in Bada *et al.* (Bada et al., 2010).

2.3.2 Ontologies

The annotations of eight ontologies, representing a wide variety of biomedical terminology, were used for this evaluation: 1-3) The three sub-ontologies of the Gene Ontology (Biological Process, Molecular Function, Cellular Component) (The Gene Ontology Consortium, 2000) 4) the Cell Type Ontology (Bard et al., 2005) 5) Chemical Entities of Biological Interest Ontology (Degtyarenko, 2003) 6) the NCBI Taxonomy (Wheeler et al., 2006) 7) the Sequence Ontology (Eilbeck et al., 2005) and 8) the Protein Ontology (Natale et al., 2011). Versions of ontologies used along with descriptive statistics can be seen in Table 2.1. CRAFT also contains Entrez Gene annotations, but these were analyzed in previous work (Verspoor et al., 2012). The Gene Ontology (GO) aims to standardize the representation of gene and gene product attributes; it consists of three distinct sub-ontologies, which are evaluated separately: Molecular Function, Biological Process, and Cellular Component. The Cell Type Ontology (CL) provides a structured vocabulary for cell types. Chemical Entities of Biological Interest (ChEBI) is focused on molecular entities, molecular parts, atoms, subatomic particles, and biochemical roles and applications. NCBI Taxonomy (NCBITaxon) provides classification and nomenclature of all organisms and types of biological taxa in the public sequence database. The Sequence Ontology (SO) aims to describe the features and attributes of biological sequences. The Protein Ontology (PRO) provides a representation of protein-related entities.

2.3.3 Structure of ontology entries

The ontologies used are from the Open Biomedical Ontologies (OBO) (biomedical ontologies) flat file format. To help to understand the structure of the file, an entry of a

Table 2.1: **Characteristics of ontologies evaluated.**

Ontology	**Version**	**# Concepts**	**Avg. Term Length**	**Avg. Words in Term**	**Avg. # Synonyms**	**% Have Punctuation**	**% Have Numerals**	**% Have Stop Words**
Cell Type	25:05:2007	838	20.0±9.5	3.0±1.4	0.5±1.1	11.6	4.8	3.3
Sequence	30:03:2009	1,610	21.6±13.3	3.1±1.0	1.4±1	91.9	6.6	9.3
ChEBI	28:05:2008	19,633	25.5±24.2	4.3±4.8	2.0±2.5	54.8	41.3	0
NCBITaxon	12:07:2011	789,538	24.6±10.2	3.6±2.0	N/A	53.7	56.0	0.3
GO-MF	28:11:2007	7,984	39.1±15.4	4.6±2.2	2.8±4.6	52.8	26.6	2.7
GO-BP	28:11:2007	14,306	40.1±19.0	5.0±2.7	2.1±2.5	23.5	7.0	45.7
GO-CC	28:11:2007	2,047	26.6±14.2	3.6±1.7	0.1±0.9	29.5	14.4	6.8
Protein	22:04:2011	26,807	38.4±18.5	5.5±2.5	3.1±3.2	68.4	74.8	4.3

concept from CL is shown below. The only parts of an entry used in our systems are the id, name, and synonym rows. Alternative ways to refer to terms are expressed as synonyms; there are many types of synonyms that can be specified with different levels of relatedness to the concept (exact, broad, narrow, and related). An ontology contain a hierarchy among its terms; these are expressed in the "is_a" entry. Terms described as "ancestors", "less specific", or "more general" lie above the specified concept in the hierarchy, while terms described as "more specific" are below the specified concept.

id: CL:0000560

name: band form neutrophil

def: "A late neutrophilic metamyelocyte in which the nucleus is in the form of a curved or coiled band, not having acquired the typical multi lobar shape of the mature neutrophil."

synonym: "band cell" EXACT

synonym: "rod neutrophil" EXACT

synonym: "band" NARROW

is_a: CL:0000776 ! immature neutrophil

relationship: develops_from CL:0000582   neutrophilic metamyelocyte

2.3.4 A note on obsolete terms

Ontologies are ever changing: new terms are added, modifications are made to others, and others are made obsolete. This poses a problem because obsolete terms are not removed from the ontology, but only marked as obsolete in the obo flat file. The dictionary-based methods used in our analysis do not distinguish between valid or obsolete terms when creating their dictionaries, so obsolete terms may be returned by the systems. A filter was incorporated to remove obsolete terms returned (discussed more below). Not filtering obsolete terms introduces many false positives. For example, the terms "GO:0005574 - DNA" and "GO:0003675 - protein" are both obsolete in the cellular component branch of the Gene Ontology and are mentioned very frequently within the biomedical literature.

2.3.5 Concept recognition systems

We evaluated three concept recognition systems, NCBO Annotator (NCBO Annotator)(Jonquet et al., 2009), MetaMap (Aronson, 2001), and ConceptMapper (Sandbox, 2009; Tanenblatt et al., 2010). All three systems are publicly available and able to produce annotations for many different ontologies but differ in their underlying implementation and amount of configurable parameters. The full evaluation results are available for download at http://bionlp.sourceforge.net/.

NCBO Annotator is a web service provided by the National Center for Biomedical Ontology (NCBO) that annotates textual data with ontology terms from the UMLS and BioPortal ontologies. The input text is fed into a concept recognition tool (MGREP) and annotations are produced. A wrapper (Roeder et al., 2010) programmatically converts annotations produced by NCBO into xml, which is then imported into our evaluation pipeline. The evaluations from NCBO Annotator were performed in October and November 2012.

MetaMap (MM) is a highly configurable program created to map biomedical text to the UMLS Metathesaurus. MM parses input text into noun phrases and generates variants (alternate spellings, abbreviations, synonyms, inflections and derivations) from these. A candidate set of Metathesaurus terms containing one of the variants is formed, and scores are computed on the strength of mapping from the variants to each candidate term. In contrast to a Web service, MM runs locally; we installed MM v.2011 on a local Linux server. MM natively works with UMLS ontologies, but not all ontologies that we have evaluated are a part of the UMLS. The optional data file builder (Rodgers et al.) allows MM to use any ontology as long as they can be formatted as UMLS database tables; therefore, a Perl script was written to convert the ontology obo files to UMLS database tables following the specification in the data file builder overview.

ConceptMapper (CM) is part of the Apache UIMA (Ferrucci and Lally, 2004) Sandbox and is available at http://uima.apache.org/d/uima-addons-current/ConceptMapper. Version 2.3.1 was used for these experiments. CM is a highly configurable dictionary lookup tool implemented as a UIMA component. Ontologies are mapped to the appropriate dictionary format required by ConceptMapper. The input text is processed as tokens; all tokens within a span (sentence) are looked up in the dictionary using a configurable lookup algorithm.

2.3.6 Parameter exploration

Each system's parameters were examined and configurable parameters were chosen. Table 2.2 gives a list of each system with the chosen parameters along with a brief description and possible values.

2.3.7 Evaluation pipeline

An evaluation pipeline for each system was constructed and run in UIMA (IBM, 2009). MM produces annotations separate from the evaluation pipeline; UIMA components were created to load the annotations before evaluation. NCBO Annotator is able to produce annotations and evaluate them within the same pipeline, but NCBO Annotator annotations were cached to avoid hitting the Web service continually. Like MM, a separate analysis engine was created to load annotations before evaluation. CM produces annotations and evaluates them in a single pipeline.

Evaluation pipelines for each system have a similar structure. First, the gold standard is loaded; then, the system's annotations are loaded, obsolete annotations are removed, and finally a comparison is made. CRAFT was not annotated with obsolete terms, so the obsolete terms filtered out are those that are obsolete in the version of the ontology used to annotate CRAFT.

CM and MM dictionaries were created with the versions of the ontologies that were used to annotate CRAFT. Since NCBO Annotator is a Web service, we do not have control over the versions of ontologies used; it uses newer versions with more terms. To remove spurious terms not present in the ontologies used to annotate CRAFT, a filter was added to the NCBO Annotator evaluation pipeline. The NCBO Annotator specific filter removes terms not present in the version used to annotate CRAFT and ensures that the term is not

Table 2.2: System parameter description and values. Parameters that were evaluated for each system along with a description and possible values are listed in all capital letters. For the most part, parameters are self-explanatory, but for more information see documentation for each system. CM (Sandbox, 2009), NCBO Annotator (Jonquet et al., 2009), MM (Aronson, 2001).

NCBO Annotator Parameters	
Parameter	**Description andPossible Values**
wholeWordOnly	Term recognition must match whole words - (YES, NO)
filterNumber	Specifies whether the entity recognition step should filter numbers - (YES, NO)
stopWords	List of stop words to exclude from matching - (PubMed- commonly found terms from PubMed, NONE)
stopWordsCaseSensitive	Whether stop words are case sensitive - (YES, NO)
minTermSize	Specifies minimum length of terms to be returned - (ONE, THREE, FIVE)
withSynonyms	Whether to include synonyms in matching - (YES, NO)

MetaMap Parameters	
Parameter	**Description and Possible Values**
model	Determines which data model is used - (STRICT - lexical, manual, and syntactic filtering are applied, RELAXED - lexical and manual filtering are used)
gaps	Specifies how to handle gaps in terms when matching - (ALLOW, NONE)
wordOrder	Specifies how to handle word order when matching - (ORDER MATTERS, IGNORE)
acronymAbb	Determines which generated acronym or abbreviations are used - (NONE, DEFAULT, UNIQUE - restricts variants to only those with unique expansions)
derivationalVars	Specifies which type of derivational variants will be used - (NONE, ALL, ONLY ADJ NOUN)
scoreFilter	MetaMap reports a score from 0-1000 for every match, with 1000 being the highest, those matches with scores ≤ will be returned - (0, 600, 800, 1000)
minTermSize	Specifies minimum length of terms to be returned - (ONE, THREE, FIVE)

ConceptMapper Parameters	
Parameter	**Description and Possible Values**
searchStrategy	Specifies the dictionary lookup strategy - (CONTIGUOUS - longest match of contiguous tokens, SKIP ANY - returns longest match of not-necessarily contiguous tokens and next lookup begin in next span, SKIP ANY ALLOW OVERLAP - returns longest match of not-necessarily contiguous tokens in the span and next lookup begin after next token)
caseMatch	Specifies the case folding mode to use - (IGNORE - fold everything to lower case, INSENSITIVE - fold only tokens with initial caps to lowercase, SENSITIVE - no folding, FOLD DIGIT - fold only tokens with digits to lower case)
stemmer	Name of the stemmer to use before matching - (Porter- classic stemmer that removes common morphological and inflectional endings from English words, BioLemmatizer- domain specific lemmatization tool for the morphological analysis of biomedical literature presented in Liu *et al.* (Liu et al., 212), NONE)
orderIndependentLookup	Specifies if ordering of tokens within a span can be ignored - (TRUE, FALSE)
findAllMatches	Specifies if all matches will be returned - (TRUE, FALSE - only the longest match will be returned)
stopWords	List of stop words to exclude from matching - (PubMed- commonly found terms from PubMed, NONE)
synonyms	Specifies which synonyms will be included when creating the dictionary - (EXACT ONLY, ALL)

obsolete in the version used to annotate CRAFT. Because the versions of the ontologies used in CRAFT are older, it may be the case that some terms annotated in CRAFT are obsolete in the current versions. All systems were restricted to only using valid terms from the versions of the ontology used to annotate CRAFT.

All comparisons were performed using a STRICT comparator, which means that ontology ID and span(s) of a given annotation must match the gold-standard annotation exactly to be counted correct. A STRICT comparator was chosen because it was our desire to see how well automated methods can recreate exact human annotations. A pitfall of the using a STRICT comparator is that a distinction cannot be made between erroneous terms vs. those along the same hierarchical lineage; both are counted as fully incorrect in our analysis. For example, if the gold standard annotation is "GO:0005515 - protein binding" and "GO:0005488 - binding" is returned by a system, partial credit should be given because "binding" is an ancestor of "protein binding". Future comparisons could address this limitation by accounting for the hierarchical relationship in the ontology by counting those less specific terms as partially correct by using hierarchical precision/recall/F-measure as seen in Verspoor *et al* (Verspoor et al., 2006).

The output is a text file for each parameter combination listing true positives (TP), false positives (FP), and false negatives (FN) for each document as well as precision (P), recall (R), and F-measure (F) (Calculations of P, R, and F can be seen in formulas 2.1, 2.2, and 2.3). Precision, recall, and F-measure are calculated over all annotations across all documents in CRAFT, i.e. as a *macro-average*.

$$P = \frac{TP}{TP + FP} \tag{2.1}$$

$$R = \frac{TP}{TP + FN} \tag{2.2}$$

$$F = 2 * \frac{P * R}{P + R} \tag{2.3}$$

2.3.8 Statistical analysis

The Kruskal-Wallis statistical method was chosen to test significance for all our comparisons because it is a non-parametric test that identifies differences between ranked group of variables. It is appropriate for our experiments because we do not assume our data follows any particular distribution and desire to determine if the distribution of scores from a particular experimental condition, such as tool or parameters, are different from the others. The implementation built into R was used (*kruskal.test*). Kruskal-Wallis was applied in three different ways:

1. For each ontology, Kruskal-Wallis was used to determine if there is a significant difference in F-measure performance between tools. The mean and variance was computed across all parameter combinations for a given tool, calculated at the corpus level using the macro-average F-measure and provided as input to Kruskal-Wallis.

2. For each tool, Kruskal-Wallis was used to determine if there is a difference in performance between parameter values for each parameter. The mean and variance was computed across all parameter values for a given parameter, calculated at the corpus level using the macro-average F-measure.

3. Results from Kruskal-Wallis only determine if there is a difference between the groups but does not provide insight into how many differences or between which groups a difference exists. When a significant difference was seen between three or more groups, Kruskal-Wallis was used between a *post hoc* test to identify the significantly different group(s).

Significance is determined at a 99% level, $\alpha = 0.01$; because there are multiple comparisons, a Bonferroni correction was used, and the new significance level is $\alpha = 0.00036$.

2.3.9 Analysis of results files

For each ontology-system pair, an analysis was performed on the maximum F-measure parameter combination. We did not analyze every annotation produced by all systems but made sure to account for ~70-80% of them. By performing the analysis this way, we are concentrating on the general errors and terms missed rather than rare errors.

For each maximum F-measure parameter combination file, the top 50-150 (grouped by ontology ID and ranked by number of annotations for each ID) of each true positive (TP), false positive (FP), and false negative (FN) were analyzed by separating them into groups of like annotations. For example, the types of bins that FPs fall into are: "errors from variants", "errors from ambiguous synonyms", "errors due to identifying less specific concepts", etc., and are different than the bins into which TPs or FNs are categorized.

Because we evaluated all parameter combinations, we were able to examine the impact of single parameters by holding all other parameters constant. The maximum F-measure producing parameter combination result file and the complementary result file with varied parameter were run through a graphical difference program, DiffMerge, to examine the annotations found/lost by varying the parameter. Examples mentioned in the Results and discussion are from this comparison.

2.4 Results and discussion

Results and Discussion are broken down by ontology and then by tool. For each ontology we present three different levels of analysis:

1. At the ontology level. This provides a synopsis of overall performance for each system with comments about common terms correct (TPs), errors (FPs), and categories missed (FNs). Specific examples are taken from the top-performing, highest F-measure parameter combination.

2. A high-level parameter analysis, performed over all parameter combinations. This allows for observation about impact on performance seen by manipulating parameter values, presented as ranges of impact.

3. A low-level analysis obtained from examining individual result files gives insight into specific terms or categories of terms that are affected by manipulating parameters.

Within a particular ontology, each system's performance is described. The most impactful parameters are explored further and examples from variations on maximum F-measure combination are provided to show the effect they have on matching. Results presented as

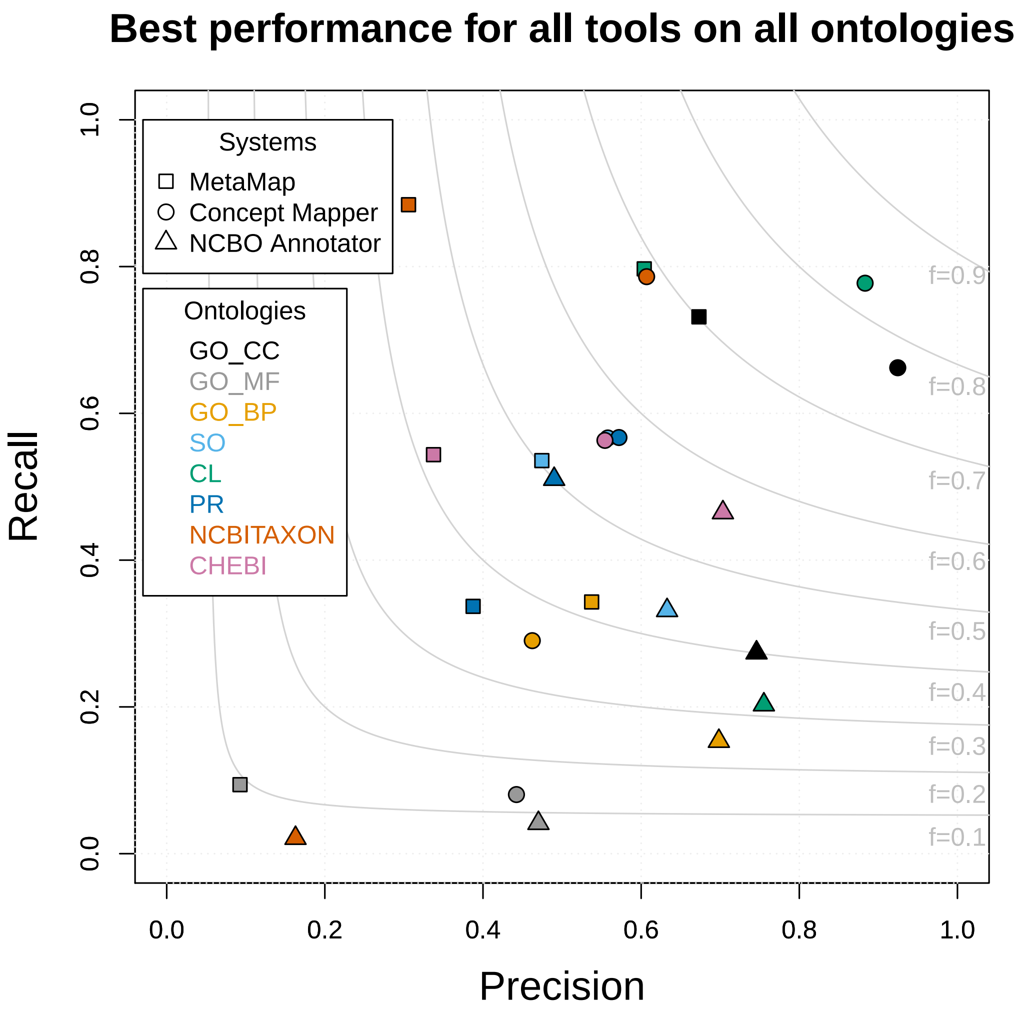


Figure 2.1: Maximum F-measure for each system-ontology pair. A wide range of maximum scores is seen for each system within each ontology.

numbers of annotations are of this type of analysis. We end the results and discussion section with overall parameter analysis and suggestions for parameters on any ontology.

The best-performing result for each system-ontology pair is presented in Figure 2.1. There is a wide range of F-measures for all ontologies, from <0.10 to 0.83. Not only is there a wide range when looking at all ontologies, but a wide range can be seen within each ontology. Two of our hypotheses are supported by this analysis: we can see that not all concept recognition systems perform equally, and the best concept recognition system varies from ontology to ontology.

2.4.1 Best parameters

Based on analysis, the suggested parameters for maximum performance for each ontology-system pair can be seen in Tables 2.3 and 2.4.

Table 2.3: Best performing parameter combinations for CL and GO subsections. Suggested parameters to use that correspond to best score on CRAFT. Parameters where choices don't seem to make a difference in performance are represented as "ANY".

Cell Type Ontology (CL)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	ANY	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	INSENSITIVE
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	Porter or BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	ONE or THREE	derivationalVariants	ALL	orderIndLookup	OFF
withSynonyms	YES	scoreFilter	0	findAllMatches	NO
		minTermSize	1 or 3	synonyms	EXACT ONLY

Gene Ontology - Cellular Component (GO_CC)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	ANY	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	INSENSITIVE
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	Porter
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	ONE or THREE	derivationalVariants	ANY	orderIndLookup	OFF
withSynonyms	ANY	scoreFilter	0 or 600	findAllMatches	NO
		minTermSize	1 or 3	synonyms	EXACT ONLY

Gene Ontology - Molecular Function (GO_MF)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	NO	model	ANY	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	ANY
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	ANY	derivationalVariants	ANY	orderIndLookup	OFF
withSynonyms	NO	scoreFilter	0 or 600	findAllMatches	NO
		minTermSize	1 or 3	synonyms	EXACT ONLY

Gene Ontology - Biological Process (GO_BP)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	ANY	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	INSENSITIVE
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	Porter
SWCaseSensitive	ANY	acronymAbb	ANY	stopWords	NONE
minTermSize	ANY	derivationalVariants	ADJ NOUN VARS	orderIndLookup	OFF
withSynonyms	YES	scoreFilter	0	findAllMatches	NO
		minTermSize	5	synonyms	ALL

2.4.2 Cell Type Ontology

The Cell Type Ontology (CL) was designed to provide a controlled vocabulary for cell types from many different prokaryotic, fungal, and eukaryotic organisms. Out of all

Table 2.4: Best performing parameter combinations for SO, ChEBI, NCBITaxon, and PRO. Suggested parameters to use that correspond to best score on CRAFT. Parameters where choices don't seem to make a difference in performance are represented as "ANY".

Sequence Ontology (SO)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	STRICT	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	INSENSITIVE
stopWords	ANY	wordOrder	ANY	stemmer	Porter or BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	THREE	derivationalVariants	NONE	orderIndLookup	OFF
withSynonyms	YES	scoreFilter	600	findAllMatches	NO
		minTermSize	3	synonyms	EXACT ONLY

Protein Ontology (PRO)					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	ANY	searchStrategy	ANY
filterNumber	ANY	gaps	NONE	caseMatch	CASE FOLD DIGITS
stopWords	PubMed	wordOrder	ANY	stemmer	NONE
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	ONE or THREE	derivationalVariants	NONE	orderIndLookup	OFF
withSynonyms	YES	scoreFilter	600	findAllMatches	NO
		minTermSize	3 or 5	synonyms	ALL

NCBI Taxonomy					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	ANY	searchStrategy	SKIP ANY or ALLOW
filterNumber	ANY	gaps	NONE	caseMatch	ANY
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	PubMed
minTermSize	FIVE	derivationalVariants	NONE	orderIndLookup	OFF
withSynonyms	ANY	scoreFilter	0 or 600	findAllMatches	NO
		minTermSize	3	synonyms	EXACT ONLY

ChEBI					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	STRICT	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	ANY
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	ONE or THREE	derivationalVariants	NONE	orderIndLookup	OFF
withSynonyms	YES	scoreFilter	0 or 600	findAllMatches	YES
		minTermSize	5	synonyms	EXACT ONLY

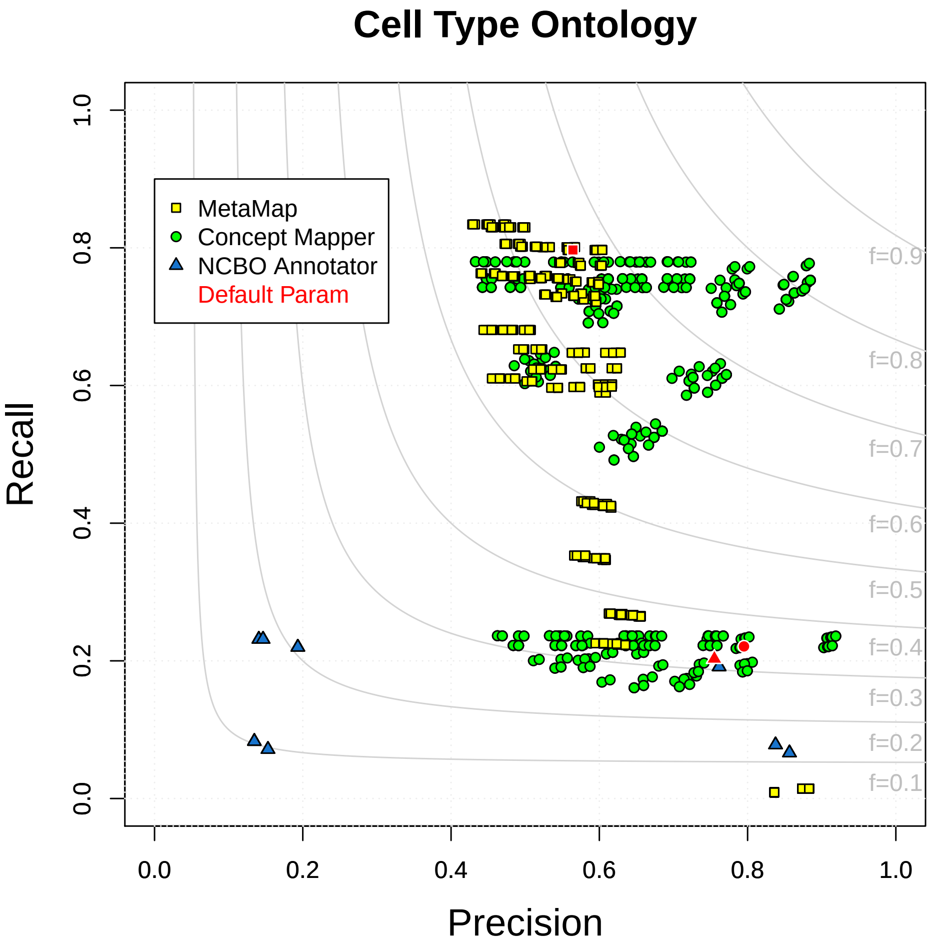


Figure 2.2: All parameter combinations for CL. The distribution of all parameter combinations for each system on CL. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

ontologies annotated in CRAFT, it is the smallest, terms are the simplest, and there are very few synonyms (Table 2.1). The highest F-measure seen on any ontology is on CL. CM is the top performer (F=0.83), MM performs second best (F=0.69), and NCBO Annotator is the worst performer (F=0.32). Statistics for the best scores can be seen in Table 2.5. All parameter combinations for each system on CL can be seen in Figure 2.2.

Annotations from CL in CRAFT are heavily weighted towards the root node, "CL:0000000 - cell"; it is annotated over 2,500 times and makes up ∼44% of all annotations. To test whether annotations of "cell" introduced a bias, all annotations of CL:0000000 were removed and re-evaluated. (Results not shown here.) We see a decrease in F-measure of 0.08

Table 2.5: Best performance for each ontology-system pair. Maximum F-measure for each system on each ontology. Bolded systems produced the highest F-measure.

Cell Type Ontology (CL)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.32	0.76	0.20	1169	379	4591
MetaMap	0.69	0.61	0.80	4590	3010	1170
ConceptMapper	0.83	0.88	0.78	4478	592	1282

Gene Ontology - Cellular Component (GO_CC)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.40	0.75	0.27	2287	779	6067
MetaMap	0.70	0.67	0.73	6111	2969	2341
ConceptMapper	0.77	0.92	0.66	5532	452	2822

Gene Ontology - Molecular Function (GO_MF)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.08	0.47	0.04	173	195	4007
MetaMap	0.09	0.09	0.09	393	3846	3787
ConceptMapper	0.14	0.44	0.08	337	425	3834

Gene Ontology - Biological Process (GO_BP)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.25	0.70	0.15	2592	1120	14321
MetaMap	0.42	0.53	0.34	5802	4994	11111
ConceptMapper	0.36	0.46	0.29	4909	5710	12004

Sequence Ontology (SO)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.44	0.63	0.33	7056	4094	14231
MetaMap	0.50	0.47	0.54	11402	12634	9885
ConceptMapper	0.56	0.56	0.57	12059	9560	9228

ChEBI						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.56	0.7	0.46	3782	1595	4355
MetaMap	0.42	0.36	0.50	4424	8689	3717
ConceptMapper	0.56	0.55	0.56	4583	3687	3554

NCBI Taxonomy						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.04	0.16	0.02	157	807	7292
MetaMap	0.45	0.31	0.88	6587	14954	862
ConceptMapper	0.69	0.61	0.79	5857	3793	1592

Protein Ontology (PRO)						
System	**F-measure**	**Precision**	**Recall**	**# TP**	**# FP**	**# FN**
NCBO Annotator	0.50	0.49	0.51	7958	8288	7636
MetaMap	0.36	0.39	0.34	5255	8307	10339
ConceptMapper	0.57	0.57	0.57	8843	6620	6751

for all systems and are able to identify similar trends in the effect of parameters when “cell” is not included. We can conclude that “cell” annotations do not introduce any bias.

Precision on CL is good overall, the highest being CM (0.88) and the lowest being MM (0.60), with NCBO Annotator in the middle (0.76). Most of the FPs found are due to partial term matching. “CL:0000000 - cell” makes up more than 50% of total FPs because it is contained in many terms and is mistakenly annotated when a more specific term cannot be found. Besides “cell”, terms recognized that are less specific than the gold standard are “CL:0000066 - epithelial cell” instead of “CL:0000082 - lung epithelial cell” and “CL:0000081

- blood cell” instead of “CL:0000232 - red blood cell”. MM finds more FPs than the other systems, many of these due to abbreviations. For example, MM incorrectly annotates the span “ES cells” with “CL:0000352 - epiblast cell” and “CL:0000034: stem cell”. By utilizing abbreviations, MM correctly annotates “NCC” with “CL:0000333 - neural crest cell”, which the other two systems do not find.

Recall for CM and MM are over 0.8 while NCBO Annotator is 0.2. The low recall seen from NCBO Annotator is due to the fact that it is unable to recognize plurals of terms unless they are explicitly stated in the ontology; it correctly finds “melanocyte” but does not recognize “melanocytes”, for example. Because CL is small and its terms are quite simple, there are only two main categories of terms missed: missing synonyms and conjunctions. The biggest category is insufficient synonyms. We find “cone” and “cone photoreceptor” annotated with “CL:0000573 - retinal cone cell” and “photoreceptor(s)” annotated with “CL:0000210 - photoreceptor cell”; these two examples make up 33% (400 out of 1,200) of annotations missed by all systems. No systems found any annotations that contained conjunctions. For example, for the text span “retinal bipolar, ganglion, and rod cells”, three cell types are annotated in CRAFT: “CL:0000748 - retinal bipolar neuron”, “CL:0000740 - retinal ganglion cell”, and “CL:0000604 - retinal rod cell”.

2.4.2.1 NCBO Annotator parameters

Two parameters were found to be statistically significant: *wholeWordsOnly* (p=2.24×10^{-6}) and *minTermSize* (p=9.68×10^{-15}). By using *wholeWordsOnly* = *YES* precision is increased 0.6-0.7 with no change in recall; because of low recall, F-measure is only increased by 0-0.1. Allowing matching to non-whole words finds ~100 more correct annotations but also ~7500 more FPs. Correct annotations found are mostly due to hyphenated spans: “one-neuron” is correctly annotated with “CL:0000540: neuron” and “fibroblast-like” with “CL:0000057 - fibroblast”. About half of the FPs found are due to plurals, “cell” is found within “cells” and “cellular”, while “neuron” is found within “neuronal” and “neurons”. There are FPs because the beginning and end of the text span do not match exactly. Synonyms can be misleading when mixed with matching non-whole words. “CL:0000502 - type D enteroendocrine cell” has a synonym “D cell”, which is found in the following spans “shaped cells”, “elongated cells”, “Disorganized cellular”. Correct terms found do not outweigh the errors introduced, so it is most effective to restrict matching to only whole words.

There is no difference between filtering terms of size *ONE* or *THREE* characters. When filtering out terms less than *FIVE* characters, recall drops 0.15 and F-measure decreases 0.1-0.2. Since “cell” is less than five characters and makes up a large proportion of the annotations, it is understandable why a large decrease in performance is seen. It is best to only filter smaller terms, less than *ONE* or *THREE* characters.

2.4.2.2 MetaMap parameters

Three MM parameters were found to be statistically significant: *scoreFilter* ($p=2.2 \times 10^{-16}$), *minTermSize* ($p=1.1\times10^{-9}$), and *gaps* ($p=8.9\times10^{-8}$). *scoreFilter* and *minTermSize* act as a filter on terms and do not effect the way matching is performed. The best F-measures on CL do not filter any terms, so a score of 0 is suggested. For the same reasons as the NCBO Annotator parameter *minTermSize* seen above, *ONE* and *THREE* are best for filtering term length.

The *gaps* parameter allows skipping tokens to facilitate matching. Allowing gaps increases R <0.05 and decreases P 0-0.2, with a loss of 0.1 in F. Allowing gaps proved useful, and found ∼200 more correct terms. We found words of some term names were not needed to fully express meaning: “apoptotic cell(s)” is correctly annotated with “CL:0000445 - apoptosis fated cell”. Allowing gaps also introduces ∼1200 more incorrect annotations. Terms that are less specific than the annotations seen in CRAFT are incorrectly recognized, such as: “hair cell(s)” being annotated with “CL:0000346 - hair follicle dermal papilla cell”. Mixing abbreviations and gaps produced an interesting incorrect result: “ES cell(s)” is now annotated with “CL:0000715 - embryonic crystal cell”. Due to such introduced errors, although allowing gaps found an increased number of correct terms, the overall F-measure decreased.

2.4.2.3 ConceptMapper parameters

Four CM parameters were found to be statistically significant: *stemmer* ($p=2.2\times10^{-16}$), *orderIndependentLookup* ($p=3.26 \times 10^{-7}$), *searchStrategy* ($p=2.2 \times 10^{-16}$), and *synonyms*

($p{=}5.24{\times}10^{-3}$). It is not conclusive as to which stemmer to use (Porter or BioLemmatizer), but the fact is clear that a stemmer should be used. Not using a stemmer decreases recall by 0.6 with no change in precision. Using a stemmer allows “neuron” to be converted to “neuronal”, “apoptosis” to be found as “apoptotic”, and most importantly allows plurals to be found. Without a stemmer, performance is very similar to the NCBO Annotator.

The parameter *searchStrategy* controls the way possible matches are found in the dictionary. *CONTIGUOUS* produces highest performance, with an increase in R of 0.1-0.3 and an increase in P of 0.2-0.4 over the other values. Allowing CM to ignore tokens converts correct annotations into incorrect annotations because the span length increases to encompass more tokens. Using *CONTIGUOUS* matching, the span “hair cells” is correctly annotated with “CL:0000855 - sensory hair cell” but when *SKIP ANY MATCH* is used we see a completely different annotation, “cochlear immature hair cells” is incorrectly annotated with “CL:0000202 - auditory hair cell”. It should be noted that this is considered to be an incorrect annotation in our analysis. By employing a comparator that takes into account the hierarchy of the ontology, this could be given partial credit since “sensory hair cell” is a parent of “auditory hair cell” (Verspoor *et al* 2006) We can also get complete nonsense when mixing *synonyms* and *searchStrategy*: “CL:0000570 - parafollicular cell” has a synonym “C cell” which gets annotated to “C) Positive control showing granule cells” and masks the correct annotation of “CL:0000120 - granule cell”.

When using *EXACT synonyms* over *ALL* we see an increase in P of 0.1-0.15 with no change in R. Using *ALL synonyms*, ~400 more incorrect annotations are produced; broad synonyms are the cause. “CL:0000562 - nucleate erythrocyte” and “CL:0000595 - enucleate erythrocyte” both have broad synonyms of “red blood cell” and “RBC”, which are found many times. Also, “CL:0000560 - band form neutrophil” has a broad synonym “band” which is not specific; it could be talking about a band on a gel or referring to muscle. Because CL doesn’t contain many synonyms, the meaning is contained well enough to only use *EXACT synonyms*.

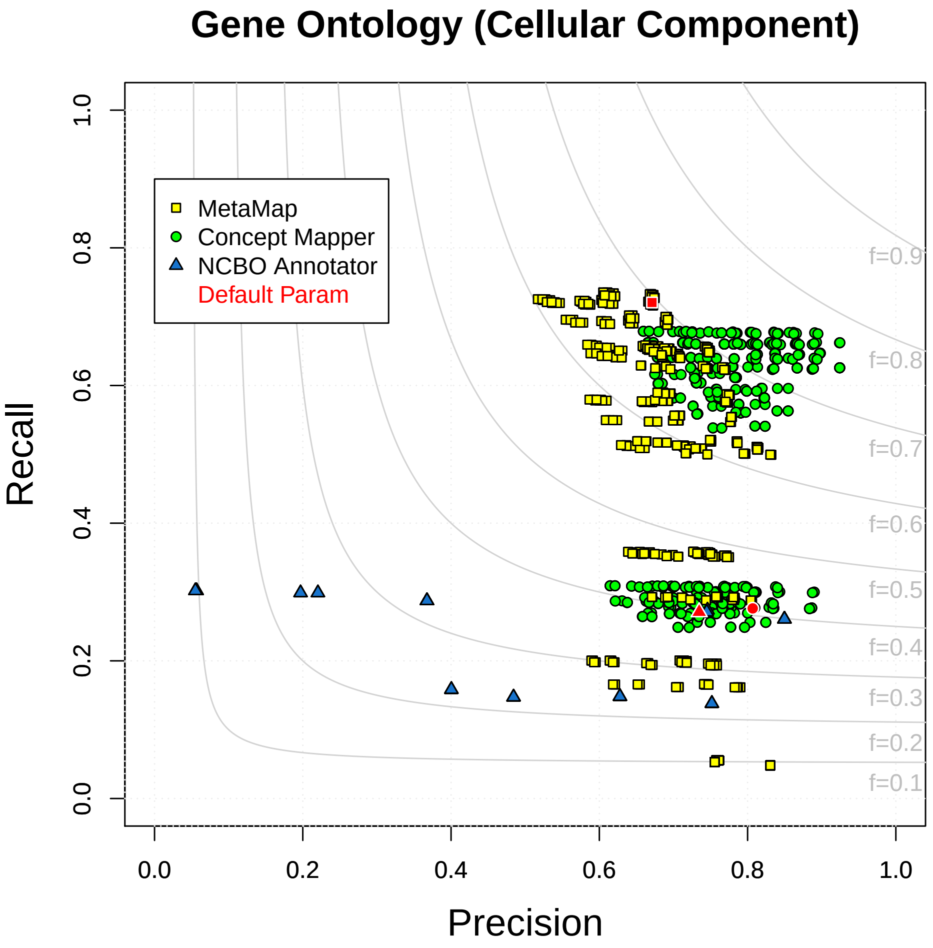


Figure 2.3: All parameter combinations for GO_CC. The distribution of all parameter combinations for each system on GO_CC. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

2.4.3 Gene Ontology - Cellular Component

The cellular component branch of the Gene Ontology describes locations at the levels of subcellular structures and macromolecular complexes. It is useful for annotating where gene products have been found to be localized. GO_CC is similar to CL in that it is a smaller ontology and contains very few synonyms, but the terms are slightly longer and more complex than CL (Table 2.1). Performance from CM (F=0.77) is the best, with MM (F=0.70) second, and NCBO Annotator (F=0.40) third (Table 2.5). All parameter combinations for each system on GO_CC can be seen in Figure 2.3.

Just as in CL, there are many annotations to “GO:0005623 - cell”, 3,647 or 44% of all 8,354 annotations. We removed annotations of “cell” and saw a decrease in performance. Unlike CL, removal of these annotations does not affect all systems consistently. CM sees a large decrease in F-measure (0.2), while MM and NCBO Annotator see decreases of 0.08 and 0.02, respectively.

Precision for all parameter combinations of CM and MM are over 0.50, with the highest being CM at 0.92. NCBO Annotator widely varies from < 0.1 to 0.85. Because precision is high, there are very few FPs that are found. The FPs in common by all systems are due to less specific terms being found and ambiguous terms; NCBO Annotator also finds FPs from broad synonyms and MM specific errors are from abbreviations. Most of the common FPs are mentions that are less specific than the gold standard, due to higher-level terms contained within lower-level ones. For instance, “GO:0016020 - membrane” is found instead of a more specific type of membrane such as “vesicle membrane”, “plasma membrane”, or “cellular membrane”. All systems find ∼20 annotations of “GO:0042603 - capsule” when none are seen in CRAFT; this is due to overloaded terms from different biomedical domains. Because NCBO Annotator is a Web service, we have no control over versions of ontologies used, so it used a newer version of the ontology than that which was used to annotate CRAFT and as inputted into CM and MM. ∼42% of NCBO Annotator FPs were because “GO:0019814 - immunoglobulin complex, circulating” has a broad synonym “antibody” added. Because MM generates variants and incorporates synonyms, we see an interesting error produced from MM: “hair(s)” get annotated with “GO:0009289 - pilus”. It is not understandable why MM would assume this because “hair” is not a synonym, but in the GO definition, pilus is described as a “hair-like appendage”.

MM achieves the highest recall of 0.73 with CM slightly lower at 0.66 and NCBO Annotator the lowest (0.27). NCBO Annotator’s inability to recognize plurals and generate variants significantly hurts recall. NCBO Annotator can annotate neither “vesicles” with “GO:0031982 - vesicle” nor “autosomal” with “GO:0030849 - autosome”, which both CM and MM correctly annotate. The largest category of missed annotations represents other ways to refer to terms not in the synonym list. In CRAFT, “complex(es)” is annotated with “GO:0032991 - macromolecular complex”, and “antibody”, “antibodies”, “immune complex", and "immunoglobulin" are all annotated with "GO:0019814 - immunoglobulin complex", but no systems are able to identify these annotations because these synonyms do not exist in the ontology. MM achieves highest recall because it identifies abbreviations that other systems are unable to find. For example, "chr" is correctly annotated with "GO:0005694 - chromosome", "ER" with "GO:0005783 - endoplasmic reticulum", and "ECM" with "GO:0031012 - extracellular matrix".

2.4.3.1 NCBO Annotator parameters

Two parameters were found to be significant: *minTermSize* ($p=5.9 \times 10^{-8}$) and *wholeWordsOnly* ($p=3.7 \times 10^{-9}$). Given that 44% of annotations in CRAFT are "cell", filtering terms less than *FIVE* characters removes many correct annotations. There is no difference between filtering *ONE* and *THREE* length terms. If using the combination of *wholeWordsOnly* = *NO* and *synonyms* = *YES*, which we do not recommend, it is better to filter terms less than *THREE*. "GO:0005783 - endoplasmic reticulum" has a synonym "ER" which is a common combination of letters seen $\sim$32,500 times in CRAFT in words such as "other", "were", "experiments", and "promoter". Using *wholeWordsOnly* = *NO* allows NCBO Annotator to find $\sim$250 more correct annotations (e.g. "membrane-bound" is correctly annotated with "GO:0016020 - membrane" and "collagen-related" with "GO:0005581 - collagen"), but it also finds $\sim$41,000 more false positives (including erroneous annotations of "er"). Examples of incorrect terms found that are not due to synonyms are plurals, "vesicles" incorrectly annotated with "GO:0031982 - vesicle". These are counted as errors because we used a strict comparator, where beginning and end of the text span must match. If a more lenient comparator were used, these types of errors would be considered correct.

2.4.3.2 MetaMap parameters

Three parameters were found to be statistically significant: *acronymAbb* ($p=1.2\times10^{-5}$), *scoreFilter* ($p=2.2\times10^{-16}$), and *minTermSize* ($p=1.4\times10^{-11}$). MM offers multiple ways to compute and use acronyms or abbreviations to help resolve ambiguous terms. We find it best to use the parameter values *DEFAULT* or *UNIQUE*. The other value, *ALL*, uses all acronyms or abbreviations. When using *ALL* instead of *UNIQUE*, we see a decrease in P of 0.05-0.2 and slight decrease in R; $\sim$80 less TPs and $\sim$1,500 more FPs are found by the maximum

F-measure parameter combination. It is unclear why using *ALL* *acronymAbb* finds fewer correct annotations than using only those with *UNIQUE* expansions. The annotations missed appear to have nothing to do with acronyms or abbreviations but actually derivations. Examples of annotations that were missed by using *ALL* instead of *UNIQUE* are "cytoplasmic" annotated with "GO:0005737 - cytoplasm" and "cytoskeletal" annotated with "GO:0005856 - cytoskeleton". Errors introduced by using *ALL* do look like they came from acronyms or abbreviations. For example, "lung(s)", "pulmonary artery", "pulmonary", "pathological", and "pathology" are all incorrectly annotated with "GO:0000407 - pre-autophagosomal structure", which has a synonym "PAS". "PAS" is an abbreviation for "periodic acid-schiff", a staining method commonly used to stain glycoproteins in the lungs, but it is unlikely that MM makes this linked logical jump; it is unclear why these terms get annotated. It is best to use *DEFAULT* or *UNIQUE* for *acronymAbb*.

It is best to not filter out many terms and use a *scoreFilter* of 0 or 600, because R decreases 0.2-0.6 when using a score of 800 or 1000. Just like the NCBO Annotator parameter examined above, filtering terms less than 5 characters removes many correct annotations of "cell"; it is best to filter less than 1 or 3.

2.4.3.3 ConceptMapper parameters

Four parameters were found to be statistically significant: *searchStrategy* (p=2.2×10^{-16}), *stemmer* (p=2.2×10^{-16}), *findAllMatches* (p=4.8×10^{-5}), and *synonyms* (p=1.3×10^{-4}). The *searchStrategy* *CONTIGUOUS* produces the best performance; we see an increase in P of 0.1 over *SKIP ANY ALLOW OVERLAP* and increase in both P of 0.1 and R of 0.05 over *SKIP ANY MATCH*. Using any other *searchStrategy* besides *CONTIGUOUS* allows correct annotations to be masked by inclusion of surrounding tokens. The span "chromatin granules and fusion of membranes" is incorrectly annotated with "GO:0042584 - chromatin granule membrane" when using *SKIP ANY MATCH*, but the underlined sub-span is correctly annotated with "GO:0042583 - chromatin granule" when using *CONTIGUOUS* matching.

It is significantly better to use a stemmer, which results in an increase in P of 0.1-0.2 and R of 0.4. It is not clear which stemmer is better. Since both stemmers have similar performance, we will only discuss one. Using Porter over *NONE* introduces ~3,000 correct

Table 2.6: **Word Length in GO - Biological Process.**

# Words in Term	# CRAFT annotations	% found by CM	% found by MM	% found by NCBO
5	7	14.3	14.3	14.3
4	109	17.4	3.7	9.2
3	317	37.2	33.4	35.0
2	2077	49.0	50.7	43.3
1	13574	27.6	34.2	11.6

annotations while only finding ∼150 errors. Plurals and variants such as "chromosomal" and "chromosomes" are correctly annotated with "GO:0005694 - chromosome" and "cytoplasmic" correctly annotated with "GO:0005737 - cytoplasm". Not all variants generated by stemming are valid, for example, "fibrillate(s)" and "fibrillation" get annotated with "GO:0043205 - fibril". Overall, the majority of variants are helpful.

Creating dictionaries using *ALL synonyms* instead of *EXACT* decreases P 0.05 with no loss of R. Broad synonyms are the source of these errors; "GO:0035003 - subapical complex" has a broad synonym of "SAC" which is seen ∼100 times in PMID 17608565 as an abbreviation for "starburst amacrine cells". "GO:0019013 - viral nucleocapsid" has a broad synonym of "core" which is found numerous times throughout CRAFT not referring to anything viral. Like CL, there are very few synonyms in GO_CC and we can conclude other types of synonyms are not used frequently in text.

2.4.4 Gene Ontology - Biological Process

Terms from GO_BP are complex; they have the longest average length, contain many words, and almost half contain stop words (Table 2.1). The longest annotations from GO_BP in CRAFT contain five tokens. Distribution of annotations broken down by number of words along with performance can be seen in Table 2.6. When dealing with longer and more complex terms, it is unlikely to see them expressed exactly in text as they are seen in the ontology. For these reasons, none of the systems performed very well. The maximum F-measures seen by each system can be seen in Table 2.5. All parameter combinations for each system on GO_BP can be seen in Figure 2.4. Examining mean F-measures for all parameter combinations, there is no difference in performance between CM (F=0.37) and MM (F=0.42), but considering only the top 25% of combinations there is a difference

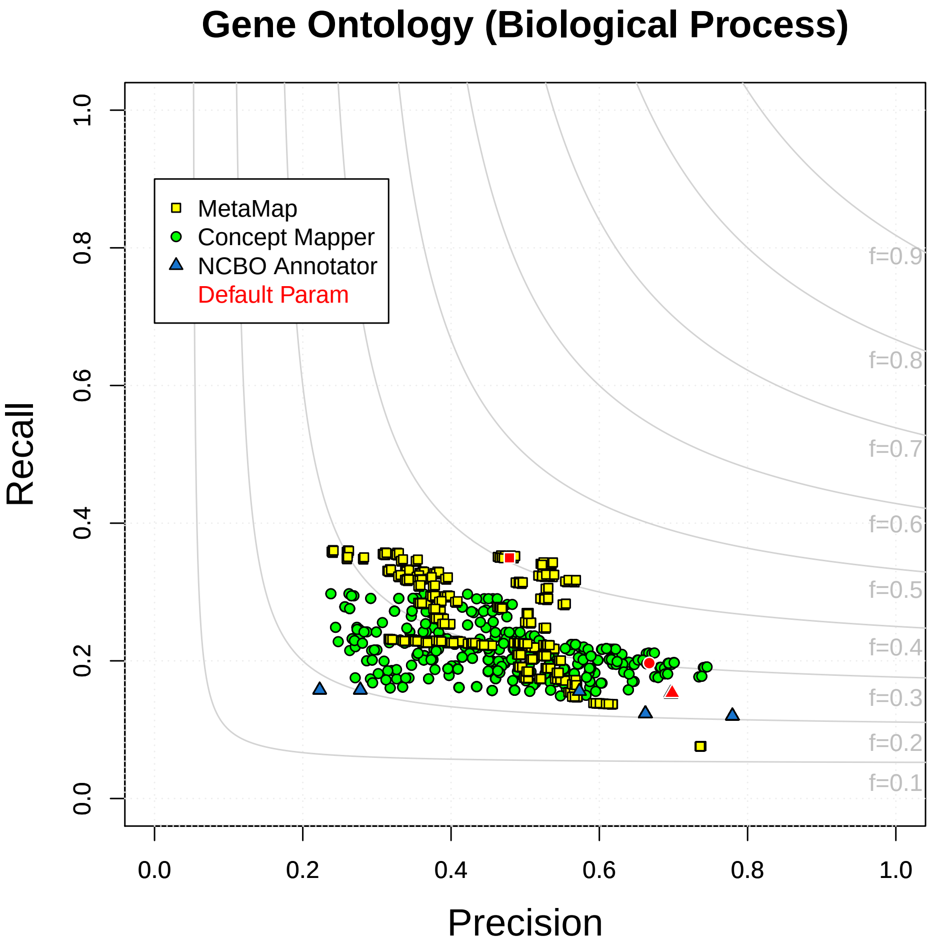


Figure 2.4: All parameter combinations for GO_BP. The distribution of all parameter combinations for each system on GO_BP. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

between the two. A statistical difference exists between NCBO Annotator (F=0.25) and all others, under all comparison conditions.

Performance by all parameter combinations for all systems are grouped tightly along the dimension of recall. Precision for all systems is in the range of 0.2-0.8, with NCBO Annotator situated on the extremes of the range and CM/MM distributed throughout. Common categories of FPs encountered by all three systems are recognizing parts of longer/more specific terms and having different annotation guidelines. As seen in the previous ontologies, high-level terms are seen in lower level terms, which introduces errors in systems that find all matches. For example, we see NCBO Annotator incorrectly annotate "GO:001625 - death" within "cell death", and both CM and MM annotate "development" with "GO:00032502 - developmental process" within the span "limb development". Different annotation guidelines also cause errors to be introduced, e.g. all systems annotate "formation" with "GO:0009058 - biosynthetic process" because it has a synonym "formation", but in CRAFT "formation" may be annotated with "GO:0032502 - developmental process", "GO:0009058 - biosynthetic process", or "GO:0022607 - cellular component assembly", depending on the context. Most of the FPs common to both CM and MM are due to variant generation, for example, CM annotates "region(s)" with "GO:003002 - regionalization" and MM annotates "regular" and "regulator(s)" with "GO:0065007 - biological regulation". Even though we see errors introduced through generating variants, many more correct annotations are produced.

In the grouping of all systems performance, recall lies between 0.1-0.4, which is low in comparison to most all other ontologies. More than ∼7,000 (>50-60%) of the FNs are due to different ways to refer to terms not in the synonym list. The most missed annotation, with over 2,200 mentions, are those of "GO:0010467 - gene expression"; different surface variants seen in text are "expressed", "express", "expressing", and "expression". There are ∼800 discontiguous annotations that no systems are able to find. An example of a discontiguous annotation is seen in the following span: the underlined text from "localization of the Ptdsr protein" gets annotated with "GO:0008104 - protein localization". Many of the annotations in CRAFT cannot be identified using the ontology alone so improvements in recall can be made by analyzing disparities between term name and the way they are expressed in text.

2.4.4.1 NCBO Annotator parameters

Only one parameter was found to be significant, *wholeWordsOnly* ($p=1.33 \times 10^{-7}$). Allowing NCBO Annotator to match non-whole words, only ∼70 more correct annotations found while allowing ∼6000 more incorrect matches, resulting in a decrease in P of 0.1-0.5 with a small increase in R. Correct annotations found are due to hyphenated text, for example, "gene expression" from the span "target-gene expression" and "one-gene expression" are correctly annotated with "GO:0010467 - gene expression". A few FP found are from finding whole terms within other words in the text, e.g. "GO:0007618 - mating" found within "estimating". Using synonyms with matching of non-whole words introduces the majority of errors seen. For instance, “GO:0031028 - septation ignition signaling cascade” has an exact synonym “SIN”, which is found ~2200 times in words such as “using”, “single”, “increasingly”, and “encompassing”. We suggest using *wholeWordsOnly* = *YES* for maximum F-measure and P.

2.4.4.2 MetaMap parameters

Three parameters were found to be significant: *gaps* ($p=1.8 \times 10^{-6}$), *derivationalVariants* ($p=2.8 \times 10^{-10}$), and *scoreFilter* ($p=2.2 \times 10^{-16}$). One way to approximate variation in complex terms is to allow MM to skip tokens to find a match. By allowing gaps, ~75 more TPs are found but the solution isn’t optimal because ~7,500 more FPs are also found; P decreases 0.2-0.3 with a small increase in R. Skipping tokens helps correctly annotate “photoreceptor morphogenesis” with “GO:0008594 - photoreceptor cell morphogenesis” and “meiotic checkpoint” with “GO:0033313 - meiotic cell cycle checkpoint”, but because of the structure of terms in GO_BP we see many more errors. Many terms share similar token patters and by allowing MM to skip tokens many incorrect annotations are produced. For example “regulated process” is incorrectly annotated with 193 different GO terms, such as “GO:0009889 - regulation of biosynthetic process”, “GO:0042053 - regulation of dopamine metabolic process”, and “GO:0045363 - regulation of interleukin-11 biosynthetic process”.

Another way to help find variants of terms in text is to use derivational variants. It is best to generate variants, but there is no significant difference between which type of variants, *ALL* or *ADJ NOUN ONLY*. Generating variants trades precision for recall. When comparing *NONE* to *ADJ NOUN ONLY*, we see an increase in R of 0.05-0.2 along with a decrease in P of 0-0.1. For the best parameter combination, ~2,000 more TPs are found along with ~1,700 more FPs. Not using variants correctly annotates “development” with “GO:0032502 - developmental process” but when adding *ADJ NOUN ONLY* variants, “developmental”, “developmentally”, “developing”, and “develop(s)” are also correctly annotated. Generating variants does not always produce semantically similar terms because ambiguities are introduced. For example, “GO:0007586 - digestion” refers to the process of breaking down nutrients into components that are easily absorbed, but variants such as “digestion(s)”, “digested”, and “digesting” also refer to the process of fragmenting DNA using enzymes. Even though errors are introduced, it is still best to generate variants of terms.

2.4.4.3 ConceptMapper parameters

Four parameters were found to be statistically significant: *searchStrategy* (p=2.2×10^{-16}), *orderIndependentLookup* (p=9.8×10^{-11}), *findAllMatches* (p=4.0×10^{-10}), and *synonyms* (p=2.4×10^{-9}). Like MM, CM also has the ability to approach matching of complex terms through the use of *searchStrategy* and *orderIndependentLookup*. Setting CM's *searchStrategy* = *SKIP ANY MATCH* we see ∼10 more correct annotations found while allowing ∼3,000 more incorrect ones to be found. This can be seen when CM correctly annotates "DNA damage repair" with "GO:0006821 - DNA repair" but also incorrectly annotates the long span "photoreceptors were significantly altered in the their expression level in the Crx-/- mouse, there are many candidates that could be important for photoreceptor morphogenesis" with "GO:'0046531 - photoreceptor cell development". It is interesting to note that the correct and incorrect annotations found by changing MM's *gaps* parameter are not seen when making the similar change in CM and vice versa; even though the parameter should have the same effect on matching, the same annotations are not produced because the systems have different underlying methods.

Besides skipping tokens, another way to approach complex terms is to allow token reordering. Allowing CM to reorder tokens decreases P 0-0.1 with varying small impact on R. In the case of the maximum F-measure parameter combination, varying token order only allows 1 more TP to be found but ∼200 more FPs. Word reordering only helped to find "endoplasmic reticulum protein retention" annotated with "GO:0006621 - protein retention in ER lumen". Finding that single term also introduces errors such as "activated cell(s)" incorrectly annotated with "GO:0001775 - cell activation" and "of apoptosis induction" incorrectly annotated with "GO:0006917 - induction of apoptosis". The benefits of finding the single correct term do not outweigh the errors also introduced; it is best to not allow reordering of tokens for GO_BP.

Stemming is useful for accounting for variations between terms in the ontology and their morphological variations see in text. Using Porter instead of BioLemmatizer or *NONE*, precision is traded for recall, but a higher F-measure is produced. Comparing Porter to *NONE*, ∼1,300 more TPs are found, but also ∼3,500 more FPs are found. CM with Porter, for example, correctly annotates "regulate", "regulating", and "regulated" with "GO:0065007 - biological regulation" and "transcriptional" with "GO:0006351 - transcription, DNA-dependent". Some of the incorrect annotations seen are "transcript(s)" annotated with "GO:0006351 - transcription, DNA-dependent" and "signal(s)" annotated with "GO:0007165 - signal transduction". It is interesting to see that for the single ontology term, "transcription DNA-dependent", both TPs and FPs can be generated by changing the endings.

2.4.5 Gene Ontology - Molecular Function

The molecular function branch of the Gene Ontology describes molecular-level functionalities that gene products possess. It is useful in the protein function prediction field and serves as the standard way to describe functions of gene products. Like GO_BP, terms from GO_MF are complex, long, and contain numerous words with 52.8% containing punctuation and 26.6% containing numerals (Table 2.1). All parameter combinations for each system on GO_MF can be seen in Figure 2.5. Performance on GO_MF is poor; the highest F-measure seen is 0.14. Besides terms being complex, another nuance of GO_MF that makes their recognition in text difficult is the fact that nearly all terms, with the primary exception of binding terms, end in "activity". This was done to differentiate the activity of a gene product from the gene product itself, for example, "nuclease activity" versus "nuclease". However, the large majority of GO_MF annotations of terms other than those denoting binding are of mentions of gene products rather than their activities.

A majority of true positives found by all systems (>70%) are binding terms such as "GO:0005488 - binding", "GO:0003677 - DNA binding", and "GO:0036094 - small molecule binding". These terms are the easiest to find because they are short and do not end in "activity". NCBO Annotator only finds binding terms while CM and MM are able to identify other types. CM identifies exact synonym matches; in particular, "FGFR" is correctly annotated with "GO:0005007 - fibroblast growth factor-activated receptor activity", which has an exact synonym "FGFR". MM correctly annotates "taste receptor" with "GO:0008527

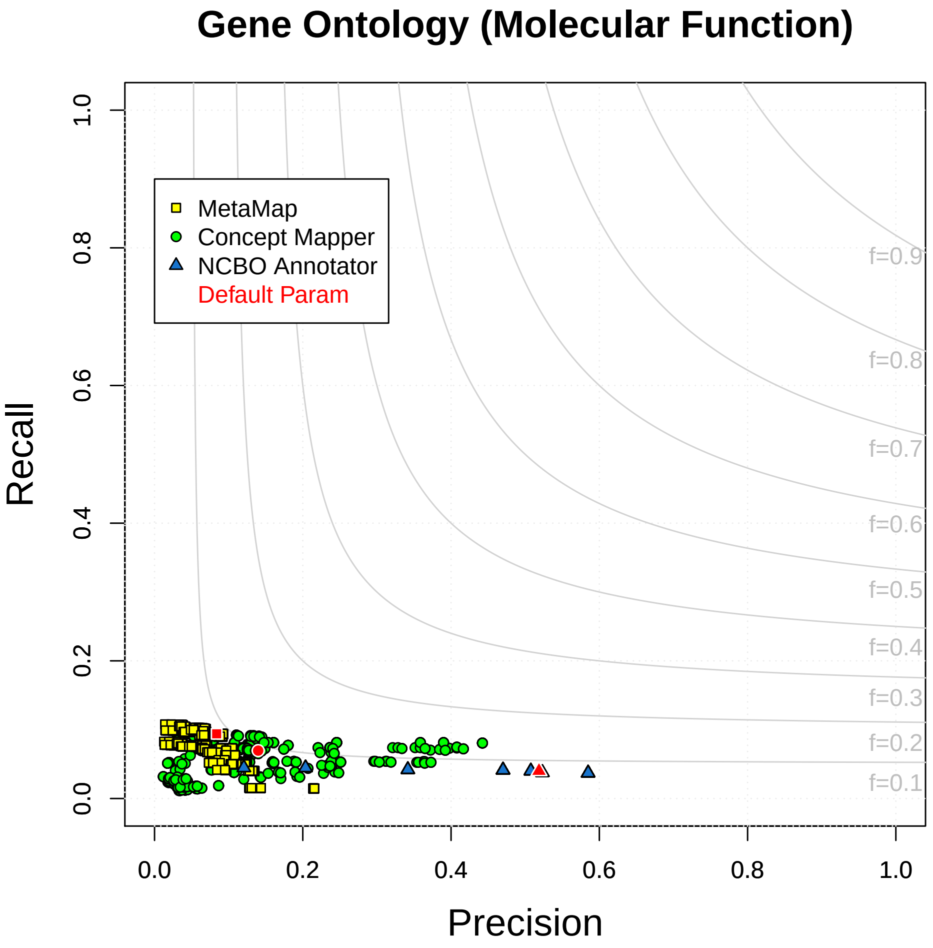


Figure 2.5: All parameter combinations for GO_MF. The distribution of all parameter combinations for each system on GO_MF. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

- taste receptor activity". These annotations are correctly found because the terms have synonyms that refer to the gene products as well as the activity. The only category of FPs seen between all systems is nested or less specific matches, but there are system-specific errors: NCBO Annotator finds activity terms that are incorrect, while MM finds many errors pertaining to synonyms. Example of incorrect nested annotations found by all systems are "GO:0005488 - binding" annotated within "transcription factor binding" and "GO:0016788 - esterase activity" within "acetylcholine esterase". Because the CRAFT annotation guidelines purposely never included the term "activity", some instances of annotating activity along with the preceding word is incorrect; for example, NCBO Annotator incorrectly annotates the span “recombinase activity” with “GO:0000150 - recombinase activity”. FPs seen only by MM are due to broad, narrow, and related synonyms. We see MM incorrectly annotate “neurotrophin” with “GO:0005165 - neurotrophin receptor binding” and “GO:0005163 nerve growth factor receptor binding” because both terms have “neurotrophin” as a narrow synonym.

Recall for GO_MF is low; at best only 10% of total annotations are found. Most of the annotations missed can be classified into three categories: activity terms, insufficient synonyms, and abbreviations. The category of activity terms is an overarching group that contains almost all of the annotations missed; we show performance can be improved significantly by ignoring the word activity in the next section. Terms that fall into the category of insufficient synonyms (∼30% of all terms not found) are not only missed because they are seen without “activity”. For instance, “hybridization(s)”, “hybridized”, “hybridizing”, and “annealing” in CRAFT are annotated with both “GO:0033592 - RNA strand annealing activity” and “GO:0000739 - DNA strand annealing activity”. These mentions are annotated as such because it is sometimes difficult to determine if the text is referring to DNA and/or RNA hybridization/annealing; thus, to simplify the task, these mentions are annotated with both terms, indicating ambiguity. Another example of insufficient synonyms is the inability of all systems to recognize “K+ channel” as “GO:00005267 - potassium channel activity”, due to the fact that the former is not listed as a synonym of the latter in the ontology. A smaller category of terms missed are those due to abbreviations, some of which are mentioned earlier in the work. For instance, in CRAFT, “Dhcr7” is annotated with “GO:0047598 - 7-dehydrocholesterol reductase activity” and “neo” is annotated with “GO:0008910 - kanamycin kinase activity”. Overall, there is much room for improvement in recall for GO_MF; ignoring “activity” at the end of terms during matching alone leads to an increase in R of 0.3.

2.4.5.1 NCBO Annotator parameters

The only parameter found to be statistically significant is *wholeWordsOnly* ($p=1.82 \times 10^{-6}$). Since most of the correct annotations found are related to binding, one can imagine that allowing to match non-whole words leads to many incorrect instances of “GO:0005488

- binding" being found. When allowing matching to non-whole words, precision decreases 0.1-0.4. Even though we see a decrease in P, F-measure is only decreased 0.01 because R is so low. $\sim$20 more TPs are found within hyphenated text, e.g. "substrate-binding" is correctly annotated with "GO:0005488 - binding". But not all hyphenated nested terms are correct. $\sim$70 more errors are also introduced; for instance, "phospholipid-binding" is incorrectly annotated with "GO:005488 - binding". We also see full terms within other words, "GO:0003774 - motor activity" is incorrectly annotated within "locomotor activity". It is difficult to provide suggestions because the highest mean F-measures, 0.075, are obtained by using *wholeWordsOnly* = NO, but using *wholeWordsOnly* = YES produces a mean F-measure of 0.070. There is a statistically significant difference between the two, but practically speaking they are both poor.

2.4.5.2 MetaMap parameters

Four parameters were found to be statistically significant: *gaps* ($p=2.2 \times 10^{-16}$), *acronymAbb* ($p=5.2 \times 10^{-5}$), *scoreFilter* ($p=2.2 \times 10^{-16}$), and *minTermSize* ($p=1.6 \times 10^{-9}$). Even though these four parameters produce statistically significant mean F-measures, it is difficult to analyze them because for most parameter combinations P, R, and F are all less than 0.1. The *gaps* parameter shows the biggest difference between F-measure in parameter values, 0.02. Allowing gaps introduces $\sim$10 more correct annotations along with $\sim$4,000 more incorrect ones. Of the few correct annotations found by allowing gaps, one example is, "Ran-binding" correctly annotated with "GO:0008536 - Ran GTPase binding". The errors introduced from allowing gaps are due to similarities in terms in the ontology. For instance, "D activity" is incorrectly annotated with 170 different GO terms, such as, "GO:0047816 - D-arabinose 1-dehydrogenase activity" and "GO:00428880 - D-glucuronate transmembrane transporter activity". For best performance, gaps should not be allowed.

scoreFilter and *minTermSize* are filters on the returned annotations and do not affect the way matching is performed. The maximum F-measures are seen when *scoreFilter* is set to 0 or 600 and *minTermSize* is set to 1 or 3. These parameter settings return most of the annotations found by MM.

2.4.5.3 ConceptMapper parameters

Four parameters were found to be statistically significant: *searchStrategy* (p=2.2×10^{-16}), *stopWords* (p=5.8×10^{-13}), *findAllMatches* (p=2.2×10^{-16}), and *synonyms* (p=4.3×10^{-16}). Using *CONTIGUOUS searchStrategy* produces the highest F-measure; an increase in P of 0.05-0.3 and an increase in R of 0-0.05 is seen when comparing to other values. Allowing CM to skip tokens when looking terms up converts TPs to FPs because more tokens are included. For example, using *CONTIGUOUS*, "GO:0005488 - binding" is correctly annotated in the span "proteins that bind", but when using *SKIP ANY MATCH*, the same span is incorrectly annotated with "GO:0005515 - protein binding". We see an interaction between *searchStrategy* and *findAllMatches*. When using a value of *searchStrategy* that allows gaps along with *findAllMatches* = *YES*, recall is increased and a higher F-measure is seen.

It is recommended to not remove stop words; both P and R are decreased when removing PubMed stop words. ~15 TPs found when not removing stop words are missed when removing PubMed stop words because more specific annotations can be made by ignoring a common word. For example, when stop words are not removed "bind" is correctly annotated with "GO:0005488 - binding", but when removing PubMed stop words, other binding annotations are produced from the same span, such as, "proteins that bind" is incorrectly annotated with "GO:0005515 - protein binding" and "receptors, which bind" is incorrectly annotated with "GO:0005102 - receptor binding". Besides the missed annotations seen above, ~1,000 more errors are introduced. Most of these errors are from synonyms with the *caseMatch* parameter set to *IGNORE* or *INSENSITIVE*. For instance, "bind(s)", "binding", "bound" are incorrectly annotated with "GO:0003680 - AT DNA binding", which has exact synonym "AT binding", which contains a stop word. Along the same lines, "activity" is incorrectly annotated with "GO:0050501 - hyaluronan synthase activity", which has a broad synonym "HAS activity", where "has" is a stop word.

Creating dictionaries with *ALL synonyms* introduces ~100 more TPs and ~11,000 more FP. Using narrow synonyms helps to correctly identify "ryanodine receptor" with "GO:0005219 - ryanodine-sensitive calcium-release channel activity". But overall, using *ALL synonyms* hurts performance. Related synonyms for some terms are common words. For example, "GO:0004066 - asparagine synthesis activity" has a related synonym "as",

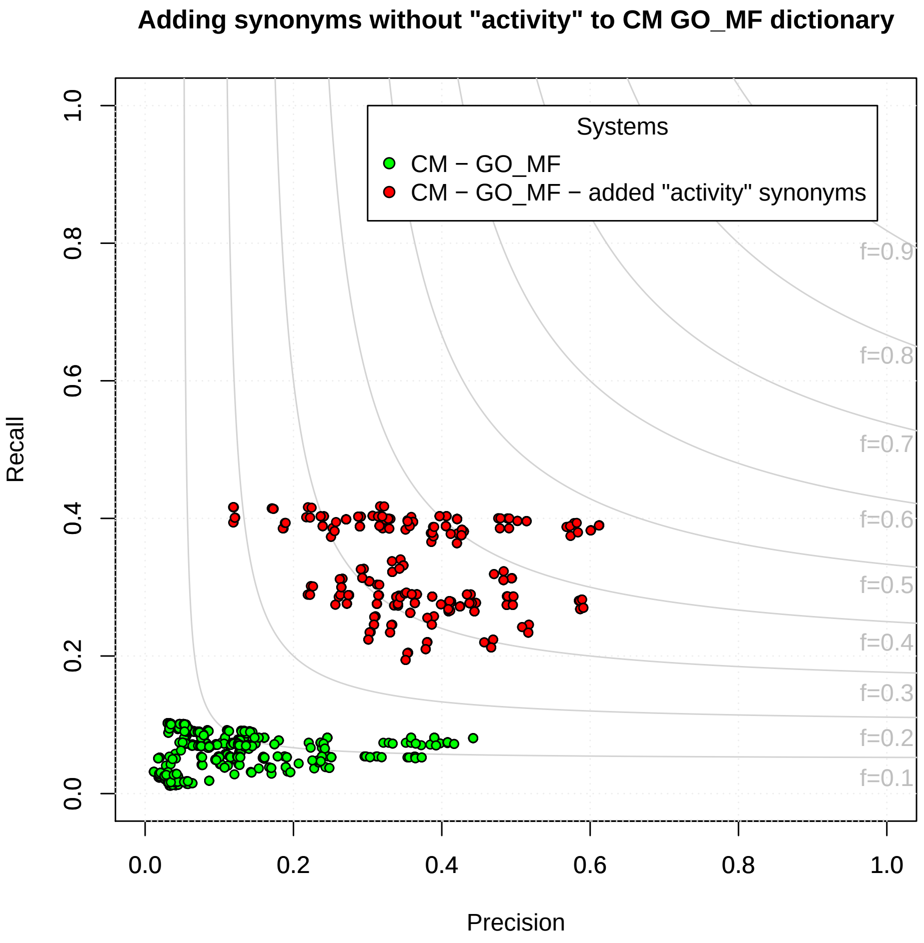


Figure 2.6: Improvement seen by CM on GO_MF by adding synonyms to the dictionary. By adding synonyms of terms without "activity" to the GO_MF dictionary precision and recall are increased.

which is found more than 2,000 times in CRAFT. We also see many interesting errors introduced when mixing a stemmer and all synonyms. "GO:0043807 - 3-methyl-2-oxobutanoate dehydrogenase (ferredoxin) activity" has a related synonym "VOR", which when run though BioLemmatizer produces the lemma "for" and is found over 4,000 times in CRAFT. We suggest using *EXACT* synonyms.

2.4.5.4 Improving performance on GO_MF

As suggested in previous work on the GO, since the word "activity" is present in most terms, its information content is very low (Verspoor et al., 2005). Also, when adding "activity" to the end of the top 20 most common other words in GO_MF terms (as seen in (McCray et al., 2002)), over half are terms themselves (Ogren et al., 2004). An experiment was performed to evaluate the impact of removing "activity" from all terms in GO_MF. For each term with "activity" in the name, a synonym was added to the ontology obo file with the token "activity" removed; for example, for "GO:0004872 - receptor activity", a synonym of "receptor" was added. We tested this only with CM; the same evaluation pipeline was run but the new obo file used to create the dictionary. Using the new dictionary, F-measure is increased from 0.14 to 0.48 and a maximum recall of 0.42 is seen (Figure 2.6). These synonyms should not be added to the official ontology because it contradicts the specific guidelines the GO curators established (Verspoor et al., 2009), but should be added to dictionaries provided as input to concept recognition systems.

2.4.6 Sequence Ontology

The Sequence Ontology describes features and attributes of biological sequences. The SO is one of the smaller ontologies evaluated, ∼1,600 terms, but contains the highest number of annotations in CRAFT, ∼23,500. ∼92% of SO terms contain punctuation, which is due to the fact that the words of the primary labels are demarcated not by spaces but by underscores. Many, but not all, of the terms have an exact synonym identical to the official name, but with spaces instead of underscores. CM is the top performer (F=0.56) with MM middle (F=0.50) and NCBO Annotator at the bottom (F=0.44). Statistically, looking at all parameter combinations mean F-measures, there is a difference between CM and the rest, while a difference cannot be determined between MM and NCBO Annotator. When looking at the top 25% of combinations, a difference can be seen between all three systems. All parameter combinations for each system on SO can be seen in Figure 2.7.

Most of the FPs can be grouped into four main categories: contextual dependence of SO, partial term matching, broad synonyms, and variants generated. In all three systems, we see the first three types, but errors from variants are specific to CM and MM. The

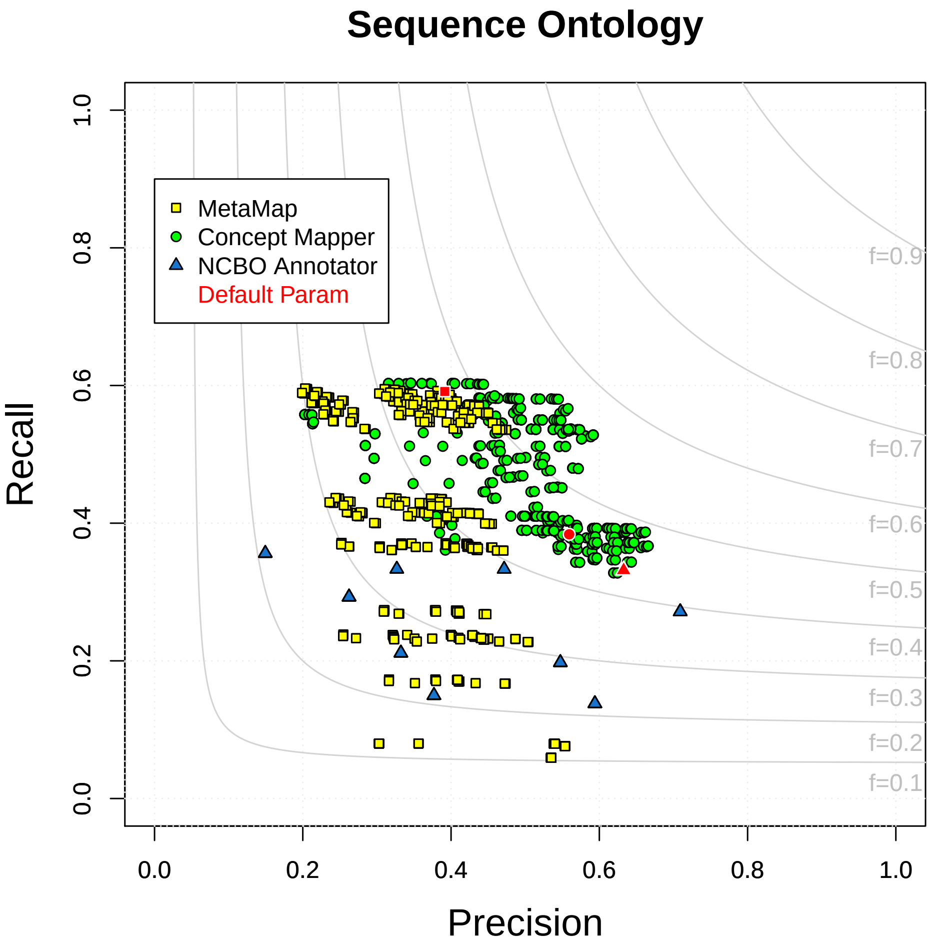


Figure 2.7: All parameter combinations for SO. The distribution of all parameter combinations for each system on SO. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

SO is sequence specific, meaning that terms are to be understood in relation to biological sequences. When the ontology is separated from the domain, terms can become ambiguous. For example, “SO:0000984 - single” and “SO:0000985 - double” refer to the number of strands in a sequence, but can also be used in other contexts, obviously. Synonyms can also become ambiguous when taken out of context. For example, “SO:1000029 - chromosomal_deletion” has a synonym “deficiency”. In the biomedical literature, “deficiency” is commonly used when discussing lack of a protein, but as a synonym of “chromosomal_deletion” it refers to a deletion at the end of a chromosome; these are not semantically incorrect, but incorrect in terms of CRAFT concept annotation guidelines. Because of the hierarchical relationships in the ontology we find the high level term "SO:0000001 - region" within other terms; when the more specific terms are unable to be recognized, "region" can still be recognized. For instance, we find "region" incorrectly annotated inside the span "coding region", when in the gold standard the span is annotated with "SO:0000851 - CDS_region". Besides being ambiguous, synonyms can also be too broad. For instance, "SO:0001091 - non_covalent_binding_site" and "SO:0100018 - polypeptide_binding_motif" both have a synonym of "binding"; as seen in GO_MF above, there are many annotations of binding in CRAFT. The last category of errors are only seen in CM and MM because they are able to generate variants. Examples of erroneous variants are MM incorrectly annotating "based", "foundation", and "fundamental" with "SO:0001236 - base" and CM incorrectly annotating "probing" and "probed" with "SO:0000051 - probe".

Recall on SO is close between CM (0.57) and MM (0.54), while recall for NCBO Annotator is 0.33. The ∼5,000 annotations found by both CM and MM that are missed by NCBO Annotator are composed of plurals and variants. The three categories that a majority of the FNs fall into are insufficient synonyms, abbreviations, and multi-span annotations. More than half of the FNs are due to insufficient synonyms or other ways to express a term. In CRAFT, "SO:0001059 - sequence_alteration" is annotated to "mutation(s)", "mutant", "alteration(s)", "changes", "modification", and "variation". It may not be the most intuitive annotation, but because of the structure of the SO version used in CRAFT, it is the most specific annotation that can be made for mutating/changing a sequence. Another example of insufficient synonyms can be seen from the annotation of "chromosomal region", "chromosomal loci", "locus on chromosome" and "chromosomal segment" with"SO:0000830 - chromosome_part". These are more intuitive than the previous example; if different "parts" of a chromosome are explicitly enumerated the ability to find them increases. Abbreviations or symbols are another category missed. For example, "SO:0000817 - wild_type" can be expressed as "WT" or "+" and "SO:0000028 - base_pair" is commonly seen as "bp". These abbreviations are more commonly seen in biomedical text than the longer terms are. There are also some multi-span annotations that no systems are able to find; for example, "homologous human MCOLN1 region" is annotated with "SO:0000853 - homologous_region".

2.4.6.1 NCBO Annotator parameters

Two parameters were found to be significant: *wholeWordsOnly* ($p=8.6 \times 10^{-11}$) and *minTermSize* ($p=2.5 \times 10^{-5}$). Allowing NCBO Annotator to match non-whole words introduces $\sim$500 more correct annotations, but as a result, $\sim$40,000 more incorrect ones are also found resulting in a decrease in P of 0.2-0.4 with a small decrease in R. Correct annotations found are from hyphenated spans of text. For example, "SO:0001026 - genome" is correctly found within "genome-wide", and "GO:0000704 - gene" is also correctly found within "gene-based" and "receptor-gene". Many errors are introduced given the ability to recognize non-whole words. Smaller terms are found within other words. "SO:0000704 - gene"; for example, is found within "morphogeneic", "general", and "degenerate".

Filtering terms less than *FIVE* characters decreases R by 0.2. This is due to the fact that two commonly found correct annotations will be filtered out, "SO:0000704 - gene" and "SO:0000352 - DNA". For best performance, terms less then length *ONE* or *THREE* should be filtered.

2.4.6.2 MetaMap parameters

Four parameters were found to be different: *model* ($p=1.6 \times 10^{-6}$), *acronymAbb* ($p=2.8 \times 10^{-9}$), *scoreFilter* ($p=2.2 \times 10^{-16}$) and *minTermSize* ($p=1.0 \times 10^{-11}$). The SO is the one of two ontologies where there is a difference between the values of MM's *model* parameter. Using the *RELAXED* model in place of *STRICT*, decreases P 0.1-0.3 with no change in R. We find that $\sim$400 more FP are introduced for the best performing parameter combination when *RELAXED* is used. A majority of the errors are from matching a capital letter at the beginning or end of a token. For example, "HKI" is incorrectly annotated with "SO:0001230 - inosine" and "SO:0001438 - isoleucince", both of which have "I" as a synonym. An error seen that was not due to matching capital letters is "DNA-binding" and "DNA binding" incorrectly annotated with "SO:0000417 - polypeptide_domain", which has a synonym "DNA_bind". We can conclude that it is better to use the *STRICT* model with SO.

There is no difference between using *DEFAULT* and *UNIQUE* values of the *acronymAbb* parameter, but there is a difference when using *ALL*. Comparing *DEFAULT* and *ALL*, P is decreased 0-0.2 with a slight increase in R. ~20 more TPs and ~4,000 more FPs are found when using *ALL* acronyms and abbreviations. One example of a term correctly recognized is due to synonymous adjectives, “short interfering RNA” is correctly annotated with “SO:0000646 - siRNA”, which has a synonym of “small interfering RNA”. Some abbreviations have more than one meaning and are not found unless *ALL* is used, e.g. “PCR product(s)” is correctly annotated with ‘SO:0000006 - PCR_product”. Unfortunately, there are many terms that have possible ambiguous abbreviations. For instance, “protein(s)” is incorrectly annotated with “SO:0001439 - proline”, which has the synonyms “P” and “Pro”; Also, “states” is incorrectly annotated with “SO:0000331 - STS”.

Filtering terms based on length has the same results as with the NCBO Annotator parameter above; filtering out terms less than 5 characters decreases R by 0.1-0.2, so it is best to filter terms less than 1 or 3. Along the same lines, it is best to use all or most of the annotations returned by MM, so setting *scoreFilter* equal to 0 or 600 is suggested.

2.4.6.3 ConceptMapper parameters

Four parameters were found to be significant: *searchStrategy* (p=7.4×10^{-8}), *stemmer* (p=2.2×10^{-16}), *stopWords* (p=1.7×10^{-4}), and *synonyms* (p=2.2×10^{-16}). As seen in many of the other ontologies before, stemming is useful for improving recall. With the SO, there is no difference between Porter or BioLemmatizer, but there is a difference between a stemmer and *NONE*. When using Porter over *NONE*, ~3,800 more TPs are found along with ~5,300 more FPs. Along with variants, such as “genomic” and “genomically” correctly annotated with “SO:0001026 - genome”, using a stemmer allows plurals to be found. Not all variants carry the same meaning as the original term. For instance, “SO:0000141 - terminator” refers to the sequence of DNA at the end of a transcript that causes RNA polymerase to fall off, while “terminal”, “terminally”, and “termination” all carry different meanings. Even though using a stemmer introduces more incorrect annotations than correct ones, F-measure is increased by 0.1-0.2.

Removing PubMed stop words has varying effects. For one group, an increase in P of 0.05-0.2 with no change in R is seen, but for the other one, slight decreases in P and R are seen. The maximum F-measure parameter combination falls in the latter group, for which ∼25 less TPs and ∼200 more FPs are found when using PubMed stop words. The correct annotations found when not using stop words and missed by removing stop words are masked by longer FPs. For instance, not removing stop words, "SO:0000151 - clone" and "SO:0000756 - cDNA" are both correctly annotated in the span "clone in these cDNA", but when removing PubMed stop words the entire span is incorrectly annotated with "SO:0000792 - cloned_cDNA" because "in" and "these" are not considered. Errors introduced are from the 9.3% of terms that contain stop words that are integral to their meaning. For example, "motif" is incorrectly annotated with "SO:0001010 - i_motif". For the best performance, it is best to not remove stop words.

Creating dictionaries with *ALL* instead of only *EXACT synonyms* allows ∼400 more TPs to be found while introducing ∼5,000 more FPs, which leads to a decrease in P of 0.1-0.4 with an increase in R of 0-0.05. Only two examples make up all of the correct annotations found: "domain(s)" correctly annotated with "SO:0000417 - polypeptide_domain" which has broad synonym "domain" and "signal(s)" correctly annotated with both "SO:0000725 - transit_peptide" and "SO:0000418 - signal_peptide", which both have broad synonym "signal"; both of these correct annotations are matches to broad synonyms. Of the errors introduced, over half, ∼2,600, of the incorrect annotations are broad synonyms from the following two examples: "region(s)", "site(s)", "position(s)", and "positional" are incorrectly annotated with "SO:0000839 - polypeptide_region" (has broad synonyms "region", "positional", and "site") and "signal(s)" incorrectly annotated with "SO:0000725 - transit_peptide" and "SO:0000418 - signal_peptide" (has broad synonym "signal"). It is interesting that the same broad synonym, "signal", produces a ∼30 TPs but many more FPs (∼1,300). We can conclude that the correct annotations found do not outweigh the errors introduced, so it is best to create dictionaries with only *EXACT synonyms*.

2.4.7 Protein Ontology

The Protein Ontology (PRO) represents evolutionarily defined proteins and their natural variants. It is important to note that although the PRO technically represents proteins strictly, the terms of the PRO were used to annotate genes, transcripts, and proteins in CRAFT. Terms from PRO contain the most words, have the most synonyms, and ∼75% of

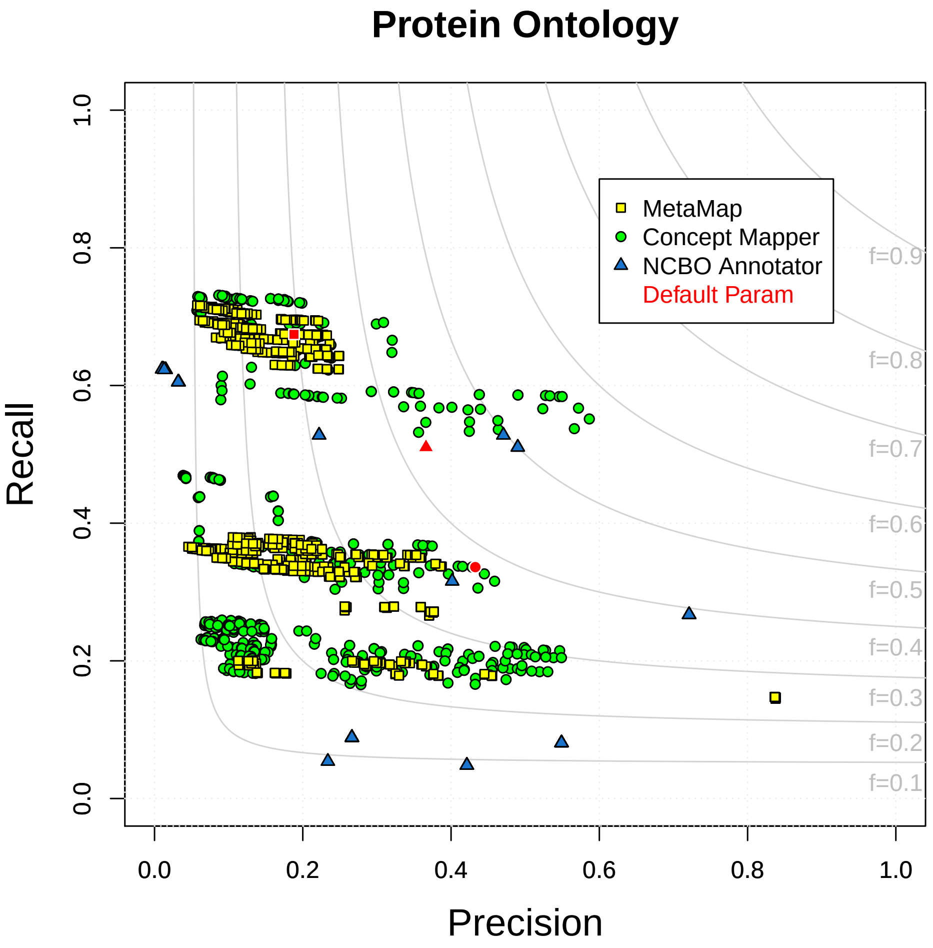


Figure 2.8: All parameter combinations for PRO. The distribution of all parameter combinations for each system on PRO. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

terms contain numerals (Table 2.1). Even though term names are complex, in text, many gene and gene product references are expressed as abbreviations or short names. These references are mostly seen as synonyms in PRO. Recognizing and normalizing gene and gene product mentions is the first step in many natural language processing pipelines and is one of the most fundamental steps. CM produces the highest F-measure (0.57), followed by NCBO Annotator (0.50), and lastly MM (0.35) produces the lowest. All parameter combinations for each system on PRO can be seen in Figure 2.8. Unlike most of the ontologies covered above, stemming terms from PRO does not result in the highest performance.

The best parameter combination for CM does not use any stemmer, which is why NCBO Annotator performs better than MM.

All systems are able to find some references to the long names of genes and gene products, such as “PR:000011459 - neurotrophin-3” and “PR:000004080 - annexin A7”. As stated previously, a majority of the annotations in CRAFT are short names of genes and gene products. For example, the long name of PR:000003573 is “ATP-binding cassette sub-family G member 8”, which is not seen, but the short name “Abcg8” is seen numerous times. The errors introduced by all systems can be grouped into misleading synonyms and different annotation guidelines, while MM also introduces errors from abbreviations and variants. Of errors common to all systems, the largest category is from misleading synonyms (>50% for CM and NCBO Annotator, ∼33% for MM). For example, ∼3,000 incorrect annotations of “PR:000005054 - caspase-14”, which has synonym “MICE”, are seen, along with mentions of the word “male” incorrectly annotated with “PR:000023147 - maltose-binding periplasmic protein”, which has the synonym “malE”. As seen in these errors, capitalization is important when dealing with short names. Differing annotation guidelines also result in matching errors, but because all systems are at the same disadvantage a bias isn’t introduced. The word “protein” is only annotated with the ChEBI ontology term “protein”, but there are many mentions of the word “protein” incorrectly annotated with a high-level term of PRO, “PR:000000001 - protein”. This term was purposely not used to annotate “protein” and “proteins”, as this would have conflicted with the use of the terms of PRO to annotate not only proteins but also genes and transcripts. MM generates abbreviations and acronyms, but they are not always helpful. For example, due to abbreviations, “MTF-1” is incorrectly annotated with “PR:000008562 - histidine triad nucleotide-binding protein 2”; because MM is a black box, it is unclear how or why this abbreviation is generated. Morphological variants of synonyms are also causes of errors. For example, “finding” and “found” are incorrectly annotated because they are variants of “FIND”, which is a synonym of “PR:000016389 - transmembrane 7 superfamily member 4”.

All systems are able to achieve recall of >0.6 on at least one parameter combination, with CM and MM achieving 0.7 by sacrificing precision. When balancing P and R, the maximum R seen is from CM (0.57). Gene and gene product names are difficult to recognize because there is so much variation in the terms — not morphological variation as seen in most other ontologies, but differences in symbols, punctuation, and capitalization. The main categories of missed annotations are due to these differences. Symbols and Greek letters are a problem encountered many times when dealing with gene and gene product names (Yu et al., 2002). These tools offer no translation between symbols so, for example, "TGF-β2" is unable to be annotated with "PR:000000183 - TGF-beta2" by any systems. Along the same lines, capitalization and punctuation are important. The hard part is knowing when and when not to ignore them; any of the FPs seen in the previous paragraph are found because capitalization is ignored. Both capitalization and punctuation must be ignored to correctly annotate the spans "mr-s" and "mrs" with "PR:000014441 - sterile alpha motif domain-containing protein 11", which has a synonym "Mr-s". As seen above, there are many ways to refer to a gene/gene product. In addition, an author can define one by any abbreviation desired and then refer to the protein in that way throughout the rest of the chapter, so attempting to capture all variation in synonyms is a difficult task. In CRAFT, for instance, "snail" refers to "PR:000015308 - zinc finger protein SNAI1" and "moonshine" or "mon" refers to "PR:000016655 - E3 ubiquitin-protein ligase TRIM33".

2.4.7.1 NCBO Annotator parameters

Only *wholeWordsOnly* (p=2.3×10^{-11}) was found to be significant. Matching non-whole words introduces $\sim$1,500 more TPs and $\sim$270,000 more FPs. The TPs that were found contained some kind of punctuation. For example, "BRCA2" from the spans " BRCA2-independent", "RAD51-BRCA2", and "BRCA2^{+}" are correctly annotated with "PR:000004804 - breast cancer type 2 susceptibility protein". Many of the FPs found are from matching smaller synonyms within longer words. An example,"PR:000008207 - synaptic glycoprotein SC2", which has an exact synonym "TER", is incorrectly found $\sim$7,000 times in words such as "determine", "promoter", and "anterior". It is best to not allow NCBO Annotator to match non-whole words.

2.4.7.2 MetaMap parameters

Four parameters were found to be significant: *gaps* (p=2.2×10^{-16}), *acronymAbb* (p=2.2×10^{-16}), *scoreFilter* (p=2.2×10^{-16}), and *minTermSize* (p=7.1×10^{-14}). Inserting gaps when matching decreases P by 0.2 and increases R slightly; varying *gaps* on maximum F-measure parameter combinations finds ∼100 more TPs while introducing ∼12,000 more FPs. *gaps* provide MM the ability to skip tokens to find matches. For example, "cyclin 5" is correctly annotated with "PR:000005258 - cell division protein kinase 5", which has a synonym "cyclin-dependent kinase 5"; the four tokens, "-dependent kinase", are skipped to allow this match. Even though skipping tokens find some TPs, many more errors are found. Some errors are close matches, but less specific terms, such as "alpha-crystallin" incorrectly annotated with "PR:000005908 - alpha-crystallin B chain". Others found are completely wrong; for instance, the span "protein-1" can be annotated with any term as long as it contains "protein", "-", and "1", in that order. "PR:000010001 - protein lyl-1" and "PR:000009230 - multisynthetase complex auxiliary component p38", which has the synonym "protein JTV-1" are examples of terms incorrectly matched with "protein-1". Not using gaps produces the highest F-measure.

The maximum F-measure is obtained by using *scoreFilter* =600 and *minTermSize* =3 or 5. A majority of the terms matched in PR are at least 3 characters long. By filtering some correct annotations will be lost, such as "Rb" annotated with "PR:000013773: retinoblastoma-associated protein", but for the most part it is safe to filter out terms less than 3 characters.

2.4.7.3 ConceptMapper parameters

Four parameters were found to be significant: *caseMatch* (p=6.1×10^{-6}), *stemmer* (p=2.2×10^{-16}), *findAllMatches* (p=1.1×10^{-9}), and *synonyms* (p=5.3×10^{-16}). PR is the only ontology where *caseMatch* is significant. The *caseMatch* value CASE FOLD DIGITS produces the highest F-measure. Only text that contains digits is folded to lower case. ∼2,000 fewer TPs and ∼40,000 fewer FPs are found when comparing folding only digits to folding everything. Some annotations are missed, for example, "MYOC", which is a synonym of "PR:000010873 - myocilin" is not matched with "Myoc". Errors introduced by folding everything are mainly from folding synonyms that are common english terms. For example, “TO” is a synonym of “PR:000016214 - tryptophan 2,3-dioxygenase”. Just from looking at the synonym, it is hard to determine when and when not to fold cases. For maximum F-measure, it is best to only fold those with digits.

Not using any type of stemmer produces the highest precision and F-measure. Using BioLemmatizer increases R by finding ∼2,000 more TPs but decreases P by finding ∼100,000 more FPs. Using a stemmer allows plurals to be found. For example, “neurotrophins” is correctly annotated with “PR:000021998 - neurotrophin”. Also, using a stemmer folds all text to lower cases; for example, “Shh” is correctly annotated with “PR:000014841 - sonic hedgehog protein”, which has a synonym “SHH”. Generating morphological and derivational variants also introduces many other errors. For instance, “PR:000008323 - general transcription factor II-I repeat domain-containing protein 1” has a synonym “BEN”, that when put through BioLemmatizer gets turned into “been”, “is”, “are”, “be”, “being”, “were” that are found incorrectly ∼15,000 times in CRAFT. Folding everything also produces incorrect annotations, such as “SAC” getting incorrectly annotated with “PR:000003752 - adenylate cyclist type 10”, which has a synonym “sAC”. Using a stemmer finds many TPs, but the many FPs introduced outweigh the TPs.

Using *ALL synonyms* produces the highest F-measure; ∼5,000 fewer TPs are found by only using *EXACT synonyms*. Because *ALL* is the best performance it tells us that the synonym list for PR is well maintained and does not contain many spurious synonyms. In addition, many of the related synonyms are synonyms for the corresponding genes of proteins which are equivalently annotated in CRAFT. An example of an annotation missed by using only *EXACT synonyms* is “Car2” correctly annotated with “PR:000004918 - carbonic anhydrase 2”; it has an exact synonym of “Ca2” and a related synonym of “Car2”. It is best to use *ALL synonyms* for PR.

2.4.7.4 Removing FP PRO annotations

In order to show that performance improvements can be made easily, we examined and removed the top five FPs from each system on PRO. The top five errors only affect precision and can be removed without any impact in recall; the impact can be seen in Figure 2.9. A

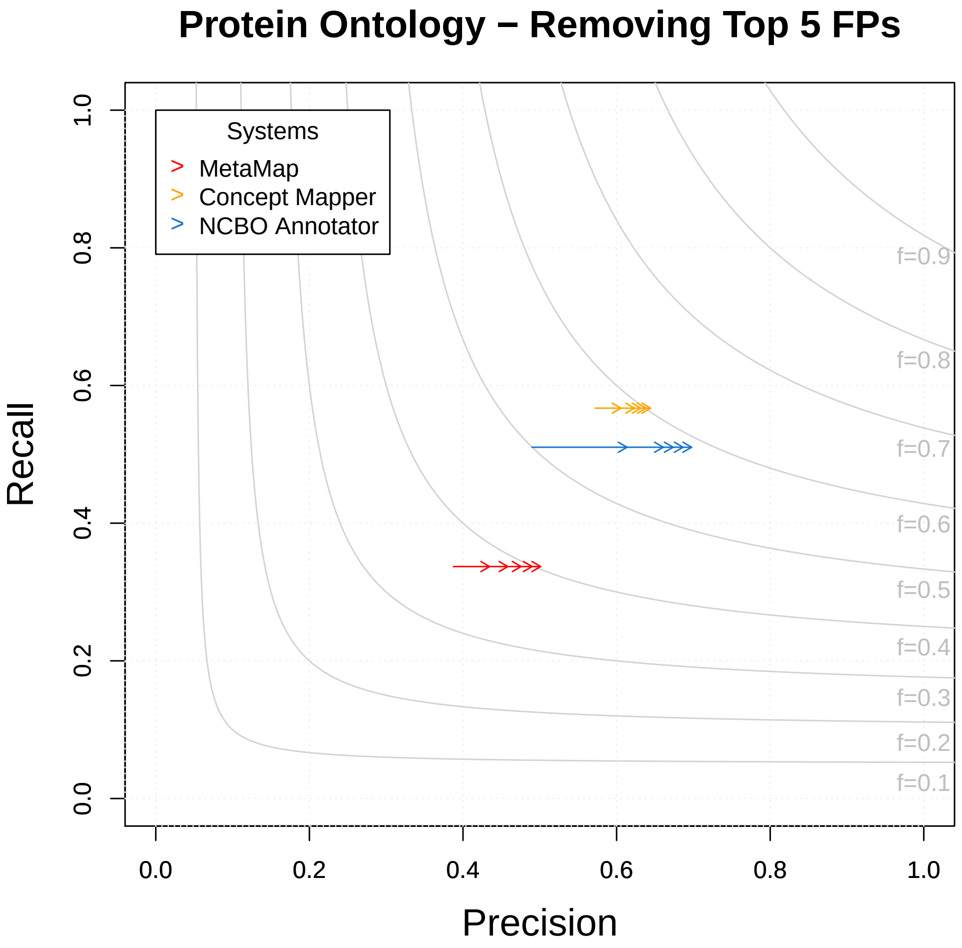


Figure 2.9: Improvement on PRO when top 5 FPs are removed. The top 5 FPs for each system are removed. Arrows show increase in precision when they are removed. No change in recall was seen.

simple process produces a change in F-measure of 0.03-0.09. A common category of FPs removed from all systems are annotations made with "PR:000000001 - protein", as the term was found ∼1,000-3,500 times. Three out of the top five errors common to MM and NCBO Annotator were found because synonym capitalization was ignored. For example, "MICE" is a synonym of "PR:000005054 - caspase-14", "FIND" is a synonym of "PR:000016389 - transmembrane 7 superfamily member 4", and "AGE" is a synonym of "PR:000013884 - N-acylglucosamine 2-epimerase". The second largest error seen in CM is from an ambiguous synonym: "PR:000012602 - gastricsin" has an exact synonym "PGC"; this specific protein is not seen in CRAFT, but the abbreviation "PGC" is seen ∼400 times referring to the

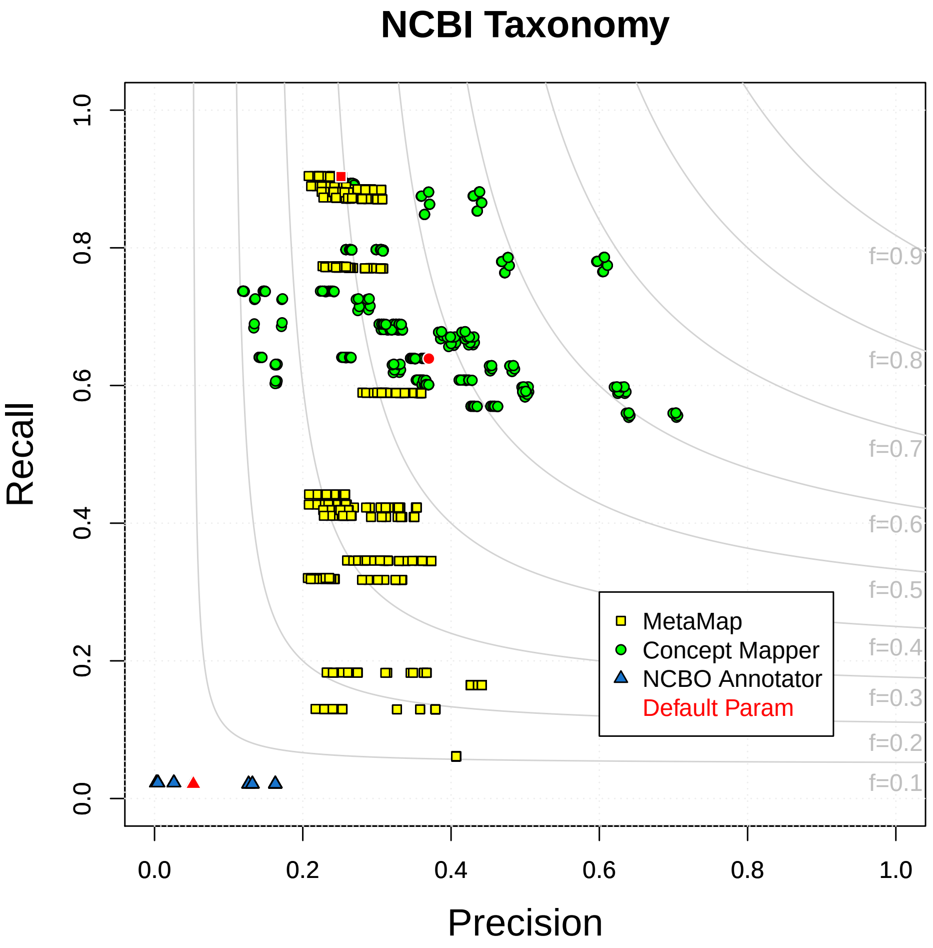


Figure 2.10: All parameter combinations for NCBITaxon. The distribution of all parameter combinations for each system on NCBITaxon. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

protein peroxisome proliferator-activated receptor-gamma. By addressing just these few categories of FPs, we can increase the performance of all systems.

2.4.8 NCBI Taxonomy

The NCBI Taxonomy is a curated set of nomenclature and classification for all the organisms represented in the NCBI databases. It is by far the largest ontology evaluated, at almost 800,000 terms, but with only 7,820 total NCBITaxon annotations in CRAFT. Performance on NCBITaxon varies widely for each system: NCBO Annotator performs poorly (F=0.04), MM performers better (F=0.45) and CM performs best (F=0.69). When looking at all parameter combinations for each system, there is generally a dimension (P or R) that varies widely among the systems and another that is more constrained (Figure 2.10).

In CRAFT, text is annotated with the most closely matching explicitly represented concept. For many organismal mentions, the closest match to an NCBI Taxonomy concept is a genus or higher-level taxon. For example, “mice” and “mouse” are annotated with the genus “NCBITaxon:10088 - Mus”. CM and MM both find mentions of “mice”, but NCBO Annotator does not. (Why will be discussed in the next paragraph.) All systems are able to find annotations to specific species; for example, “Takifugu rubripes” is correctly annotated with “NCBITaxon:31033 - Takifugu rubripes”. The FPs found by all systems are from ambiguous terms and terms that are too specific. Since the ontology is large and names of taxa are diverse, the overlap between terms in the ontology and common words in English and biomedical text introduces these ambiguous FPs. For example, “NCBITaxon:169495 - this” is a genus of flies, and “NCBITaxon:34205 - Iris germanica”, a species of monocots, has the common name “flag”. Throughout biomedical text there are many references to figures that are incorrectly annotated with “NCBITaxon:3493 - Ficus”, which has a common name of “figs”. A more biologically relevant example is “NCBITaxon:79338 - Codon” which is a genus of eudicots but also refers to a set of three adjacent nucleotides. Besides ambiguous terms, annotations are produced that are more specific than those in CRAFT. For example, “rat” in CRAFT is annotated at the genus level “NCBITaxon:10114 - Rattus”; while all systems incorrectly annotate “rat” with more specific terms such as, “NCBITaxon:10116 - Rattus norvegicus” and “NCBITaxon:10118 - Rattus sp.”. One way to reduce some of these false positives is to limit the domains in which matching is allowed, however, this assumes some previous knowledge of what the input will be.

Recall of >0.9 is achieved by some parameter combinations of CM and MM, while the maximum F-measure combinations are lower (CM - R=0.79 and MM - R=0.88). NCBO Annotator produces very low recall (R=0.02) and performs poorly due to a combination of: the way CRAFT is annotated and the way NCBO Annotator handles linking between ontologies. In NCBO Annotator, for example, the link between “mice” and “Mus” is not inferred directly, but goes through the MaHCO ontology (DeLuca et al., 2009), an ontology of major histocompatibility complexes. Because we limited NCBO Annotator to only using ontology directly tested, the link between “mice” and “Mus” is not used, and therefore are not found. For this reason, NCBO Annotator is unable to find many of the NCBITaxon annotations in CRAFT. On the other hand, CM and MM are able to find most annotations, the annotations missed are due to different annotation guidelines or specific species with a single-letter genus abbreviation. In CRAFT, there are ∼200 annotations of the ontology root, with text such as “individual” and “organism”; these are annotated because the root was interpreted as the foundational type of organism. An example of a single-letter genus abbreviation seen in CRAFT is “D. melanogaster” annotated with “NCBITaxon:7227 - Drosophila melanogaster”. These types of missed annotations are easy to correct for through some synonym management or post-processing step. Overall, most of the terms in NCBITaxon are able to be found and focus should be on increasing precision without losing recall.

2.4.8.1 NCBO Annotator parameters

Two parameters were found to be significant: *wholeWordsOnly* (p=2.2×10^{-16}) and *minTermSize* (p=7.2×10^{-5}). Matching non-whole words decreases P by 0.1 with a slight increase in R. Varying *wholeWordsOnly* on the maximum F-measure parameter combination finds ∼15 more TPs and ∼5,500 more FPs. All correct annotations found contain connected punctuation that hinder recognition. For example, “Danio rerio” from the span “(Danio rerio [Dr])” is correctly annotated with “NCBITaxon:7955 - Danio rerio”. Unfortunately, many errors are introduced by matching terms within longer words. For instance, the genus of bony fish, “NCBITaxon:385272 - Conta”, is seen within “contain” and its variants. It is suggested to only allow matching to whole words.

Filtering terms that are less than *FIVE* characters leads to the best performance on NCBITaxon, increasing P by 0.1 with no loss of recall over other parameter values. Comparing lengths of *ONE* to *FIVE*, ∼250 more FPs are found when not removing terms less than *FIVE* characters. For example, “lens” is incorrectly annotated with the genus of flowering plants “NCBITaxon:3863 - Lens”. For the reasons stated in the previous paragraph on recall of NCBITaxon, NCBO Annotator does not find any correct annotations that are less than *FIVE* characters.

2.4.8.2 MetaMap parameters

Four parameters were found to be significant: wordOrder (p=1.4×10^{-6}), *derivationalVariants* (p=1.0×10^{-8}), *scoreFilter* (p=2.2×10^{-16}), and *minTermSize* (p=2.2×10^{-16}). Allowing MM to ignore the order of tokens varies P slightly but decreases R by 0-0.3. Unfortunately, changing wordOrder on the maximum F-measure combination only introduces ~5 more FPs, so the full effect of the parameter is not really seen. Even though the effect cannot be seen, keeping the order of the tokens produces the maximum F-measure.

Generating variants of terms helps performance of most other ontologies evaluated, but not using any derivational variants produces highest F-measure for NCBITaxon. Allowing MM to use variants decreases P 0.05-0.1 with only slight increase in R. Using *ADJ NOUN ONLY* variants finds ~150 more TPs along with ~5,000 more FPs. There are some cases where variants are valid, such as "mammalian" correctly annotated with "NCBITaxon:40674 - Mammalia". For the the most part, nomenclature variants do not follow the same rules for English words. For example, a genus name of birds is "NCBITaxon:189528 - Indicator"; when variants of this are generated the words "indicate(s)", "indicated", "indicating", and "indication" are incorrectly annotated with it. Even though *derivationalVariants* are not used, variants such as "mice" → "mouse" are still correctly found; this shows that inflectional variants are apparently handled by MM even when *derivationalVariants* are not used and suggests that this behavior cannot be controlled with a parameter. For best performance on NCBITaxon, do not use any variants.

Unlike NCBO Annotator, it is best to filter terms less than 1 or 3 characters in length. There is no difference between removing terms less than 1 or 3, but filtering terms less than 5 decreases R by 0.1-0.6. This is the case because many correct annotations found, e.g. "mice", are less than 5 characters.

2.4.8.3 ConceptMapper parameters

All but one parameter were found to be significant: *searchStrategy* (p=3.9×10^{-9}), *caseMatch* (p=9.9×10^{-14}), *stemmer* (p=2.2×10^{-16}), *stopWords* (p=3.5×10^{-4}), *findAllMatches* (p=2.2×10^{-16}), and *synonyms* (p=2.9×10^{-7}). *caseMatch* is an interesting parameter; for the best performing combinations, it does not matter because BioLemmatizer is used and it natively changes everything to lower case. Also, allowing CM to find all matches, instead of only the longest one, leads to a decrease in P of 0.3-0.4 with only a slight increase in R.

The maximum F-measure by CM uses BioLemmatizer as a *stemmer*. An increase in R of 0.2 and varying effects on P are seen by using BioLemmatizer over *NONE* or Porter. ∼1,700 more TPs and ∼2,000 more FPs are found by varying BioLemmatizer vs *NONE* on the maximum F-measure combination. A majority of the correct annotations found (∼1,100) are from the variant "mouse" being correctly normalized to "NCBITaxon:10088 - Mus". Not all variants generated are correct. For example, "area" is incorrectly annotated with "NCBITaxon:293506 - Areae" and the gene that controls coat color, "agouti", is an incorrect variant of "agoutis", which is the common name of "NCBITaxon:34845 - Dasyprocta". Even though more FPs are found, the increase in R outweighs the loss of P and a total increase in F of 0.07 is seen.

Removing PubMed stop words produces differing results; for some parameter combinations there is an increase in P of 0.05-0.2 and for others there is a decrease in P of 0-0.05, while stop words doesn't seem to effect R. Not removing stop words finds ∼2,600 more incorrect annotations. A majority (∼1,800) of the errors introduced by not removing stop words are due to the word "this" being incorrectly annotated with "NCBITaxon:169495 - This". Not removing stop words and then allowing stemming introduces errors as well. For example, "can", "could", and "cannot" are incorrectly annotated with "NCBITaxon:4627 - Canna". Removing stop words produces the highest F-measure because these common English words are ignored.

2.4.9 ChEBI

The Chemical Entities of Biological Interest (ChEBI) Ontology focuses on the representation of molecular entities, molecular parts, atoms, subatomic particles, and biochemical rules and applications. The complexity of terms in ChEBI varies from the simple single-word compound "CHEBI:15377 - water" to very complex chemicals that contain numerals and punctuation, e.g., "CHEBI:37645 - luteolin 7-O-[(beta-D-glucosyluronic acid)-(1->2)-(beta-D-glucosiduronic acid)] 4'-O-beta-D-glucosiduronic acid". The maximum F-measure on ChEBI is produced by CM and NCBO Annotator (F=0.56) with MM (F=0.42) not

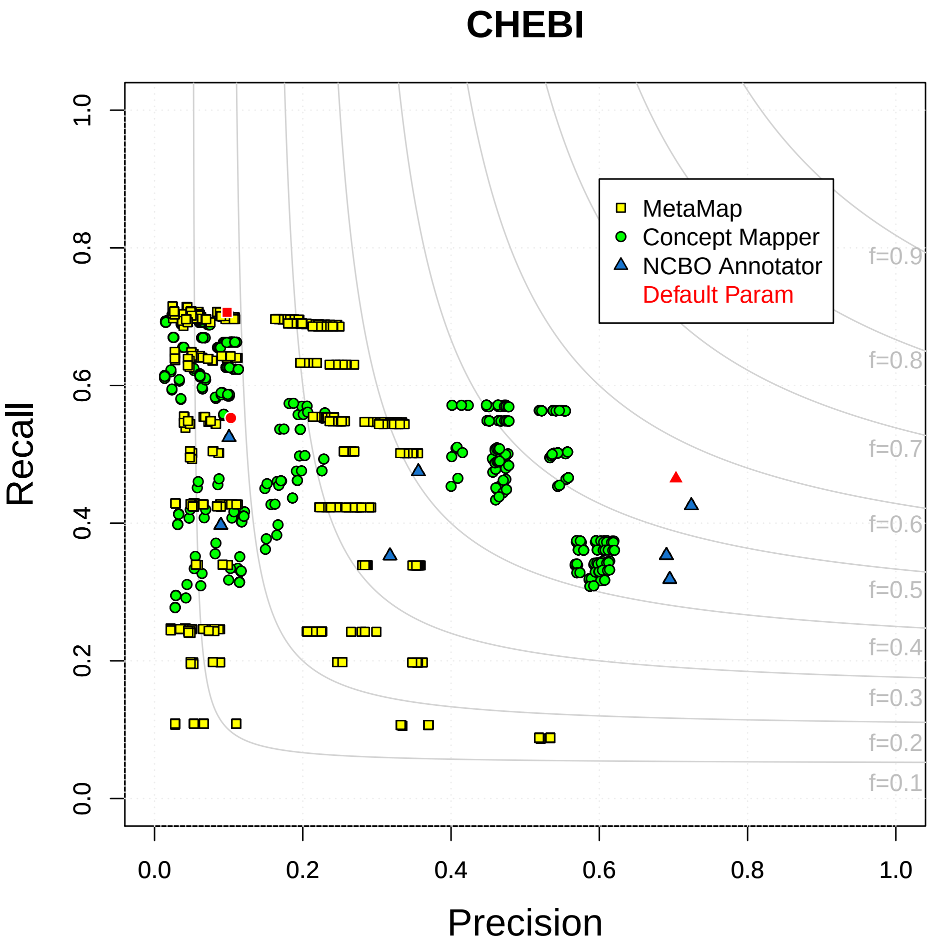


Figure 2.11: All parameter combinations for ChEBI. The distribution of all parameter combinations for each system on ChEBI. (MetaMap - yellow square, ConceptMapper - green circle, NCBO Annotator - blue triangle, default parameters - red.)

performing as well. CM and MM both find ∼4,500 TPs, but because MM finds ∼5,000 more FPs its overall performance suffers (Table 2.4). All parameter combinations for each system on ChEBI can be seen in Figure 2.11.

There are many terms that all systems correctly find, such as “protein” with “CHEBI:36080 - protein” and “cholesterol” with “CHEBI:16113 - cholesterol”. Errors seen from all systems are due to differing annotation guidelines and ambiguous synonyms. Errors from both CM and MM come from generating variants while MM produces some unexplained errors. Different annotation guidelines lead to the introduction of both FPs and FNs. For example, in CRAFT, “nucleotide” is annotated with “CHEBI:25613 - nucleotidyl group”, but all systems incorrectly annotate “nucleotide” with “CHEBI:36976 - nucleotide” because they exactly match. (Mentions of “nucleotide(s)” that refer to nucleotides within nucleic acids are not annotated with “CHEBI:36976 - nucleotide” because this term specifically represents free nucleotides, not those as parts of nucleic acids.) Many FPs and FNs are produced by a single nested annotation; four gold-standard annotations are seen within “amino acid(s)”. Of these four annotations, two are found by all systems, “CHEBI:37527 - acid” and “CHEBI:46882 - amino”, while one introduces a FP: “CHEBI:33709 - amino acid” incorrectly annotated instead of “CHEBI:33708 - amino-acid residue”, while “CHEBI:32952 - amine” is not found by any system. Ambiguous synonyms also lead to errors; for example, “lead” is a common verb but also a synonym of “CHEBI:25016 - lead atom” and “CHEBI:27889 - lead(0)”. Variants generated by CM and MM do not always carry the same semantic meaning as the original term, such as “based” and “basis” from “CHEBI:22695 - base”. MM also produces some interesting unexplainable errors. For example, “disease” is incorrectly annotated with “CHEBI:25121 - maleoyl group”, “CHEBI:25122 - (Z)-3-carboxyprop-2-enoyl group”, and “CHEBI:15595 - malate(2-)”; all three terms have a synonym of “Mal”, but we could find no further explanations.

Recall for maximum F-measure combinations are in a similar range, 0.46-0.56. The two most common categories of annotations missed by all systems are abbreviations and a difference between terms and the way they are expressed in text. Many terms in ChEBI are more commonly seen as abbreviations or symbols. For instance, “CHEBI:29108 - calcium(2+)” is more commonly seen as “Ca2+”; even though it is a related synonym, the systems evaluated are unable to find it. A more complicated example can be seen when talking about the chemicals that lie on the ends of amino acid chains. In CRAFT, “C” from “C-terminus” is annotated with “CHEBI:46883 - carboxyl group” and “CHEBI:18245 - carboxylato group” (the double annotation indicating ambiguity among these), which all systems are unable to find; The same principle also applies for the N-terminus. One simple annotation that should be easy to get is “mRNA” annotated with “CHEBI:33699 - messenger RNA”, but CM and NCBO Annotator miss it. There is not always an overlap between the term names and their expression in text. For instance, the term “CHEBI:36357 - polyatomic entity” was chosen to annotate general “substance” words like “molecule(s)”, “substances”, and “compounds" and "CHEBI:33708 - amino-acid residue" is often expressed as "amino acid(s)" and "residue".

2.4.9.1 NCBO Annotator parameters

Only *wholeWordsOnly* ($p=2.3 \times 10^{-11}$) was found to be significant. Like all other ontologies above, it is best to only match whole words. When allowing to match non-whole words, P decreases 0.4-0.6 with a slight decrease in R. $\sim$500 more TPs and $\sim$36,000 more FPs are found when NCBO Annotator recognizes non-whole words. Correct annotations found by matching non-whole words contain punctuation. For example, "CHEBI:37527 - acid" and "CHEBI:30879 - alcohol" are correctly found in "Fekete's acid-alcohol-formalin fixative". ChEBI contains small terms that are found within longer words such as "CHEBI:24870 - ion", which is found incorrectly $\sim$17,000 times in words such as "proliferation", "mutation", and "localization". Also many errors are introduced from mixing synonyms and matching non-whole words. For instance, "CHEBI:27007 - tin atom" has a synonym "tin", which is found $\sim$4,000 times within words like "blotting", "continious", and "intestinal". Both of these examples are small and would be filtered out if *minTermSize* =*FIVE* was used, but there are also examples that are longer; for example, "CHEBI:35701 - ester" is incorrectly found within "cholesterol" and "western". Overall, it is best to not match non-whole words.

2.4.9.2 MetaMap parameters

Four parameters were found to be significant: *model* ($p=2.4 \times 10^{-10}$), *acronymAbb* ($p=2.2 \times 10^{-16}$), *scoreFilter* ($p=2.2 \times 10^{-16}$), *minTermSize* ($p=7.2 \times 10^{-14}$). ChEBI is one of two ontologies where a difference is seen between values of the *model* parameter. Using the *STRICT* model instead of *RELAXED* increases P 0-0.5 with no change in R, which leads to an increase in F-measure of 0-0.1. Changing the best parameter combination's model to *RELAXED* finds $\sim$200 more FPs with no more TPs. It is unclear why the errors seen are thought to be correct by MM. For example, the text "Ndrg1", which looks to be a protein name, is incorrectly annotated with terms like "CHEBI:30226 - azanldylidene group", "CHEBI:33268 - monoatomic nitrogen", and "CHEBI:36934 - nitrogen-15 atom". The only thing in common between those three ChEBI terms is they all have a synonym of "N". To achieve the best performance, the *STRICT* model should be used.

For best performance, terms less than 5 characters should be filtered out. By doing this P is increased 0.3-0.5, but R is decreased by 0.2; F-measure is increased by 0.05. Comparing the lengths of terms filtered (3 vs. 5), we find that ∼1,000 TPs are missed but ∼8,000 FPs are avoided. It makes sense that the TPs missed are terms and abbreviations that are 3-4 characters in length such as "CHEBI:37527 - acid", "CHEBI:33290 - food", and "EDTA", which is a synonym of "CHEBI:42191 - ethylenediaminetetraacetic acid". The errors filtered out are mostly due to synonyms that contain ambiguous abbreviations. For example "PGC" is incorrectly annotated with "CHEBI:26336 - prostaglandins C" and "male" is incorrectly annotated with "CHEBI:30780 - maleate(2-)". Along the same lines, the *acronymAbb* parameter can introduce many more erroneous abbreviations if the value is set to *ALL*. In order to minimize errors introduced through abbreviations, it is best to use set *acronymAbb* to *DEFAULT* or *UNIQUE* and to also set *minTermSize* to filter out 5 or less characters.

2.4.9.3 ConceptMapper parameters

Only one parameter was found to be statistically significant, *synonyms* ($p=2.2\times10^{-16}$). This does not mean that this parameter is the only one that matters and that any combination will perform well. What we see happening is that the *synonyms* parameter separates the data into two distinct groups and that the effect of other parameters on each group is widely different. For example, we find that *stemmer* performance is directly tied to which synonyms are used. When *ALL synonyms* are used, there is no difference between any of them, but when using *EXACT synonyms*, the stemmers cluster into three distinct groups, with BioLemmatizer achieving the best performance.

Using *ALL synonyms* decreases P by 0.4-0.6 with varying effects on R. Examining the highest F-measure performing combination, ∼1,000 more TPs and ∼365,000 more FPs are introduced by creating the dictionary with *ALL synonyms* instead of *EXACT*. Correct annotations found are mostly from abbreviations. For example "NaCl" is correctly annotated with "CHEBI:26710 - sodium chloride", "MgSO4" is correctly annotated with "CHEBI:32599 - magnesium sulfate", and "mRNA" is correctly annotated with "CHEBI:33699 - messenger RNA". Abbreviations for chemicals can introduce many errors; for example, "CHEBI:30430

- indium atom" and "CHEBI:30433 - indium(1+)" both have a synonym of "In", which is a common English word and seen ∼56,000 times in CRAFT. Mixing *ALL synonyms* and any stemmer produces interesting errors also. For example, "CHEBI:33783 - beryllium(0)" has a synonym "Be", which is incorrectly annotated to "am", "is", "are", "was", "been", etc... We can conclude that the non-exact synonyms for ChEBI are not helpful for concept recognition.

2.4.10 Overall parameter analysis

Here we present overall trends seen from aggregating all parameter data over all ontologies and explore parameters that interact. Suggestions for parameters for any ontology based upon its characteristics are given. These observations are made from observing which parameter values and combinations produce the highest F-measures and not from statistical differences in mean F-measures.

2.4.10.1 NCBO Annotator

Of the six NCBO Annotator parameters evaluated, only three impact performance of the system: *wholeWordsOnly*, *withSynonyms*, and *minTermSize*. Two parameters, *filterNumber* and *stopWordsCaseSensitive*, did not impact recognition of any terms, while removing stop words only made a difference for one ontology (PRO).

A general rule for NCBO Annotator is that only whole words should be matched; matching whole words produced the highest F-measure on seven out of eight ontologies and on the eighth, the difference was negligible. Allowing NCBO Annotator to find terms that are not whole words greatly decreases precision while minimally, if at all, increasing recall.

Using synonyms of terms makes a significant difference in five ontologies. Synonyms are useful because they increase recall by introducing other ways to express concepts. It is generally better to use synonyms, as only one ontology performed better when not using synonyms (GO_MF).

minTermSize does not effect the matching of terms but acts as a filter to remove matches of less than a certain length. A safe value of *minTermSize* for any ontology would be *ONE* or *THREE* because only very small words (< 2 characters) are removed. Filtering terms less than length *FIVE* is useful, not so much for finding desired terms, but for removing undesired terms. Undesired terms less than *FIVE* characters can be introduced either through synonyms or small ambiguous terms that are commonly seen and should be removed to increase performance. (e.g. “NCBITaxon:3863 - Lens” and “NCBITaxon:169495 - This”)

2.4.10.2 Interacting parameters - NCBO Annotator

Because half of NCBO Annotator’s parameters do not affect performance, we only see interaction between two parameters: *wholeWordsOnly* and *synonyms*. The interactions between these parameters come from mixing *wholeWordsOnly* = *NO* and *synonyms* = *YES*. As noted in the discussion of ontologies above, using this combination of parameters introduces anywhere from 1,000 to 41,000 FPs, depending on the test scenario and ontology. These errors are introduced because small synonyms or abbreviations are found within other words.

2.4.10.3 MetaMap

We evaluated seven MM parameters. The only parameter value that remained constant between all ontologies was *gaps*; we have come to the consensus that gaps between tokens should not be allowed when matching. By inserting gaps, precision is decreased with no or slight increase in recall.

The *model* parameter determines which type of filtering is applied to the terms. The difference between the two values for model is that *STRICT* performs an extra filtering step on the ontology terms. Performing this filtering increases precision with no change in recall for ChEBI and NCBITaxon with no differences between the parameter values on the other ontologies. Because it is best performing on two ontologies and in MM documentation is said to produce the highest level of accuracy, the *STRICT* model should be used for best performance.

One simple way to recognize more complex terms is to allow the reordering of tokens in the terms. Reordering tokens in terms helps MM to identify terms as long as they are syntactically or semantically the same. For example, “GO:0000805 - X chromosome” is equal to “chromosome X”. Practically, the previous example is an exception, as most reorderings are not syntactically or semantically similar; by ignoring token order, precision is decreased without an increase in recall. Retaining the order of tokens produces highest F-measure on six out of eight ontologies, while there was no difference on the other two. We conclude for best performance it is best to retain token order.

One unique feature of MM is that it is able to compute acronym and abbreviation variants when mapping text to the ontology. MM allows the use of *ALL* acronym/abbreviations (-a), only those with *UNIQUE* expressions (-u) and the *DEFAULT* (no flags). For all ontologies, there is no difference between using the *DEFAULT* or only those with *UNIQUE* expressions, but both are better than using *ALL*. Using *ALL* acronyms and abbreviations introduces many erroneous matches; precision is decreased without an increase in recall. For best performance, use *DEFAULT* or *UNIQUE* values of acronyms and abbreviations.

Generating derivational variants helps to identify different forms of terms. The goal of generating variants is to increase recall without introducing ambiguous terms. This parameter produces the most varied results. There are three parameter values (*ALL*, *NONE*, and *ADJ NOUN ONLY*), and each of them produces the highest F-measure on at least one ontology. Generating variants hurts the performance on half of the ontologies. Of these ontologies, variants of terms from PRO and ChEBI do not make sense because they do not follow typical English language rules while variants of terms in NCBITaxon and SO introduce many more errors than correct matches. *ALL* variants produce highest F-measure on CL, while *ADJ NOUN ONLY* variants are best-performing on GO_BP. There is no difference between the three values for GO_CC and GO_MF. With these varied results, one can decide which type of variants to use by examining the way they expect terms in their ontology to be expressed. If most of the terms do not follow traditional English rules, like gene/protein names, chemicals, and taxa, it is best to not use any variants. For ontologies where terms could be expressed as nouns or verbs, a good choice would be to use the default value and generate *ADJ NOUN ONLY* variants. This is suggested because it generates the most common types of variants, those between adjectives and nouns.

The parameters *minTermSize* and *scoreFilter* do not affect matching but act as a post-processing filter on annotations returned. *minTermSize* specifies the minimum length, in characters, of annotated text; text shorter than this is filtered out. This parameter acts exactly like that of the NCBO Annotator parameter with the same name presented above.

MM produces scores in the range of 0 to 1000, with 1000 being the most confident. For all ontologies, a score of 1000 produces the highest P and the lowest R, while a score of 0 returns all matches and has the highest R with the lowest P, with 600 and 800 somewhere between. Performance is best on all ontologies when using most of the annotations found by MM, so a score of 0 or 600 is suggested. As input to MM, we provided the entire document; it is possible that different scores are produced when providing a phrase, sentence, or paragraph as input. The scores are not as important as the understanding that most of the annotations returned by MM are used.

2.4.10.4 ConceptMapper

We evaluated seven CM parameters. When examining best performance, all parameter values vary but one: *orderIndependentLookup* = *OFF*, which does not allow the reordering of tokens when matching, is set in the highest F-measure parameter combination for all ontologies. As for MM, it is best to retain ordering of tokens.

searchStrategy affects the way dictionary lookup is performed. *CONTIGUOUS* matching returns the longest span of contiguous tokens, while the other two values (*SKIP ANY MATCH* and *SKIP ANY ALLOW OVERLAP*) can skip tokens and differ on where the next lookup begins. Performance on six out of eight ontologies is best when only contiguous tokens are returned. On NCBITaxon, the behavior of *searchStrategy* is unclear and unintuitive: By returning non contiguous tokens, precision is increased without loss of recall. For most ontologies, only selecting contiguous tokens produces the best performance.

The *caseMatch* parameter tells CM how to handle capitalization. The best performance on four out of eight ontologies uses *INSENSITIVE* case matching while there is no difference between the values of *caseMatch* on three ontologies. There is no difference on those three because the best parameter combination utilizes the BioLemmatizer, which automatically ignores case. Thus, best performance on seven out of eight ontologies ignores case. PRO is the exception; its best-performing combination only ignores case on those tokens containing digits. For most ontologies, it is best to use *INSENSITIVE* matching.

Stemming and lemmatization allow matching of morphological term variants. Performance on only one ontology, PRO, is best when no morphological variants are used; this is the case because PRO terms annotated in CRAFT are mostly short names which do not behave and have the properties of normal English words. The best-performing combination on all other ontologies use either the Porter stemmer or the BioLemmatizer. For some ontologies, there is a difference between the two variant generators, and for others there was not. Even ontologies like ChEBI and NCBITaxon perform best with morphological variants because they are needed for CM to identify inflectional variants such as plurals. For most ontologies, morphological variants should be used.

CM can take a list of stop words to be ignored when matching. Performance on seven out of eight ontologies is better when stop words are not ignored. Ignoring PubMed stop words from these ontologies introduces errors without an increase in recall. An example of one error seen is the span "proteins that target" incorrectly annotated with "GO:0006605 - protein targeting". The one ontology, NCBITaxon, where ignoring stop words results in best performance is due to a specific term, "NCBITaxon:169495 - this". By ignoring the word "this", ∼1,800 FPs are prevented. If there is not a specific reason to ignore stop words, such as the terms seen in NCBITaxon, we suggest not ignoring stop words for any ontology.

By default CM only returns the longest match; all matches can be returned by setting *findAllMatches* to *TRUE*. Seven out of eight ontologies perform better when only the longest match is returned. Returning all matches for these ontologies introduces errors because higher-level terms are found within lower-level ones and the CRAFT concept annotation guidelines specifically prohibit these types of nested annotations. CHEBI performs best when all matches are returned because it contains such nested annotations. If the goal is to find all possible annotations or it is known that there are nested annotations we suggest to set *findAllMatches* to *TRUE*, but for most ontologies, only the longest match should be returned.

There are many different types of synonyms in ontologies. When creating the dictionary with the value *ALL*, all synonyms (exact, broad, narrow, related, etc...) are used; the value *EXACT* creates dictionaries with only the exact synonyms. The best performance on six out of eight ontologies uses only *EXACT synonyms*. On these ontologies, using only exact instead of all synonyms increases precision with no loss of recall; use of broad, related, and narrow synonyms mostly introduce errors. Performance on PRO and GO_BP is best when

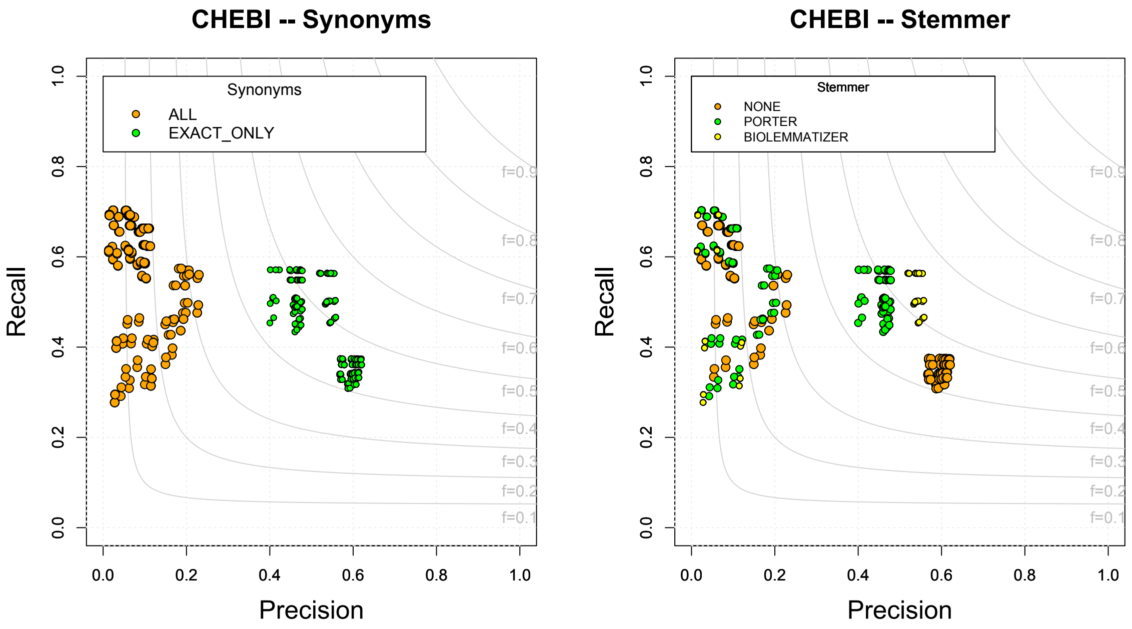


Figure 2.12: Two CM parameter that interact on CHEBI Synonyms (left) and *stemmer* (right) parameter interact. The *stemmer* produce distinct clusters when only *EXACT synonyms* are used. When *ALL synonyms* are used, it is hard to distinguish any patterns in the *stemmer*.

using all synonyms. On these two ontologies, the other types of synonyms are useful for recognition and increase recall. For most ontologies using only exact synonyms produces the best performance.

2.4.10.5 Interacting parameters - ConceptMapper

We see the most interaction between parameters in CM. There are two different interactions that are apparent in certain ontologies: 1) *stemmer* and *synonyms* and 2) *stopWords* and *synonyms*. The first interaction found is in ChEBI. We find the *synonyms* parameter partitions the data into two distinct groups. Within each group, the stemmer parameter has two completely different patterns (Figure 2.12). When only *EXACT synonyms* are used all three stemmers are clustered, with BioLemmatizer performing best, but when *ALL synonyms* are used it is hard to find any difference between the three stemmers. The second interaction found is between the *stopWords* and *synonyms* parameters. In GO_MF several terms have synonyms that contain two words, with one being in the PubMed stop word list. For example, all mentions of "activity" are incorrectly annotated with "GO:0050501 -

hyaluronan synthase activity", which has a broad synonym "HAS activity"; "has" is contained in the stop word list and therefore is ignored.

Not only do we find interactions within CM, but some parameters also mask the effect of other parameters. It is already known and stated in the CM guidelines that the *searchStrategy* values *SKIP ANY MATCH* and *SKIP ANY ALLOW OVERLAP* imply that *orderIndependentLookup* is set to *TRUE*. Analyzing the data, it was also discovered that BioLemmatizer converts all tokens to lower case when lemmas are created, so the parameter *caseMatch* is effectively set to *IGNORE*. For these reasons, it is important to not only consider interactions but also the masking effect that a specific parameter value can have on another parameter.

2.4.11 Substring matching and stemming

Through our analysis we have shown that accounting for morphology of ontological terms has an important impact on the performance of concept annotation in text. Normalizing morphological variants is one way to increase recall by reducing the variation between terms in an ontology and their natural expression in biomedical text. In NCBO Annotator, morphology can only be accommodated in the very rough manner of either requiring that ontology terms match whole (space or punctuation-delimited) words in the running text, or allowing any substring of the text whatsoever to match an ontology term. This leads to overall poorer performance by NCBO Annotator for most ontologies, through the introduction of large numbers of false positives. It should be noted that some substring annotations may appear to be valid matches, such as the annotation of the singular "cell" within "cells". However, given our evaluation strategy, such an annotation would be counted as incorrect since the beginning and end of the span do not directly match the boundaries of the gold CRAFT annotation. If a less strict comparator were used, these would be counted as correct, thus increasing recall, but many FPs would still be introduced through substring matching from e.g., short abbreviation strings matching many words.

MM always includes inflectional variants (plurals and tenses of verbs) and is able to include derivational variants (changing part of speech) through a configurable parameter. CM is able to ignore all variation (*stemmer* = *NONE*), only perform rough normalization

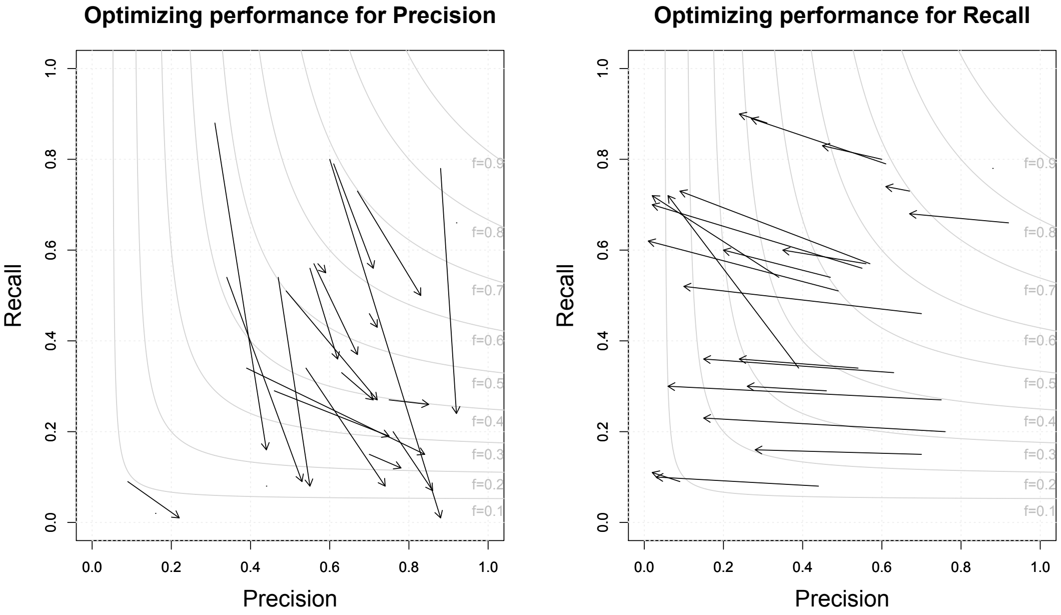


Figure 2.13: Differences between maximum F-measure and performance when optimizing one dimension. Arrows point from best performing F-measure combination to the best precision/recall parameter combination. All systems and all ontologies are shown.

by removing common word endings (*stemmer* = Porter), and handle inflectional variants (*stemmer* = BioLemmatizer). We currently do not have a domain-specific tool available for integration into CM to handle derivational morphology, as well, but a tool that could handle both inflectional and derivational morphology within CM would likely provide benefit in annotation of terms from certain ontologies. If NCBO Annotator were to handle at least plurals of terms, its recall on CL and GO_CC ontologies would greatly increase because many terms are expressed as plurals in text. For ontologies where terms do not adhere to traditional English rules (e.g.,ChEBI or PRO), using morphological normalization actually hinders performance.

2.4.12 Tuning for precision or recall

We acknowledge that not all tasks require a balance between precision and recall; for some tasks high precision is more important than recall, while for others the priority is high recall and it is acceptable to sacrifice precision to obtain it. Since all the previous results are based upon maximum F-measure, in this section we briefly discuss the tradeoffs between precision and recall and the parameters that control it. The difference between the

Table 2.7: Best parameters for optimizing performance for precision or recall.

High precision annotations					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	YES	model	STRICT	searchStrategy	CONTIGUOUS
filterNumber	ANY	gaps	NONE	caseMatch	SENSITIVE
stopWords	ANY	wordOrder	ORDER MATTERS	stemmer	NONE
SWCaseSensitive	ANY	acronymAbb	DEFAULT or UNIQUE	stopWords	NONE
minTermSize	THREE or FIVE	derivationalVariants	NONE	orderIndLookup	OFF
withSynonyms	NO	scoreFilter	1000	findAllMatches	NO
		minTermSize	3 or 5	synonyms	EXACT ONLY

High recall annotations					
NCBO Annotator		**MetaMap**		**ConceptMapper**	
Parameter	**Value**	**Parameter**	**Value**	**Parameter**	**Value**
wholeWordOnly	NO	model	RELAXED	searchStrategy	SKIP ANY or ALLOW
filterNumber	ANY	gaps	ALLOW	caseMatch	IGNORE or INSENSITIVE
stopWords	ANY	wordOrder	IGNORE	stemmer	Porter or BioLemmatizer
SWCaseSensitive	ANY	acronymAbb	ALL	stopWords	PubMed
minTermSize	ONE or THREE	derivationalVariants	ALL or ADJ NOUN	orderIndLookup	ON
withSynonyms	YES	scoreFilter	0	findAllMatches	YES
		minTermSize	1 or 3	synonyms	ALL

maximum F-measure parameter combination and performance optimized for either precision or recall for each system-ontology pair can be seen in Figure 2.13. By sacrificing recall, precision can be increased between 0 and 0.45. On the other hand, by sacrificing precision, recall can be increased between 0 and 0.38.

The best parameter combinations for optimizing performance for precision and recall can be seen in Table 2.7. Unlike the previous combinations seen above, parameters that produce the highest recall or precision do not vary widely between the different ontologies. To produce the highest precision, parameters that introduce any ambiguity are minimized; for example, word order should be maintained and stemmers should not be used. Likewise, to find as many matches as possible, the loosest parameter settings should be used; for example, all variants and different term combinations should be generated along with using all synonyms. The combination of parameters that produce the highest precision or recall are very different from the maximum F-measure-producing combinations.

2.5 Conclusions

After careful evaluation of three systems on eight ontologies, we can conclude that ConceptMapper is generally the best-performing system. CM produces the highest F-measure on seven out of eight total ontologies, while NCBO Annotator and MM both produce the highest F-measure on only one ontology (NCBO Annotator and MM produce equal F-measures on ChEBI). Out of all systems CM balances precision and recall the best; it produces the highest precision on four ontologies and the highest recall on three ontologies. The other systems perform well in one dimension but suffer in the other. MM produces the highest recall on five out of eight ontologies but precision suffers because it finds the most errors; the three ontologies for which it did not achieve highest recall are those where variants were found to be detrimental (SO, ChEBI, and PRO). On the other hand, NCBO Annotator produces the highest precision for four ontologies but falls behind in recall because it is unable to recognize plurals or variants of terms. Overall, CM performs best out of all systems evaluated on the concept normalization task. For this reason, for the rest of the dissertation I utilize the ConceptMapper pipeline described here along with best parameters discovered for concept recognition.

Besides performance, another important thing to consider when using a tool is the ease of use. In order to use CM, one must adopt the UIMA framework. Transforming any ontology for matching is easy with CM with a simple tool that converts any OBO ontology file to a dictionary. MM is a standalone tool that works only with UMLS ontologies natively; getting it to work with any arbitrary ontology can be done but is not straightforward. MM is the most like a black box of all the systems, which results in some annotations that are unintuitive and cannot be traced to their source. NCBO Annotator is the easiest to use as it is provided as a Web service, with large retrieval occurring through a REST service. NCBO Annotator currently works with any of the 330+ BioPortal ontologies. Drawbacks of NCBO Annotator are due to it being provided as a Web service, they include changes in the underlying implementation, resulting in different annotations returned over time; there is also no control over the version of the ontologies used or the ability to add an ontology.

Using the default parameters for any tool is a common practice, but as seen here, the defaults often do not produce the best results. We have discovered that some parameters do not impact performance, while others greatly increase performance when compared to defaults. As seen in the Results and Discussion section, we have provided parameter suggestions for not only the ontologies evaluated but also provide general suggestions that can be applied to any ontology. We can also conclude that parameters cannot be optimized individually. If we didn't generate all parameter combinations and instead examined parameters individually, we would be unable to see these interacting parameters and could have chosen a non-optimal parameter combination as the best.

Complex multi-token terms are seen in many ontologies and are more difficult to normalize than the simpler one- or two-token terms. Inserting gaps, skipping tokens, and reordering tokens are simple methods currently implemented in both CM and MM. These methods provide a simple heuristic but do not always produce valid syntactic structures or retain the semantic meaning of the original term. From our analysis above, we can conclude that more sophisticated, syntactically valid methods need to be implemented to recognize complex terms seen in ontologies such as GO_MF and GO_BP. Chapter III focuses on the improvement of these types of complex Gene Ontology concepts through the implementation of synonym generation rules.

Our results demonstrate the important role of linguistic processing, in particular morphological normalization of terms, during matching. Several of the highest-performing sets of parameters take advantage of stemming or handling of morphological variants, though the exact best tool for this job is not yet entirely clear. In some cases, there is also an important interaction between this functionality and other system parameters, leading to some spurious results. It appears that these problems could be addressed in some cases through more careful integration of the tools and in others through simple adaptation of the tools to avoid some common errors that have occurred.

In this work, we established baselines for performance of three publicly available dictionary-based tools on eight biomedical ontologies, analyzed the impact of parameters for each system, and made suggestions for parameter use on any ontology. We can conclude that of the tested tools, the generic ConceptMapper tool generally provides the best performance on the concept normalization task, despite not being specifically designed for use in the biomedical domain. The flexibility it provides in controlling precisely how terms are matched in text makes it possible to adapt it to the varying characteristics of different ontologies.

CHAPTER III

IMPROVING PERFORMANCE OF CONCEPT RECOGNITION[3]

3.1 Introduction

One of the trends seen in Chapter II was that complex multi-token terms are more difficult to recognize due to increasing variation within natural text. Simple heuristics such as inserting gaps, skipping tokens, and reordering tokens are built into some of the concept recognition systems, but they can only help to a certain point. The most complex terms are those from the Gene Ontology, specifically Molecular Function and Biological Process branches; F-measure performance supports that with 0.14 and 0.42, respectively. The motivating idea behind this chapter is to increase the performance in recognizing these concepts, by expanding their synonym list with both syntactic and derivational variants, by exploiting their underlying compositional nature. The synonym generation rules presented are evaluated both intrinsically on the CRAFT gold-standard corpus along with extrinsically validated on a large collection of 1 million full text articles via manual inspection of random sampling. If these complex processes and functions could be recognized with higher accuracy and recall, automatic methods for both aiding in manual curation and biomedical prediction would greatly be improved.

3.2 Background

The Gene Ontology (GO) represents the standard by which we refer to functions and processes that genes/gene products participate in. Due to its importance in biology and the exponential growth in the biomedical literature over the past years, there has been much effort in utilizing GO for text mining tasks (Hirschman et al., 2005; Mao et al., 2014). Performance on these recognition tasks is lacking; it has been previously seen that there is a large gap between the way concepts are represented in the ontology and the many different ways they are expressed in natural text (Verspoor et al., 2003; Cohen et al., 2008; Brewster et al., 2004).

[3] The work presented in this chapter was submitted to Journal of Biomedical Semantics in April 2015 under the title *Gene Ontology synonym generation rules lead to increased performance in biomedical concept recognition.*

There have been very few evaluations assessing the ability to recognize and normalize Gene Ontology concepts from the literature; this is mostly due to lack of gold-standard annotations. Previous work evaluated concept recognition systems utilizing the Colorado Richly Annotated Full Text Corpus (CRAFT). (Funk et al., 2014a) evaluated three prominent dictionary-based systems (MetaMap, NCBO Annotator, and ConceptMapper) and found Cellular Component was able to be recognized the best (F-measure 0.77). The more complex terms from Biological Process (F-measure 0.42) and Molecular Function (F-measure 0.14) were much more difficult to recognize in text. Campos *et al.* present a framework called *Neji* and compare it against *Whatizit* on the CRAFT corpus (Campos et al., 2013); they find similar best performance (Cellular Component 0.70, Biological Process/Molecular Function 0.35). Other work explored the impact of case sensitivity and information gain on concepts recognition and report performance in the same range as what has previously been published (Cellular Component 0.78, Biological Process/Molecular Function 0.40) (Groza and Verspoor, 2015). Since all methods utilized dictionary based systems it appears that a plateau has been reached utilizing the information contained within the Gene Ontology itself. For further progress to be made, the gap between concept representation and their expression in literature needs to be reduced, which serves as major motivation for the work presented in this manuscript. There have also been sub-tasks within the BioCreative I and IV (Blaschke et al., 2005; Mao et al., 2014) community challenges that involve a task similar, but more difficult, to GO term recognition – relating relevant GO concepts given protein-document pairs – these are not addressed further here, but methods presented here could also be applied to that task.

There are two main applications of biomedical literature mining where improved performance of the GO can greatly help. 1) It is well known that manual curation can no longer keep up with the annotation of gene and protein function (Baumgartner et al., 2007a). Automatic annotation is not our direct goal, but utilizing automatic methods to highlight functions could provide input to curators to help speed up manual curation. 2) The mining of GO concepts from large collections of biomedical literature has been show to be useful for biomedical discovery, for example, pharmacogenomic gene prediction (Funk et al., 2014c) and protein function prediction (Sokolov et al., 2013b; Funk et al., 2015). We hypothesize that we can utilize the compositional nature of the Gene Ontology (Ogren et al., 2004) to develop a small number of language generation rules that will have a large impact on the ability to recognize concepts from biomedical text. We are aware that our method might overgenerate, but we also hypothesize that those synonyms probably will not be found in the biomedical literature, and therefore, will not hinder performance.

In this work we present a number of manually created recursive syntactic and derivational rules to facilitate generation of synonyms for Gene Ontology terms. We evaluate these generated synonyms both intrinsically on a gold standard corpus to show these rules increase performance over any published results for recognition of GO concepts and extrinsically through manual validation of annotations produced on a large collection of literature to illustrate the accuracy and impact of the rules have at a larger scale.

3.2.1 Natural language generation sub-task

We can view this problem as a subset of natural language generation (NLG) field. NLG is the task of automatically generating language from some knowledge source. A good review of NLG and its use in the semantic web space can be seen in Bouayad-Agha *et al.* (Bouayad-Agha et al., 2012). NLG has been used in the biomedical domain to write genetic counseling letters (Green, 2005), question answering (Athenikos and Han, 2010; Chaudhri et al., 2013), and creating textual summaries from knowledgebases (Banik et al., 2013). Language generation has also been applied specifically to ontologies for tasks such as populating an allergen ontology and generating it in multiple languages (Karkaletsis et al., 2006) and presenting textual summaries or descriptions of groups or individual ontological classes (Davies et al., 2008; Karakatsiotis et al., 2008).

Most NLG tasks must be very specific in the language and content they provide to the reader and have to decipher both 'what to say' and 'how to say it' before they can generate language (Mitkov, 2005; Dale et al., 2000), i.e. what information should be conveyed from the input text, how should sentences/phrases be ordered, and what specific syntax and words convey the information the clearest. Our task is more simple than many other NLG tasks, as we are only focused on the generation of all variability within a phrase with the caveat that everything generated by the system should have the same semantics as the original input.

Additionally, the output of our system is not intended for human interpretation, although it can be read, and may contain some syntactically incorrect entities. The designed goal is to use the output to modify a dictionary which is fed into a concept recognition system.

3.2.2 Difficulties in identifying phrasal lexicons

The identification of Gene Ontology terms is more difficult than many other types of named entities such as genes, proteins, or species mainly due to the length (Funk et al., 2014a) and complexity of the concepts along with the many variations that can occur in natural language. To help illustrate this we explore the length of terms along with the ability to recognize them in Figure 3.1. Overall, performance decreases as the complexity increases; additionally, the occurrence of terms decreases significantly after a length of two.

In CRAFT there are about 100 unique single token concepts annotated, which can be recognized with macro-averaged F-measure of about 0.70 (performance is taken from best performing parameter combination from Funk *et al* (Funk et al., 2014a). The highest number of annotations ($\sim$500) is of concepts with a length of two, we see a dramatic decrease in performance when examining these terms (F-measure = 0.33). This decreasing trend continues until spikes in performance are seen due to recognition of a single complex term with only a few total instances.

3.2.3 Compositionality of the Gene Ontology

The structure of concepts from the Gene Ontology has been been noted by many to be compositional (Ogren et al., 2004, 2005; Hill et al., 2002). A term such as "GO:1900122 - positive regulation of receptor binding" contains another concept "GO:0005102 - receptor binding"; not only do the strings overlap, but the terms are also connected by relationships within the ontology. Ogren *et al.* explore more in detail terms as proper substring of other terms (Ogren et al., 2004). Additionally, previous work examined the compositionality of the GO and employed finite state automata (FSA) to represent sets of GO terms (Ogren et al., 2005). An abstracted FSA described in that work can be seen in Figure 3.2. This example shows how terms can be decomposed into smaller parts and how many different terms share similar compositional structure.

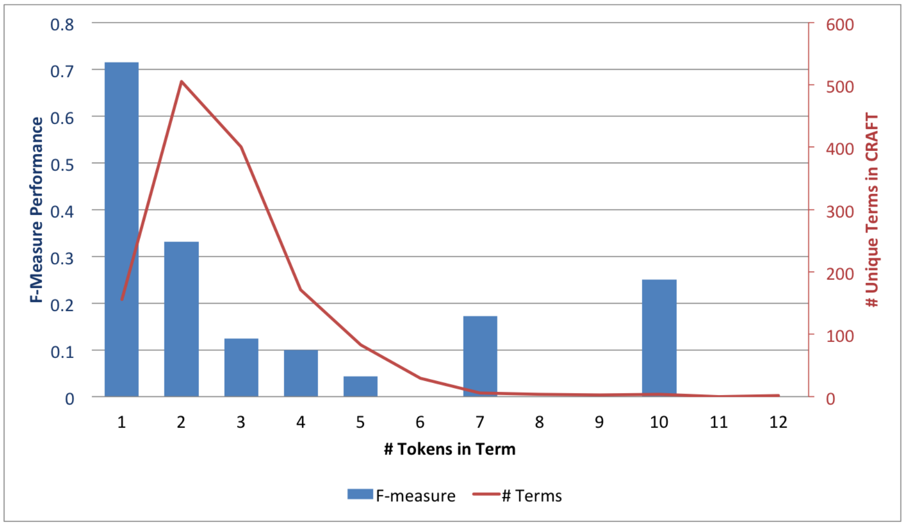


Figure 3.1: Complexity of terms in CRAFT. Plotting performance and number of terms vs the number of tokens within a GO concept name in the CRAFT corpus. Blue bars correspond to macro-averaged F-measure performance broken down by complexity of terms in number of tokens. Red line corresponds to the number of unique GO terms annotated in the CRAFT.

To facilitate generation of meaning (cross-product definitions) and consistency within the ontology, a system called *Obol* (Mungall, 2004) was developed. This work involved analyzing the structure of terms through the creation of grammars to decompose and understand the formal language underlying the GO. An example grammar describing the positive regulation of a molecular function term follows: *process(P that positively_regulates(F))* ⇒ *[positive],regulation(P),[of],molecular_function(F)*. These grammars serve as templates for the decompositional rules utilized in this work. Recently, GO has been moving away from pre-computed term, towards post-computed *on-the-fly* creation of terms for annotations using cross-products (Huntley et al., 2014). Additionally, TermGenie (Dietze et al., 2014) was developed, using a pattern-based approach, to automatically generate new terms and place them appropriately within the Gene Ontology. This work dealt with the analysis and generation of new terms for curation, but no work has been focused on synonym generation. It is evident that compositionality is a prevalent phenomenon within the Gene Ontology.

There has been previous work using the compositional nature and common syntactic patterns within the Gene Ontology itself to automatically generate lexical elementary synonym sets (Hamon and Grabar, 2008). This method generates a total of 921 sets of synonyms with a majority being generated from 1-3 terms; 80% of the sets refer to orthographic

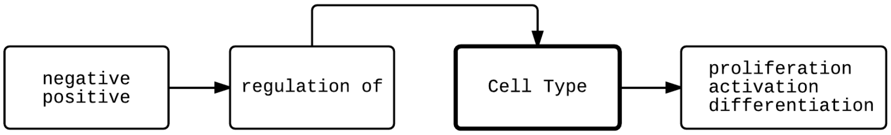


Figure 3.2: Finite state automata representing activation, proliferation, and differentiation GO terms. An abstracted FSA adapted from a figure in Ogren *et al.* (Ogren et al., 2005) that shows how a particular term can be decomposed into its smaller components; where "cell type" can be any specific type of cell.

{synthase, sythetase}, chemical products {gallate, gallic acid}, or Latin inflection {flagella, flagellum}. We believe this method is complementary to what we present here. We manually created these sets through analysis of Gene Ontology annotations in unstructured text. Additionally we go beyond and incorporate derivational variants, i.e. flagella⇒flagellar, which have been shown to be very useful for capturing the natural language of concepts. We were currently unable to find them publicly available, but if we should, they could be seamlessly integrated within the synonym generation rules we present here.

Other work takes advantage of the structure of the Gene Ontology and relationships between GO terms to show that these properties can aid in the creation of lexical semantic relationships for use in natural language processing applications (Verspoor et al., 2003). Besides compositionality, previous work tries to identify GO terms that express similar semantics that use distinct linguistic conventions (Verspoor et al., 2009). They find that, in general, concepts from the Gene Ontology are very consistent in their representation (there are some exceptions but are quality issues that the consortium would like to avoid or fix). This signifies that the Gene Ontology is an excellent starting point for rule-based synonym generation. The consistency of term representation along with the underlying compositional structure suggests the effective generation of synonyms for many terms using only a small number of rules.

3.2.4 Objectives of this work

The goal of this work is to take advantage of the consistency and compositionality underlying Gene Ontology concepts to create the fewest rules that have the largest impact on recognition recall, thus the rules presented here take into account the largest classes of concepts. We leave creation of rules that affect smaller classes of concepts for future work. Also, it is possible that some synonyms generated are not linguistically or syntactically correct; this should not impact performance because they are being recognized from a dictionary and then should not appear in free text. Is is the goal to generate and release these synonyms for the larger biomedical natural language processing community. Currently, we do not suggest that all generated synonyms be considered for addition to GO or for other uses, but we have ideas on ways that we can filter them to suggest the best generated terms as synonyms.

3.3 Methods

3.3.1 Structure of the Gene Ontology

The ontologies used were obtained from the Open Biomedical Ontology (obo) To help to understand the structure of the obo file, an entry of a concept from GO is shown in Figure 3.3. The only parts of an entry used in our systems are the id, name, and synonym rows. Alternative ways to refer to terms are expressed as synonyms; there are many types of synonyms that can be specified with different levels of relatedness to the concept (exact, broad, narrow, and related). Currently, we treat all synonyms as equal, but can alter that thinking quite easily. An ontology specification can express a hierarchy among its terms; these are expressed in the "is_a" entry. Terms described as "ancestors", "less specific", or "more general" lie above the specified concept in the hierarchy, while terms described as "more specific" are below the specified concept.

id: GO:0001764

name: neuron migration

namespace: biological_process

def: "The characteristic movement of an immature neuron from germinal zones to specific positions where they will reside as they mature."

synonym: "neuron chemotaxis" EXACT

synonym: "neuron guidance" RELATED

is_a: GO:0016477 ! cell migration

relationship: part_of GO:0048699 ! generation of neurons

Within this work we utilize the id, name, namespace, and synonym directly. Indirectly, we use the is_a and relationships through decomposition and matching of the name field.

3.3.2 CRAFT corpus

The gold standard used is the Colorado Richly Annotated Full-Text (CRAFT) Corpus (Bada et al., 2012; Verspoor et al., 2012). The full CRAFT corpus consists of 97 completely annotated biomedical journal articles, while the "public release" set, which consists of 67 documents, was used for this evaluation. CRAFT includes over 100,000 concept annotations from eight different biomedical ontologies. Even though the collection is small, there is no other corps that has text-level annotations of Gene Ontology concepts.

3.3.3 Large literature collection

To test generalization and for further analysis the impact our concept recognition can have, we utilized a large collection of one million full-text articles from Elsevier. This is a private collection of full-text documents from a wide variety of biomedical Elsevier journals. We are aware of the difficulties associated with obtaining or even accessing full-text documents from subscription journal publishers in a machine readable format, primarily XML. This collection of articles was procured negotiating a licensing deal mediated through on campus librarians, a possible solution and lessons learned from that process is described in Fox *et al.* (Fox et al., 2014).

3.3.4 Concept recognition pipeline and baselines

The baseline for GO recognition was established in previous work (Funk et al., 2014a) through parameter analysis of three different concept recognition systems. The top performing system, ConceptMapper (CM), is used for the following test because it produced highest F-measures on 7 out of 8 ontologies in the CRAFT corpus. CM takes an obo file and converts it to an xml dictionary, which is used to recognize concepts in free text. In analyzing the results there are two different baselines that were provided, 1) using best parameter combination and only information contained within the ontology obo file and 2) using best parameter combination and modifying the dictionary to account for the "activity" terms. Both baseline numbers are presented for comparison.

For the intrinsic evaluation pipeline, we use the version of GO used to annotate CRAFT from November 2007. We are aware of the great number of changes made, but this was purposefully done to keep the concepts available to the dictionary the same that were available to the annotators when they marked up the gold standard. To show that the rules created are able to generalize and apply to much more of the Gene Ontology terms added since 2007, for the extrinsic evaluation on large collection we use an updated version of the GO from 09/25/2014.

3.3.5 ConceptMapper

ConceptMapper (CM) is part of the Apache UIMA Sandbox (Ferrucci and Lally, 2004) and is available at http://uima.apache.org/d/uima-addons-current/ConceptMapper. Version 2.3.1 was used for these experiments. CM is a highly configurable dictionary lookup tool implemented as a UIMA component. Ontologies are mapped to the appropriate dictionary format required by ConceptMapper. The input text is processed as tokens; all tokens within a span (sentence) are looked up in the dictionary using a configurable lookup algorithm.

For each branch of GO we used the highest performing parameter combination previously identified (Funk et al., 2014a). Table 3.1 provides a summation of the different type of ConceptMapper parameters and shows the exact parameter combinations used for recognition of each sub-branch of the Gene Ontology.

3.3.6 Evaluation of generated synonyms

To evaluate the synonyms given we use the same pipelines described in (Funk et al., 2014a). Synonyms are generated by each method and then only those that are unique (both within the generated synonyms and GO itself) are inserted into a temporary obo file. The temporary obo file is then used to create an xml dictionary used by ConceptMapper (Tanenblatt et al., 2010) for concept recognition. The CRAFT corpus is used as the gold standard and precision, recall, and macro-averaged F-measure are reported for each branch of the GO.

Table 3.1: Summarization of ConceptMapper parameters.

Parameter	Description	MF	BP	CC
Search strategy	CONTIGUOUS - returns longest match of contiguous tokens in the span, SKIP_ANY - returns longest match of not-necessarily contiguous tokens in the span, SKIP_ANY_ALLOW_OVERLAP - returns longest match of not-necessarily contiguous tokens in the span, this implies orderIndLookup	CONTIGUOUS	CONTIGUOUS	CONTIGUOUS
Case match	IGNORE - fold everything to lowercase more matching, INSENSITIVE - fold only tokens with initial caps to lowercase, SENSITIVE - performs no case folding, FOLD_DIGIT -fold only (and only) tokens with a digit	IGNORE	INSENSITIVE	INSENSITIVE
Stemmer	specifics which stemmer to use - PORTER, BIOLEMMATIZER, or NONE	BIOLEMMATIZER	PORTER	PORTER
Stop words	a list of stopwords to remove - PUBMED or NONE	NONE	NONE	NONE
Order independent lookup	if set to TRUE token ordering within the sentence is ignored (”box top” would match“top box”) - TRUE or FALSE	FALSE	FALSE	FALSE
Find all matches	If TRUE all dictionary matches within the sentence are returned, otherwise only the longest is returned - TRUE or FALSE	FALSE	TRUE	FALSE
Synonyms	specifies which synonyms will be included when making the dictionary - EXACT_ONLY or ALL	EXACT_ONLY	ALL	EXACT_ONLY

3.3.7 Manually created rules

Each of our rules was manually created through the analysis of concept annotations within the gold standard CRAFT corpus and through discussions with an ontologist and biologist about how they most frequently represent certain concepts. Every corpus will have its set of annotation guidelines that are specific The set of derivational rules has been tuned, through error analysis, to produce high performance on the CRAFT corpus. We show that these rules are able to recognize many different terms not only in CRAFT but also in a large collection of the biomedical literature, but it is possible that depending on the task these rules should be modified.

3.4 Results and discussion

3.4.1 Current synonyms are not sufficient for text-mining

To illustrate that current synonyms are insufficient for text mining we utilize the CRAFT corpus and present both a specific example of variability within a single GO concept and follow by presenting insufficiency at a the corpus level.

3.4.1.1 Evaluation of an individual concept synonyms

We examined all variations in the CRAFT corpus of the concept “GO:0006900 - membrane budding”; the entry for the concept in the ontology file is inserted below. Like most other terms, the concept name appears as a noun and the entry contains a few synonyms (Figure 3.3).

```
id: GO:0006900
name: membrane budding
namespace: biological\_process
def: ‘‘The evagination of a membrane resulting in formation of a vesicle.’’
synonym: ‘‘membrane evagination’’ EXACT
synonym: ‘‘nonselective vesicle assembly’’ RELATED
synonym: ‘‘vesicle biosynthesis’’ EXACT
synonym: ‘‘vesicle formation’’ EXACT
is\_a: GO:0016044 ! membrane organization and biosynthesis
relationship: part\_of GO:0016192 ! vesicle-mediated transport
```

Figure 3.3: Example ontology entry for the concept “membrane budding”.

There were eight varying expressions of “membrane budding” in all of CRAFT, five of which are contained within a single article about expression and localization of Annexin A7 (Herr et al., 2003). In Table 3.2 we list the CRAFT annotations along with sentential context. The first two examples can be identified with context from the ontology file, but the others cannot.

By analyzing the different ways “membrane budding” is expressed in CRAFT, we find that a majority of the annotations are phrased around the end product, the vesicle. To help recognize these (currently) un-recognizable annotations we can reorder words and change the syntax (“budding of vesicles”). We can also generate derivational variants of “vesicle” (“vesiculation” and “vesiculate”). This one example illustrates that a rather simple term can be expressed in natural language in many different ways, that convey identical semantic meaning. We also illustrate how a few simple rules can help create synonyms to recognize them.

Table 3.2: Examples of the "membrane budding" concept within a single document.

Lipid rafts play a key role in **membrane budding**. . .
Having excluded a direct role in **vesicle formation**. . .
. . . involvement of annexin A7 in **budding of vesicles**
. . . Ca2+-mediated **vesiculation process** was not impaired
Red blood cells which lack the ability to **vesiculate** cause. . .

3.4.1.2 Full corpus evaluation of concept synonyms

The overlap between Gene Ontology concepts and their synonyms and the how concepts appear in the entire CRAFT corpus can be seen in Table 3.3, broken down by sub-ontology. There are a total of 1,108 unique GO terms annotated within CRAFT. Of those 1,108 terms, 353 (31.9%) contain at least one synonym in the ontology. Of those 353 terms that contain synonyms, only 126 (36% of those that have synonyms, 11.4% of all terms in CRAFT) terms have synonyms that appear in CRAFT. The numbers presented should be taken with consideration that CRAFT is a small collection of 67 full text articles, but 26.3% (291 unique terms) of the terms annotated in CRAFT cannot be mapped back to the official term name or a synonym, indicating that even in a small coherent corpus that the ontologies themselves do not contain enough information to accurately identify concepts within biomedical text; this problem will be magnified when scaled to the entire biomedical literature.

Table 3.3: Synonym analysis using CRAFT version of the Gene Ontology (11/2007).

Ontology	Unique GO terms	Term name matched	Contain synonym(s)	Matched synonym(s)	No synonyms & not matched
Cellular Component	205	114 (55.6%)	84 (40.9%)	17 (10.2%)	52 (25.4%)
Molecular Function	270	12 (4.4%)	176 (65.2%)	16 (9.1%)	89 (33.0%)
Biological Process	633	228 (36.0%)	93 (14.7%)	93 (100.0%)	150 (23.7%)
Total	1,108	354 (31.9%)	353 (31.9%)	126 (35.7%)	291 (26.3%)

We performed the same synonym analysis after we applied our compositional rules to the CRAFT version of the Gene Ontology (Table 3.4). Our compositional rules, described in detail below, introduce almost three times as many synonyms that are contained in the original version of the ontology. Additionally, utilizing the synonym generate rules, almost half of concepts contain synonyms that are seen in the text. These two facts lead to a large increase in terms that are able to be mapped back to the text; with our rules only 10% of unique terms in CRAFT are unable to be mapped by to their identifier. From this, we briefly show that our compositional rules help to decrease the gap between the ontological concepts and the many ways they can be expressed in natural text.

Table 3.4: Synonym analysis using compositional rules applied to CRAFT version of the Gene Ontology (11/2007).

Ontology	Unique GO terms	Term name matched	Contain synonym(s)	Matched synonym(s)	No synonyms & not matched
Cellular Component	205	98 (47.8%)	164 (80.0%)	77 (47.0%)	29 (14.1%)
Molecular Function	270	12 (4.4%)	259 (95.9%)	126 (48.6%)	11 (4.1%)
Biological Process	633	221 (34.9%)	528 (83.4%)	215 (40.7%)	77 (12.2%)
Total	1,108	331 (29.8%)	951 (85.8%)	418 (44.0%)	117 (10.6%)

Because the version of GO used to annotated CRAFT is from November 2007 and we wanted to explore the changes made over the years, the same analysis was performed using a more recent version of GO from September 2014 (Table 3.5). Examining the differences between the tables using only information from GO (Table 3.3 and 3.5) we see the vast amount of work the Gene Ontology Consortium and curators have put into the Gene Ontology. We find that over a 7 year period, almost twice as many terms have at least one synonym, with major contribution in the Biological Process branch. We also find that the official name of some concepts have changed over time. This is evident by the decrease in term name matched and an increase in number of matched synonyms; most likely the name was made a synonym and a new official name was added. Despite the improvements made, there are still 200 unique terms (vs. 291 in the CRAFT version of GO) annotated in CRAFT with something other than their term name or synonym. This analysis of synonyms supports the notion that there is still a large gap in the way terms are represented in the ontology and the way they are expressed in natural language.

Even though we see great strides by the GO Consortium over the past seven years, our use of the Gene Ontology for text-mining differs from its intended use for functional annotation. It is apparent that work is needed to help recognize GO concepts from the

Table 3.5: Synonym analysis using most recent version of the Gene Ontology (09/2014).

Ontology	Unique GO terms	Term name matched	Contain synonym(s)	Matched synonym(s)	No synonyms & not matched
Cellular Component	205	110 (53.7%)	109 (53.1%)	28 (25.9%)	25 (12.2%)
Molecular Function	270	12 (4.4%)	198 (73.3%)	17 (8.6%)	61 (22.6%)
Biological Process	633	210 (33.2%)	398 (62.9%)	117 (29.4%)	114 (18.0%)
Total	1,108	332 (30.0%)	705 (63.6%)	152 (17.3%)	200 (18.1%)

biomedical text and is a major motivation for this synonym generation rules presented in this work.

3.4.2 Sources of synonyms

We explore three different methods for generating synonyms for concepts in the Gene Ontology: 1) Importing synonyms from other biomedical resources through curated mappings, 2) using recursive rules to dissect GO concepts into their most basic forms then combining utilizing syntactic rules, and 3) using derivational variant rules to generate synonyms of the most basic composite terms. The overall results for all methods performance on CRAFT can be seen in Table 3.6 with more detailed analysis of each of methods following. More details about how we evaluated performance of each method can be seen in *Evaluation of generated synonyms.*

Table 3.6: Micro-averaged results for each synonym generation method on the CRAFT corpus.

Method	TP	FP	FN	Precision	Recall	F-measure
Baseline (B1)	10,778	6,280	18,669	0.632	0.366	0.464
Baseline (B2)	12,217	7,367	17,230	0.624	0.415	0.498
All external synonyms	12,747	11,682	16,704	0.522	0.433	0.473
Recursive syntactic rules	12,411	7,587	17,036	0.621	0.422	0.502
Recursive syntactic and derivational rules	18,611	10,507	10,836	0.639	0.632	**0.636**

We present two different baselines for comparison: 1) B1, a dictionary containing only the information within in the GO obo file and 2) B2, a dictionary that takes into account a known property of molecular function terms to counteract the separation of the protein and the function of the protein; for example, for term “GO:0016787 - hydrolase activity” a synonym of “hydrolase” is added.

The best results are obtained by using both syntactic recursive and derivational rules; an increase in F-measure of 0.112 is seen (0.610 vs 0.498). This increase is the result of a large increase in recall (0.225) with only a modest decrease in precision (0.049). Examining the overall performance we find that all synonyms generation methods perform better than B1, while all but the external synonyms perform better than B2. Overall, all generation methods trade precision for recall, which is to be expected when adding synonyms. We now discuss each method of synonym generation individually.

3.4.3 Mappings between ontologies and other biomedical resources

The first source of synonyms we used were those linked directly to Gene Ontology concepts through manually curated external mappings. We imported synonyms from four different sources, the Brenda database (Enzyme Commission numbers), UniProt knowledgebase, UniProt subcellular localization, and Wikipedia. Each of these resources contains classes or entities that are manually assigned indexes to identical, similar, or related GO terms (http://geneontology.org/page/download-mappings). It is noted by the GO Consortium that the mappings should not be taken as exact or complete. For both UniProt and Wikipedia, the official mappings were downloaded from the GO mapping website while Brenda was accessed through the provided SOAP server; all synonyms linked to the GO concepts were added as synonyms and re-evaluated on the CRAFT corpus.

The results of each external synonym source, broken down by sub-branch of GO, can be seen in Table 3.7. Overall, for the CC and MF branches, we find that the baselines provide the best performance because of a large decrease in precision (P) without a corresponding increase in recall (R) when using external mappings. With respect to the BP branch, we find a slight, 0.01, improvement in overall performance when using synonyms from UniProt. This slight improvement comes from a 0.03 increase (483 more true positives) in R and a 0.03 decrease (1,438 more false positives) in P. An error analysis was performed on the many false positives introduced through using external mappings but unfortunately no systematic method to improve or filter them was discovered (data not shown). Overall, based upon these results, external mappings introduce significantly more errors than correctly recognized concepts and are not suggested to be useful as a whole for concept recognition.

With more analysis, it is possible that filters could be created to reduce the false positives before external synonyms are inserted into the dictionary, but we leave that towards future

Table 3.7: Results for each external mapping source on the CRAFT corpus.

Cellular Component

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	5,532	452	2,822	0.925	0.662	**0.772**
Baseline (B2)	X	X	5,532	452	2,822	0.925	0.662	**0.772**
Brenda (EC)	0	0	5,532	452	2,822	0.925	0.662	**0.772**
UniProt	348	330	5,547	709	2,807	0.887	0.664	0.759
Wikipedia	210	210	5,519	1,014	2,835	0.845	0.661	0.742
All Combined	471	419	5,534	1,271	2,820	0.813	0.662	0.730

Molecular Function

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	337	146	3,843	0.698	0.081	0.145
Baseline (B2)	X	X	1,772	964	2,408	0.648	0.424	**0.512**
Brenda (EC)	22,158	2,870	1,768	2,773	2,412	0.389	0.423	0.406
UniProt	111	105	1,773	2,608	2,407	0.404	0.424	0.414
Wikipedia	31	31	1,772	2,666	2,408	0.399	0.424	0.411
All Combined	22,258	3,006	1,773	3,015	2,411	0.370	0.424	0.395

Biological Process

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	4,909	5,682	12,004	0.464	0.290	0.357
Baseline (B2)	X	X	4,913	5,951	12,000	0.452	0.291	0.354
Brenda (EC)	0	0	4,913	5,951	12,000	0.452	0.291	0.354
UniProt	361	346	5,392	7,120	11,521	0.431	0.319	**0.367**
Wikipedia	343	338	4,969	6,227	11,944	0.444	0.294	0.354
All Combined	660	600	5,440	7,396	11,473	0.424	0.322	0.366

All Gene Ontology

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	10,778	6,280	18,669	0.632	0.366	0.464
Baseline (B2)	X	X	12,217	7,367	17,230	0.624	0.415	**0.498**
Brenda (EC)	22,158	2,870	12,213	9,176	17,234	0.571	0.415	0.480
UniProt	720	781	12,712	10,437	16,735	0.549	0.432	0.483
Wikipedia	584	579	12,260	9,907	17,187	0.553	0.416	0.475
All Combined	23,389	4,025	12,747	11,682	16,704	0.522	0.433	0.473

work. Additionally, we are aware that there are many other sources of external mappings that each need to be examined individually to evaluate their usefulness as synonyms.

3.4.4 Recursive syntactic rules

The idea behind the recursive rules is to decompose a larger term to its smallest components, then compositionally combine the components utilizing varying recursive syntactic rules to generate synonyms for the original term. The recursive rules were developed through studying the grammars used in *Obol* (Mungall, 2004) and examining common patterns within Gene Ontology concepts. The syntactic recombination and common phrase enumeration rules were obtained by studying the gold standard annotations in CRAFT and through discussion with biologists on variations in the expression of the same concept. We have identified 11 specific cases when terms can be broken down into smaller composite terms; we acknowledge that there are many more, but we focused on the ones that affected the highest number of concepts, and leave the rest for future work. Through our analysis we have developed an ordering for rule application, to generate the most possible synonyms. The 11 cases and the order in which they are applied are presented below;

We name our rules based upon the type of concepts they apply to. The first step in all these rules is to decompose the concept further by making sure the left and right hand side do not match any other rules. When no more rules are matched, syntactic synonyms are generated and then compositionally combined. The words or phrases on the left hand side of the concept are now referred to as *LHS* and words on the right hand side are referred to as *RHS*; these can be replaced by both generated synonyms and the current synonyms in the ontology.

1. Terms containing prepositions
    1.1. if preposition is “via”
        1.1.1. *LHS* via conversion to *RHS*
    1.2. if preposition is “involved in”
        1.2.1. *LHS* associated *RHS*
2. “regulation of” terms
    2.1. if preceded by “positive”
        2.1.1. positive regulation of *RHS*
        2.1.2. up(-)regulation of *RHS*
        2.1.3. activation of *RHS*
        2.1.4. *RHS* activation
        2.1.5. stimulation of *RHS*
        2.1.6. *RHS* stimulation
        2.1.7. promotion of *RHS*
        2.1.8. promote *RHS*
        2.1.9. *RHS* promotion
        2.1.10. induction of *RHS*
        2.1.11. *RHS* induction

        - 2.1.12. enhancement of *RHS*
        - 2.1.13. enhance *RHS*
        - 2.1.14. *RHS* enhancement
    - 2.2. if preceded by “negative”
        - 2.2.1. negative regulation of *RHS*
        - 2.2.2. down(-)regulation of *RHS*
        - 2.2.3. *RHS* down regulation
        - 2.2.4. anti(-)*RHS*
        - 2.2.5. repression of *RHS*
        - 2.2.6. *RHS* repression
        - 2.2.7. inhibition of *RHS*
        - 2.2.8. *RHS* inhibition
        - 2.2.9. suppression of *RHS*
        - 2.2.10. suppress *RHS*
        - 2.2.11. *RHS* suppression

- 3. “response to” terms
    - 3.1. if only *RHS*
        - 3.1.1. response to *RHS*
        - 3.1.2. *RHS* response
    - 3.2. if both *LHS* and *RHS*
        - 3.2.1. *LHS* response to *RHS*
        - 3.2.2. *RHS* responsible for *LHS*
        - 3.2.3. *RHS* resulting in *LHS*
    - 3.3. if *RHS* is an ion
        - 3.3.1. *RHS*(-)responsive
        - 3.3.2. *RHS*(-)response
        - 3.3.3. *RHS* sensitivity
        - 3.3.4. *RHS* resistance
        - 3.3.5. *RHS* ion(-)responsive

3.3.6. *RHS* ion(-)response

3.3.7. *RHS* ion sensitivity

3.3.8. *RHS* ion resistance

4. “signaling” terms

4.1. if *LHS* contains “receptor” *RHS* equals “pathway”

4.1.1. *LHS* pathway

4.1.2. *LHS* signaling

4.1.3. *LHS* signalling

4.1.4. *LHS* signaling pathway

4.1.5. *LHS* signalling pathway

4.1.6. *LHS* signaling process

4.1.7. *LHS* signalling process

4.1.8. *LHS* receptor signaling

4.1.9. *LHS* receptor signalling

4.1.10. *LHS* receptor signaling process

4.1.11. *LHS* receptor signalling process

4.1.12. *LHS* receptor pathway

4.1.13. *LHS* receptor signaling pathway

4.1.14. *LHS* receptor signalling pathway

4.2. if *RHS* equals “patway”

4.2.1. *LHS* pathway

4.2.2. *LHS* signaling

4.2.3. *LHS* signalling

4.2.4. *LHS* signaling pathway

4.2.5. *LHS* signalling pathway

4.2.6. *LHS* signaling process

4.2.7. *LHS* signalling process

4.2.8. *LHS* receptor signaling

4.2.9. *LHS* receptor signalling

        4.2.10. *LHS* receptor signaling process

        4.2.11. *LHS* receptor signalling process

        4.2.12. *LHS* receptor pathway

        4.2.13. *LHS* receptor signaling pathway

        4.2.14. *LHS* receptor signalling pathway

5. "biosynthetic process" terms

    5.1. if both *LHS* and *RHS*

        5.1.1. *LHS* biosynthesis *RHS*

        5.1.2. *LHS* biosynthesis pathway *RHS*

        5.1.3. biosynthesis of *LHS RHS*

        5.1.4. *LHS* synthesis *RHS*

        5.1.5. synthesis of *LHS RHS*

        5.1.6. *LHS* production *RHS*

        5.1.7. *LHS* production pathway *RHS*

        5.1.8. production of *LHS RHS*

        5.1.9. *LHS* generation *RHS*

        5.1.10. generation of *LHS RHS*

        5.1.11. *LHS* formation *RHS*

        5.1.12. formation of *LHS RHS*

    5.2. if only *LHS*

        5.2.1. *LHS* biosynthesis

        5.2.2. *LHS* biosynthesis pathway

        5.2.3. biosynthesis of *LHS*

        5.2.4. *LHS* synthesis

        5.2.5. synthesis of *LHS*

        5.2.6. *LHS* production

        5.2.7. *LHS* production pathway

        5.2.8. production of *LHS*

        5.2.9. *LHS* generation

        - 5.2.10. generation of *LHS*
        - 5.2.11. *LHS* formation
        - 5.2.12. formation of *LHS*
    - 5.3. if only *RHS*
        - 5.3.1. biosynthesis *RHS*
        - 5.3.2. biosynthesis pathway *RHS*
        - 5.3.3. synthesis *RHS*
        - 5.3.4. production *RHS*
        - 5.3.5. generation *RHS*
        - 5.3.6. formation*RHS*

6. "metabolic process" terms
    - 6.1. if both *LHS* and *RHS*
        - 6.1.1. *LHS* metabolism *RHS*
        - 6.1.2. metabolism of *LHS RHS*
    - 6.2. if only *LHS*
        - 6.2.1. *LHS* metabolism
        - 6.2.2. metabolism of *LHS*
    - 6.3. if only *RHS*
        - 6.3.1. metabolism *RHS*
    - 6.4. if only "metabolic process"
        - 6.4.1. metabolism

7. "catabolic process" terms
    - 7.1. if both *LHS* and *RHS*
        - 7.1.1. *LHS* catabolism *RHS*
        - 7.1.2. catabolism of *LHS RHS*
        - 7.1.3. *LHS* degradation *RHS*
        - 7.1.4. degradation of *LHS RHS*
        - 7.1.5. *LHS* breakdown *RHS*
        - 7.1.6. breakdown of *LHH RHS*

- 7.2. if only *LHS*
  - 7.2.1. *LHS* catabolism
  - 7.2.2. catabolism of *LHS*
  - 7.2.3. *LHS* degradation
  - 7.2.4. degradation of *LHS*
  - 7.2.5. *LHS* breakdown
  - 7.2.6. breakdown of *LHH*
- 7.3. if only *RHS*
  - 7.3.1. catabolism *RHS*
  - 7.3.2. degradation *RHS*
  - 7.3.3. breakdown *RHS*
- 7.4. if only “catabolic process”
  - 7.4.1. catabolism

8. “binding”
   - 8.1. if *LHS* and *RHS* equals “complex”
     - 8.1.1. complex that bind *LHS*
   - 8.2. if only *LHS*
     - 8.2.1. binding of *LHS*
     - 8.2.2. binds *LHS*
     - 8.2.3. if *LHS* contains “receptor”
       - 8.2.3.1. *LHS*(-)binding receptor

9. “transport” terms
   - 9.1. if *LHS* contains “transmembrane” and *RHS* equals “activity”
     - 9.1.1. *LHS* transporter
     - 9.1.2. transporter of *LHS*
     - 9.1.3. transporting *LHS* transmembrane
     - 9.1.4. transporting *LHS* across a membrane
     - 9.1.5. transporting *LHS* across the membrane
     - 9.1.6. transportation of *LHS* transmembrane

        - 9.1.7. transportation of *LHS* across a membrane
        - 9.1.8. transportation of *LHS* across the membrane
        - 9.1.9. *LHS*
    - 9.2. *LHS* and *RHS* equals “activity”
        - 9.2.1. *LHS* transporter
        - 9.2.2. transporter of *LHS*

- 10. “differentiation” terms
    - 10.1. if only *LHS*
        - 10.1.1. differentiation into *LHS*
        - 10.1.2. differentiation into *LHS* cell
    - 10.2. if *LHS* is found within Cell Ontology, grab all synonyms, *CLSYNS*
        - 10.2.1. differentiation into *CLSYNS*
        - 10.2.2. differentiation into *CLSYNS* cell
        - 10.2.3. *CLSYNS* differentiation
- 11. “activity” terms
    - 11.1. if comma after “activity”
        - 11.1.1. *LHS* - *RHS*
        - 11.1.2. *LHS* that *RHS*
        - 11.1.3. *RHS LHS*
    - 11.2. if only *LHS*
        - 11.2.1. *LHS*

3.4.4.1 Example of recursive syntactic rules applied

A decomposed concept along with varying syntactic rules applied can be seen in Figure 3.4. We begin with the original term at the top. In this example it is “GO:0032332 - positive regulation of chondrocyte differentiation”. As it is run through the specific rules in the order listed above, the first one it matches is the *regulation of* rule. This causes the term to be decomposed into two parts, 1) “positive regulation of” and 2) “chondrocyte differentiation”, the latter is another GO term. This causes the recursive synonym generation process to start on the new term. The rules are followed in order the new term “GO:0002062 - chondrocyte differentiation" and the first one it matches is the *differentiation terms* rule. This term can then be decomposed into two smaller concepts: 1) chondrocyte and 2) differentiation. Neither of these are GO terms therefore, we are finished with the recursive breakdown and begin combining the pieces using syntactic rules. "Chondrocyte" is a cell type and we can link the word to the cell type ontology (Bard et al., 2005) through derivational variant synonyms, as described in the next section. There are many different syntactic ways that "X differentiation" can be expressed, which are listed in the figure; there are 2 synonyms generated for "chondrocyte" and 2 synonyms generated for "X differentiation". When we combinatorially combine them we generate 4 synonyms of "chondrocyte differentiation" (i.e. "differentiation into chondrocyte"). We then combine those 4 synonyms with the 17 different ways enumerated to express "positive regulation of", resulting in 68 synonyms for the original term "positive regulation of chondrocyte differentiation". This example utilized 3 specific decompositional rules along with the syntactic rules to re-compose the original concept (full enumeration is in Additional File 3).

Briefly exploring the literature, we have identified instances of these newly generated synonyms in a paper on disease "ACVR1 R206H receptor has a milder **stimulation of cartilage cell differentiation** compared to caACVR1 Q207D." (Shore, 2012) and, interestingly, within a patent "The DNA of the present invention can be used in the antisense RNA/DNA technique or the triple helix technique to inhibit type II collagen expression promotion and/or **chondrocyte differentiation promotion** mediated by the protein of the present invention." (Goichi et al., 2003). From this example, some of the 'odd' synonyms generated, that we hypothesized wouldn't affect performance, could be applicable to certain sublanguages, like the large collection of biomedical patents.

3.4.4.2 Impact of recursive syntactic rules on CRAFT

We apply only the recursive syntactic rules to all terms within the Gene Ontology and evaluate on the full-text CRAFT corpus using our dictionary based lookup system ConceptMapper; performance can be seen in Table 3.8. For Cellular Component, only a few new synonyms are generated, which is not surprising because concepts from this branch

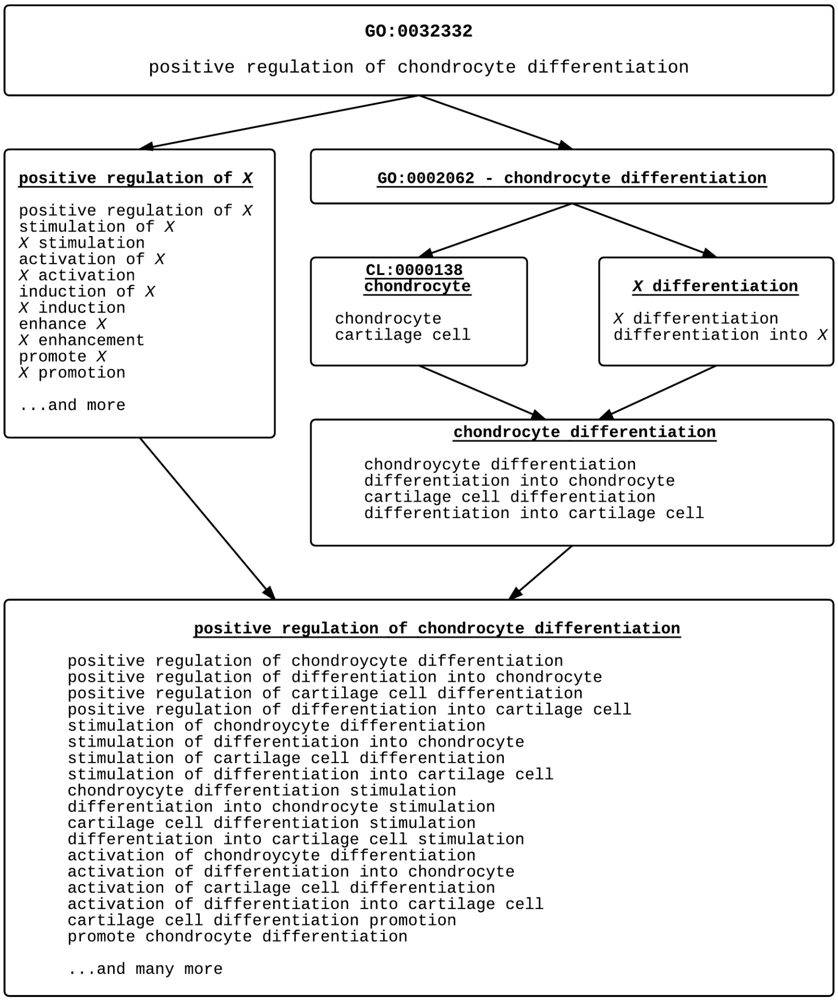

GO:0032332
positive regulation of chondrocyte differentiation
positive regulation of X
positive regulation of X
stimulation of X
X stimulation
activation of X
X activation
induction of X
X induction
enhance X
X enhancement
promote X
X promotion
...and more
GO:0002062 - chondrocyte differentiation
CL:0000138
chondrocyte
chondrocyte
cartilage cell
X differentiation
X differentiation
differentiation into X
chondrocyte differentiation
chondroycyte differentiation
differentiation into chondrocyte
cartilage cell differentiation
differentiation into cartilage cell
positive regulation of chondrocyte differentiation
positive regulation of chondroycyte differentiation
positive regulation of differentiation into chondrocyte
positive regulation of cartilage cell differentiation
positive regulation of differentiation into cartilage cell
stimulation of chondroycyte differentiation
stimulation of differentiation into chondrocyte
stimulation of cartilage cell differentiation
stimulation of differentiation into cartilage cell
chondroycyte differentiation stimulation
differentiation into chondrocyte stimulation
cartilage cell differentiation stimulation
differentiation into cartilage cell stimulation
activation of chondroycyte differentiation
activation of differentiation into chondrocyte
activation of cartilage cell differentiation
activation of differentiation into cartilage cell
cartilage cell differentiation promotion
promote chondrocyte differentiation
...and many more

Figure 3.4: Decomposition and syntactic synonym generation of a biological process. A single GO concept broken down into its composite parts (bolded and underlined), synonyms generated for each part (text underneath the part), then combination of all synonyms from all composite parts to form complete synonym of the original concept.

normally do not appear compositional in nature. These new CC synonyms do not impact performance compared to the baselines.

86% (7,353 out of 8,543) of terms within Molecular Function had at least one new synonym added by the recursive syntactic rules. Unexpectedly, performance on MF slightly decreases; this occurs when a true positive in the baseline is converted to a false positive and false negative(s) are introduced because a longer term is identified through one of the new synonyms. It is possible that these are missing annotations within the gold standard. For example, one of the new synonyms generated for “GO:0019838 - growth factor binding” is “binding growth factor”. In the gold standard, the text text “bound growth factor” is annotated with both “GO:0005488 - binding” and “GO:0008083 - growth factor activity”. With our new synonyms added to the dictionary, the same text span is only annotated with “GO:0019838 - growth factor binding” which results in the removal of two true positives and the introduction of one false positive, thus reducing overall performance. If we recognize this as a wide-spread issue, we can change the parameters for our dictionary and allow it to find all concepts, which would identify all three annotations instead of only the longest one.

Unlike Molecular Function, the performance on Biological Process slightly increases with the addition of the recursive syntactic rules. BP sees the largest increase in the number of new synonyms generated, with over 180,000 new synonyms for 46% (6,847 out of 14,767) of concepts. The syntactic and enumerated rules are helpful in generating synonyms that match instances within CRAFT. For example, 74 more correct instances of “GO:0016055 - Wnt receptor signaling pathway”, expressed in the gold standard as “Wnt signaling” and “Wnt signaling pathway”, are able to be identified with the new synonyms; these are generated through the *signaling terms* rule which found that both the words “receptor” and “pathway” were uninformative. Another example of syntactic rules helping is in identification of the term “GO:0046686 - response to cadmium ion”, which is seen 14 times in CRAFT as “cadmium response” and “cadmium-responsive”. Like the MF false positives, some of the FPs introduced in the BP look accurate. For example, a synonyms of “helicase activation” is generated for term “GO:0051096 - positive regulation of helicase activity”, which is seen in the text spans “The extent of **helicase activation** depends on the sequence context of the 3‘-tail. . . ” and “. . . replacement of thymines with guanines abolished the **helicase activation**.”. Some of the rules introduce FPs that are obviously incorrect. e.g. from “GO:0032781 - positive regulation of ATPase activity” a synonym of “ATPase activation” is generated; due to our dictionary lookup algorithm utilizing a stemmer, the text “ATPase activities” is incorrectly annotated with “positive regulation of ATPase activity”. Additionally, similar false positives are introduced for both “GO:0009896 - positive regulation of catabolic process” (catabolic activation) and “GO:0009891 - positive regulation of biosynthetic process” (anabolism activation).

Overall, despite the decrease in performance in the Molecular Function branch, the recursive syntactic rules slightly improve concept recognition of the Gene Ontology on the CRAFT corpus over baseline 2 (~200 more TPs and ~200 more FPs introduced). Because the CRAFT corpus contains only a small portion of the whole GO (1,108) and these rules only account for reordering of tokens and enumeration of common phrases within GO, we did not expect to see a large increase in concept recognition performance.

Table 3.8: Performance of Gene Ontology syntactic recursion rules on CRAFT corpus.

Cellular Component

Method	**Synonyms added**	**Affected terms**	**TP**	**FP**	**FN**	**P**	**R**	**F**
Baseline (B1)	X	X	5,532	452	2,822	0.925	0.662	**0.772**
Baseline (B2)	X	X	5,532	452	2,822	0.925	0.662	**0.772**
Syntactic recursion rules	23	21	5,532	452	2,822	0.925	0.662	**0.772**

Molecular Function

Method	**Synonyms added**	**Affected terms**	**TP**	**FP**	**FN**	**P**	**R**	**F**
Baseline (B1)	X	X	337	146	3,843	0.698	0.081	0.145
Baseline (B2)	X	X	1,772	964	2,408	0.648	0.424	**0.512**
Syntactic recursion rules	11,637	7,353	1,759	977	2,421	0.643	0.421	0.509

Biological Process

Method	**Synonyms added**	**Affected terms**	**TP**	**FP**	**FN**	**P**	**R**	**F**
Baseline (B1)	X	X	4,909	5,682	12,004	0.464	0.290	0.357
Baseline (B2)	X	X	4,913	5,951	12,000	0.452	0.291	0.354
Syntactic recursion rules	182,617	6,847	5,120	6,158	11,793	0.454	0.303	**0.363**

All Gene Ontology

Method	**Synonyms added**	**Affected terms**	**TP**	**FP**	**FN**	**P**	**R**	**F**
Baseline (B1)	X	X	10,778	6,280	18,669	0.632	0.366	0.464
Baseline (B2)	X	X	12,217	7,367	17,230	0.624	0.415	0.498
Syntactic recursion rules	194,277	14,221	12,411	7,588	17,036	0.621	0.422	**0.502**

3.4.5 Derivational variant rules

Once the original concept is broken down to its most basic components, through the rules presented above, we can apply derivational variant generation rules to help generate synonyms beyond what is expressed within the Gene Ontology. There are a total of 7 different specific cases when we apply these derivational generation rules. These rules were developed by examining the transformations needed to create the text spans annotated in the CRAFT gold standard from the information contained within the GO.

We follow similar naming trends for the derivational rules. For these rules we do not substitute phrases, but rather generate derivational variants for individual words. We slightly modify the terminology since we are assured these will be the most basic concepts, we substitute *X* for capturing specific words and we can additionally specify which word in the concept we change *w1* or *w2* for the first and second word, respectively. Any of these can be substituted for the base form of the word, most likely a noun, or any of the derivational variants generated through WordNet (Fellbaum, 1998a) or lexical variant generator (LVG) (of Medicine, 2012), adjective or verb.

1. Single word terms
   1.1. {NN} ⇒ {VB}
   1.2. {NN} ⇒ {JJ}
2. Double word terms
   2.1. {NN_1 NN_2} ⇒ {NN_1}, {VB_2 NN_1}, {JJ_1 NN_2}, {NN_1 JJ_2}
      2.1.1. *w2* of *w1*
      2.1.2. *w2* of a(n) *w1*
      2.1.3. if *w2* equals “development”
         2.1.3.1. specific development term ending in -genesis or -ization
      2.1.4. if *w2* is not a broad functional category(“binding”, “transport”, “secretion”,etc. . .)
         2.1.4.1. *w1*
   2.2. {NN_1 JJ_2} ⇒ {NN_1}
      2.2.1. if *w2* equals “complex” and *w1* is not one of “mediator”, “receptor”, or “integrator”
         2.2.1.1. *w1*
      2.2.2. if *w2* equals “complex” and *w1* equals “immunoglobulin”

2.2.2.1. antibody

2.2.2.2. antibodies

2.2.2.3. Ab

2.3. {JJ_1 NN_2} ⇒ {JJ_1}, {JJ_1 JJ_2}

2.3.1. *w1 w2*

2.3.2. if *w2* equals “perception”, “response”, “region”, “process” and *w1* does not equal “cellular”

2.3.2.1. *w1*

3. Triple word terms

3.1. {NN_1 NN_2 NN_3} ⇒ {NN_1 NN_3}, {NN_3 NN_1}, {VB_3}

3.1.1. *w3* of *w1 w2*

3.1.2. if *w2* equals “cell” or “nerve” and *w3* equals “morphogenesis” or “development”

3.1.2.1. *w1 w3*

3.1.2.2. *w3* of *w1*

3.1.3. if *w1* equals “cell” and *w3* does not equal specific terms associated with cells such as “site”, “determination”, “formation”, “assembly”, etc…

3.1.3.1. *w3*

3.1.3.2. *w1 w3*

3.1.3.3. *w3* of *w1*

4. “cell part” terms

4.1. if concept has parent of “cell part” or “organelle part”, *RHS* corresponds to specific part of cell

4.1.1. *LHS RHS*

4.1.2. *RHS* of *LHS*

5. “sensory perception” terms

5.1. generate other forms of *w4*, i.e. “taste”⇒“gustory”

6. “transcription, *X*-dependent” terms

6.1. *X*(-)reverse transcription

6.2. *X*(-)RT

6.3. *X*(-)dependent reverse(-)transcription

6.4. *X*(-)dependent RT

6.5. if *X* equals RNA

6.5.1. reverse(-)transcription

6.5.2. RT

7. “*X* strand annealing activity” terms

7.1. *X* annealing

7.2. *X* hybridization

7.3. annealing

7.4. hybridization

3.4.5.1 Examples of derivational rules applied

In Figure 3.5 we walk through the synonyms generation process through recursive decomposition and derivational variant generation of “GO:00507678 - negative regulation of

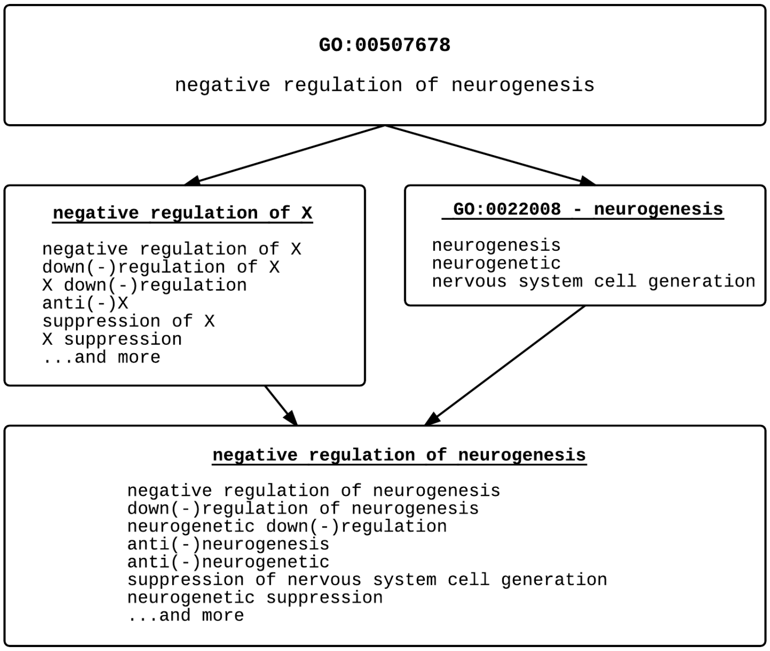

GO:00507678
negative regulation of neurogenesis
negative regulation of X
negative regulation of X
down(-)regulation of X
X down(-)regulation
anti(-)X
suppression of X
X suppression
...and more
GO:0022008 - neurogenesis
neurogenesis
neurogenetic
nervous system cell generation
negative regulation of neurogenesis
negative regulation of neurogenesis
down(-)regulation of neurogenesis
neurogenetic down(-)regulation
anti(-)neurogenesis
anti(-)neurogenetic
suppression of nervous system cell generation
neurogenetic suppression
...and more

Figure 3.5: Syntactic and derivational synonyms generation example. A single GO concept broken down into its composite parts (bolded and underlined), synonyms generated for each part (text underneath the part), then combination of all synonyms from all composite parts to form complete synonym of the original concept.

neurogenesis”. It is first decomposed into “positive regulation of” and “GO:0022008 - neurogenesis”. We cannot decompose any of these terms further, so we begin to generate synonyms then combinatorially combine all synonyms. We generate derivational variants for the term “neurogenesis” utilizing the *single word term* rule; we see if any verb or adjective forms exist in WordNet (Fellbaum, 1998a) or can be generated through LVG (lexical variant generator) (of Medicine, 2012). We take the three ways to express “neurogenesis” and combine them with the 12 different enumerations of “negative regulation of” to form 36 synonyms for the original term; the Gene Ontology currently only has 4 synonyms for this concept. It is important to generate and include derivational variants because many times a stemmer/lemmatizer is not sufficient for dictionary lookup. In this example, using the Porter stemmer (Porter, 1980) different stems are produced depending on if the noun or adjective form are stemmed: “neurogenesis”⇒“neurogenesi” and “neurogenetic”⇒“neurogenet”. These new generated synonyms are found throughout the biomedical literature many times; we identify mentions of “negative regulation of neurogenesis” within “This suggests that TNF-α would in fact have an **anti-neurogenetic** effect when allowed to act on a functional receptor.” (Van der Borght et al., 2011) and “The COX-2 inhibitors, meloxicam and nimesulide, **suppress neurogenesis** in the adult mouse brain” (Goncalves et al., 2010).

In Figure 3.6 we provide another example of utilizing our rules for a more complex and difficult to recognize term, “GO:0061005 - cell differentiation involved in kidney development”. The original concept is first decomposed through the recursive *terms containing preposition* rule; both sides of the prepositional phrase, “involved in”, will be decomposed further if possible. The left hand side, “GO:0030154 - cell differentiation”, can be decomposed using the syntactic *differentiation terms* rule and other synonyms will be generated using the *double word* derivational rule; the first three synonyms listed in the example are from “differentiation” while the last three are generated from derivations of the words “cell” and “differentiation”. The right hand side, “GO:0001822 - kidney development”, synonyms are generated solely from the *double word* rule. All synonyms generated for the left and right side are compositionally combined with the varying ways to express “involved in” to generate 60 synonyms for the original concept.

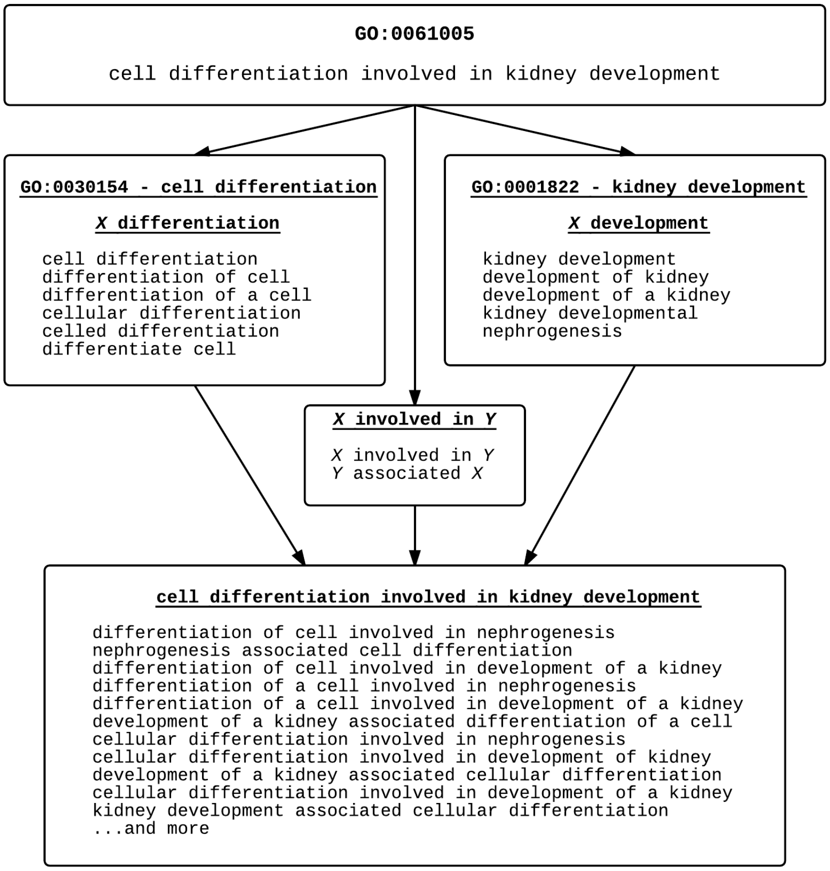

GO:0061005
cell differentiation involved in kidney development
GO:0030154 - cell differentiation
X differentiation
cell differentiation
differentiation of cell
differentiation of a cell
cellular differentiation
celled differentiation
differentiate cell
GO:0001822 - kidney development
X development
kidney development
development of kidney
development of a kidney
kidney developmental
nephrogenesis
X involved in Y
X involved in Y
Y associated X
cell differentiation involved in kidney development
differentiation of cell involved in nephrogenesis
nephrogenesis associated cell differentiation
differentiation of cell involved in development of a kidney
differentiation of a cell involved in nephrogenesis
differentiation of a cell involved in development of a kidney
development of a kidney associated differentiation of a cell
cellular differentiation involved in nephrogenesis
cellular differentiation involved in development of kidney
development of a kidney associated cellular differentiation
cellular differentiation involved in development of a kidney
kidney development associated cellular differentiation
...and more

Figure 3.6: Syntactic and derivational synonyms generation example. A single GO concept broken down into its composite parts (bolded and underlined), synonyms generated for each part (text underneath the part), then combination of all synonyms from all composite parts to form complete synonym of the original concept.

Unfortunately, none of the generated synonyms or original term can be recognized within the literature because of the constrained syntax contained in the original concept. Although, we believe that the newly generated synonyms can be helpful in creating new rules for synonyms generation and could be useful for concept recognition utilizing semantic similarity, such as GOCat (Gobeill et al., 2013b). For example, the following sentences express the concept "cell differentiation involved in kidney development": "A question of central importance is whether **differentiation into different cell types occurs during** the earliest stages of **nephrogenesis**" (Bacallao and Fine, 1989) and "Canonical Wnt9b signaling balances **progenitor cell** expansion and **differentiation during kidney development**" (Karner et al., 2011). Examining these sentences, it appears that we can express “involved in” as “during”. This new expression would be added to the *terms with prepositions* rule and be applied to all other terms that share that pattern. Even though we generate these synonyms, they are unable to be identified in the literature and therefore will not hinder the performance of the system. We’ve shown how the derivational rules can lessen the gap between the ontology and concepts expressed in text through examples and sentences directly from the literature. There is certainly more literature to be analyzed and rules to be crafted to help generate synonyms that will more likely appear in the literature.

Table 3.9: Performance of derivational variant and recursive Gene Ontology rules on CRAFT corpus.

Cellular Component

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	5,532	452	2822	0.925	0.662	0.772
Baseline (B2)	X	X	5,532	452	2822	0.925	0.662	0.772
Both Rules	4,083	724	6,585	969	1,769	0.872	0.788	**0.828**

Molecular Function

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	337	146	3,843	0.698	0.081	0.145
Baseline (B2)	X	X	1,772	964	2,408	0.648	0.424	0.512
Both Rules	14,413	7,401	2,422	1,074	1,758	0.693	0.579	**0.631**

Biological Process

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	4,909	5,682	12,004	0.464	0.290	0.357
Baseline (B2)	X	X	4,913	5,951	12,000	0.452	0.291	0.354
Both Rules	272,535	8,675	9,604	8,464	7,309	0.532	0.568	**0.549**

All Gene Ontology

Method	Synonyms added	Affected terms	TP	FP	FN	P	R	F
Baseline (B1)	X	X	10,778	6,280	18,669	0.632	0.366	0.464
Baseline (B2)	X	X	12,217	7,367	17,230	0.624	0.415	0.498
Both Rules	291,031	16,800	18,611	10,507	10,836	0.640	0.632	**0.636**

3.4.5.2 Impact of derivational rules on CRAFT

The performance of both recursive and derivational synonym generation rules on CRAFT can be seen in Table 3.8 (Note that we do not evaluate the derivational rules on their own, due to dependencies on the concepts being decomposed fully). When aggregated over the entire Gene Ontology, an increase in F-measure of 0.14 (0.498 vs. 0.636) is seen; this comes from both an increase in recall (0.22) and precision (0.02). Our rules generate ∼291,000 new synonyms which cover 66% (16,800 out of 25,471) of all terms within GO. Evaluating all branches individually, we see an increase in F-measure for all. This increase is due to a large increase in recall (up to 0.27). For both Biological Process and

Molecular Function, precision also increases, while precision slightly decreases for Cellular Component.

We now explore which synonyms contribute to the increase in performance seen on the gold standard corpus. The top 5 concepts that impact these performance numbers are presented in 3.10. For Cellular Component, the most helpful synonym "immunoglobulin"⇒"antibody" is seem many times within CRAFT and is enumerated within the *double word* rule. The other four are generated using the *single word* rule, specifically converting from the noun form seen within the ontology to the adjective form. Through examining Molecular Function terms, it became clear that "hybridize" and "anneal" did not have adequate representation within the Gene Ontology; this is changed using the *annealing* rule. Two of the next most helpful synonyms are due to low information containing words and derivational variations. Through analysis, it was identified that "protein" can be omitted within terms to produce correct annotations; for some of these it appears that more false positives are introduced, but with more work we can refine our current rules or create filters to remove some of the common false positives. It should be noted that within Molecular Function an even larger increase in in performance is seen between baseline 1 and 2 (Table 3.8), which takes into account the many "activity" terms. These types of synonyms are also accounted for in our rules and are compositionally combined into other terms. For Biological Process we observe that the most helpful synonyms are generated using the *double word* and *single word* derivational rules. Like the word "protein" from MF, "gene" and "cell" occur with numerous terms in BP and therefore, contain low information content and can most likely be omitted. We also find that generating different lexical forms of both single word concepts and within longer terms helps to introduce many true positive annotations.

From examining the top most helpful synonyms, we provide evidence that the derivational synonyms improve performance on a manually annotated corpus through the introduction of more linguistic variability, which decreases the gap between the concepts in the ontology and the way they are expressed in natural language. These variants are also included when generating the compositional synonyms mentioned in the previous section. Overall, the top synonyms that improve performance are not too interesting by themselves because they don't take into account much of the compositional nature of GO terms. We

Table 3.10: The top 5 derivational synonyms that improve performance on the CRAFT corpus. The GO terms that increase performance the most on CRAFT are along with the change(Δ) in number of true positives(TP), false positives(FP), and false negatives(FN) from the the baseline. The generated synonyms that result in this increase are shown under 'Helpful synonyms'.

Cellular Component

GO ID	**Term name**	**ΔTP**	**ΔFP**	**ΔFN**	**Helpful synonyms**
GO:0019814	immunoglobulin complex	+548	+0	-548	antibody, antibodies
GO:0005634	nucleus	+218	+35	-218	nuclear, nucleated
GO:0005739	mitochondrion	+135	+0	-135	mitochondrial
GO:0031982	vesicle	+11	+3	-11	vesicular
GO:0005856	cytoskeleton	+15	+0	-15	cytoskeletal

Molecular Function

GO ID	**Term name**	**ΔTP**	**ΔFP**	**ΔFN**	**Helpful synonyms**
GO:0000739	DNA strand annealing activity	+327	+1	-327	hybridized, hybridization, annealing, annealed
GO:0033592	RNA strand annealing activity	+327	+1	-327	hybridized, hybridization, annealing, annealed
GO:0031386	protein tag	+6	+79	-6	tag
GO:0005179	hormone activity	+1	+0	-1	hormonal
GO:0043495	protein anchor	+1	+10	-1	anchor

Biological Process

GO ID	**Term name**	**ΔTP**	**ΔFP**	**ΔFN**	**Helpful synonyms**
GO:0010467	gene expression	+2235	+361	-2235	expression, expressed, expressing
GO:0007608	sensory perception of smell	+445	+1	-445	olfactory
GO:0008283	cell proliferation	+97	+71	-97	cellular proliferation, proliferative
GO:0007126	meiosis	+93	+2	-93	meiotic, meiotically
GO:0006915	apoptosis	+173	+2	-173	apoptotic

believe this is due to two aspects 1) The annotation guidelines used to define what constitutes a correct mention of a GO concept in CRAFT (Bada et al., 2010) and 2) the small representation of what is contained within the entire biomedical literature. This small representation is due to the paper content (only mouse papers resulting in functional annotation of at least one protein), small corpus size (67 full text documents), and appearance of only a small subsection of the Gene Ontology (only 1,108 unique GO terms). To fully evaluate our rules without the aforementioned drawbacks, in the next section, we explore the impact our rules make on a large collection of the biomedical literature.

3.4.6 Evaluation of annotations on a large full text collection

Besides the intrinsic evaluation presented above, we evaluated the impact that both syntactic decompositional and derivational rules have on the ability to recognize GO concepts within a large collection of one million full text documents. Unlike the previous evaluation, these documents do not have any manual annotation or markup of Gene Ontology terms so we are unable to calculate precision/recall/F-measure. However, we can provide calculate descriptive statistics and provide manual evaluation of a random sample of the differences in annotations produced when our rules are applied. For these we used a version of GO from Oct 9th, 2014. Applying our rules generates ∼1.5 million new synonyms for 66% of all GO concepts (27,610 out of 41,852).

Table 3.11: Statistics of annotations produced on the large literature collection by information content. Shows the number of unique terms and total number of annotations produced through only OBO information, both derivational and syntactic recursive rules applied, and the impact the rules have overall. The change is percent change in total annotations.

	Only OBO information		With rules		Impact of rules on concepts recognized		
IC	# Terms	# Annotations	# Terms	# Annotations	New concepts	New Annotations	Change
undefined	3,548	16,929,911	4,303	23,653,066	755	6,723,155	+39.7%
[0,1)	7	3,202,114	7	3,177,333	0	-24,781	-0.1%
[1,2)	16	2,655,365	17	2,801,431	1	146,066	+0.1%
[2,3)	43	7,332,003	44	8,016,573	1	684,570	+0.1%
[3,4)	94	4,474,422	101	5,188,968	7	714,546	+0.2%
[4,5)	178	4,185,438	191	9,340,757	13	5,155,319	+123.8%
[5,6)	354	13,547,423	373	22,284,670	19	8,737,247	+64.4%
[6,7)	666	9,533,940	715	12,060,499	49	2,526,559	+26.3%
[7,8)	1,044	18,354,299	1,154	21,251,834	110	2,897,535	+16.8%
[8,9)	1,465	7,932,937	1,648	15,316,476	183	7,383,539	+92.4%
[9,10)	1,551	4,813,153	1,813	7,671,601	262	2,858,448	+58.3%
[10,11)	1,396	2,390,061	1,690	4,291,831	294	1,901,770	+79.1%
[11,12)	942	1,246,758	1,162	2,279,005	220	1,032,247	+83.3%
[12,13)	732	578,501	953	1,257,956	221	679,455	+117.2%
Total	12,036	97,176,325	14,171	138,592,000	2,135	41,415,675	+42.5%

Since one of the primary focuses of the Gene Ontology is functional annotation of proteins, we imparted some of that knowledge into the large scale analysis by calculating information content of each concept with respect to the experimental UniProt GOA annotations (Camon et al., 2004). We calculated the information content (IC) described in Resnik *et al.* (Resnik, 1995). IC scores range from 0-12.25; a lower score corresponds to a term that many proteins are annotated with and should appear many times in the literature while a high scoreing term is much more specific and might have only one or two annotations in GOA. For example, a common term such as "GO:0005488 - binding" has a score of 0.80 while a more informative term "GO:0086047 - membrane depolarization during Purkinje myocyte cell action potential" has a score of 12.25. A score of "undefined" corresponds to a concept that is not currently annotated to any protein with GOA. It is our hypothesis that the most informative terms (higher IC) would be more difficult to identify text due to their complicated syntactic construction and that our rules, described above, would help increase the frequency at which we can recognize correct mentions of these highly informative terms.

Descriptive statistics for both the concepts recognized using the ontology (baseline 2 presented above) and rules applied along with the differences broken down by information content can be seen in Table 3.11. Utilizing only the information contained within the Gene Ontology we find that ∼97 million mentions of ∼12,000 unique GO concepts are identified. When we apply both the recursive syntactic and derivational rules, our system is able to identify ∼138 million mentions of ∼14,100 unique GO concepts; they aid in the recognition of ∼41 million more mentions for all GO concepts (∼42% increase) along with the ability to recognize ∼2,000 unique GO concepts (∼18% increase) that are not previously identified using the ontology information alone. There were a total of ∼2.5 million mentions associated with the 2,135 unique concepts that were only found when the synonym generation rules were applied. The other ∼39 million new mentions are associated with the ∼12,000 concepts both dictionaries recognize.

Next, we show that our rules aid most in recognition of those concepts that are highly informative for functional annotation. We find that the distribution of mentions per concept is skewed. When our rules are applied there are an average of 9,700±14,500 mentions per concept (without rules 8,000±14,500). For example, the top 10 concepts, in terms of counts, represent ∼1/3 of the total number of mentions. The term "GO:0005623 - cell" is found around 13 million times and despite its higher IC score of 7, contains little information; this unexpected and high IC score is due to not many proteins being annotated in GOA with this concept. Other concepts found many times, but containing much lower IC scores, are "GO:0010467 - gene expression" (5 million), "GO:0004872 - receptor activity" (2.8 million), and "GO:0032502 - developmental process" (2.6 million). While we keep all concepts and instances of them for this analysis, for practical applications, these highly recognized concepts should be discarded or weighted using a frequency based metric.

Examining the overall numbers of concepts and mentions recognized provides some insights into how useful the synonyms generated are. Since most mentions identified using only the ontology information were also found when the rules were applied, this indicates that our rules aid in identification of many new concepts along with new mentions of concepts, thus leading to an increase in recall. We saw in evaluation on CRAFT that both precision and recall were increased; we explore through manual validation the accuracy on a large scale in the following section, *Manual validation of Gene Ontology mentions.* An exception to this is the concept “GO:0005215 - transporter activity”, where our generation rules identify ∼75,000 fewer instances, but instead are able to recognize more specific types of transporters and their activity. For instance, in the following sentence, the bold text corresponds to the concept recognized using the baseline, while the underlined text is identified through the use of our rules: “The present study was aimed to evaluate whether intraperitoneal carnitine (CA), a **transporter** of fatty acyl-CoA into the mitochondria. . . .” (Rajasekar and Anuradha, 2007). This illustrates that expressing GO terms in language that resembles natural language more closely helps to capture more specific concepts in the text. This suggests their potential usefulness for practical applications such as protein function prediction (Sokolov et al., 2013b).

Table 3.12: Results of manual inspection of random samples of annotations. Accuracy of random subsets of concepts recognized from the large literature collections. We sampled 1% of concepts, with up to 15 randomly sampled specific text spans per concept, from concepts identified using only OBO information. We sampled 10% of concepts, with up to 15 randomly sampled text spans per concept, from the new concepts recognized through the presented synonym generation rules.

	Only OBO information			With rules			Overall
IC	# Terms	# Annotations	Accuracy	# Terms	# Annotations	Accuracy	Accuracy
undefined	35	231	0.98	75	363	0.70	0.81
[0,1)	1	15	0.20	0	0	0.00	0.20
[1,2)	1	15	1.00	1	4	1.00	1.00
[2,3)	1	15	1.00	1	4	1.00	1.00
[3,4)	1	4	1.00	1	1	0.00	0.80
[4,5)	2	30	0.60	2	24	0.88	0.72
[5,6)	4	60	0.97	2	13	0.23	0.84
[6,7)	7	79	0.99	5	41	0.49	0.82
[7,8)	10	136	0.89	11	116	0.65	0.78
[8,9)	15	197	0.98	19	163	0.83	0.91
[9,10)	16	175	0.97	26	205	0.79	0.87
[10,11)	14	119	0.83	30	217	0.80	0.81
[11,12)	10	103	0.97	22	141	0.77	0.86
[12,13)	8	93	0.98	22	156	0.72	0.82
Total	125	1272	0.94	217	1448	0.74	0.83

3.4.6.1 Manual validation of Gene Ontology mentions

Although we found an improvement in performance on the CRAFT corpus and on the larger corpus a significant number of additional concepts and mentions were identified through our synonym generation rules, we are hesitant to reach any further conclusions without some manual validation of the accuracy of these introduced synonyms. There are too many concepts and annotations produced to manually validate them all, so we performed validation of a randomly distributed subset of concepts and instances of those concepts within text. For cases where the validity of the term was unclear from the matched term text alone we went back to the original paper and viewed the annotation in sentential context. For a baseline of performance, we validated a random sample of 1% of baseline concepts (125 concepts with ∼1,200 randomly sampled mentions) from each IC range and a random sample of 10% of all new concepts (217 terms with ∼1,450 randomly sampled mentions) recognized through our rules; these results are presented in Table 3.12. Examining the results, we find that overall accuracy is very high (0.94) for the concepts recognized only utilizing the ontology information. A majority of these text spans identified are exact, or very near, matches to the official ontological name or current synonyms. The only variation introduced is through a stemmer or lemmatizer used in the concept recognition pipeline (see Additional File 1 for more details). The annotations produced through our synonym generation rules do not have as high of accuracy (0.74) but still produce reasonable results. While performing the manual evaluation we noted common errors and explore them in the following section.

When describing the rules (in section *Examples of derivational rules applied*) we provide examples mostly having to do with regulation of a function. To give the reader more exposure to the breadth of the synonyms generated, we provide a few more examples of mentions with sentential context that would not be recognizable without our compositional rules. The concept "GO:0036473 - cell death in response to oxidative stress" is identified within the text "LLC-PK 1, a renal-tubular epithelial cell line, is susceptible to **oxidative stress, resulting in cell death** or injury" (Park and Han, 2013). Another concept, "GO:0097153 - cysteine-type endopeptidase activity involved in apoptotic process", would most likely never be seen exactly as it is represented within the ontology and the only synonyms are broad or narrow; these are not helpful. Our rules generate synonyms that aid in the identification of this concept within the following text: "Cytokines…can be particularly toxic, however, by initiating activation of **apoptotic-associated caspases** and production of reactive nitrogen species…" (Williams et al., 2005) and "…reported role of NO in the negative regulation of **caspase activities involved in apoptotic responses**, we hypothesize…" (authors, 2001). Additionally, for both sentences we are able to recognize the positive and negative regulation terms.

One conclusion reached from this manual validation addressed our previously mentioned hypothesis on overgeneration of synonyms. Based upon these results, we do not believe that the 1.5 million new synonyms generated introduce many false positives from overgeneration; a majority of the errors introduced come from the process of dictionary lookup. Synonyms that contain incorrect syntactic format and those that are not lexically sound do not appear within the text we are searching. An interesting observation we have made is that sometimes generating a phrase or synonym that initially appears incorrect due to using uncommon forms of words. An example is the different adjective forms of "protein"; most would use the form "proteinaceous", but another form is generated through Lexical Variant Generator, "protenic". This appears multiple times within translated articles, for example, the concept "GO:0042735 - protein body" is seen within the following sentence "The activity is exhibited through a **protenic body** of NBCF…" (Miwa, 1990).

3.4.6.2 Error analysis of Gene Ontology mentions

There were three main types of errors introduced by our synonym generation rules. Some of these are also seen in the baseline evaluation. We explain and provide examples of each type then re-evaluate after a simple fix.

1. Stemming/lemmatization creating incorrect concepts
2. Incorrect level of specificity due to information loss
3. Inclusion of incorrect punctuation

The most common type of error makes up 60% (225 out of 377) of all errors and is introduced through stemming during the concept recognition step. One of our syntactic recursive rules, *regulation of* terms, has a syntactic variant of *X activation*, which keeps the same semantics as the original concept. The error is introduced by using a stemmer within our concept recognition pipeline, that has been shown to increase performance (Funk et al., 2014a), but because both the words "activation" and "activity" stem to "activ" there are many incorrect spans of *X activity* normalized to the *positive regulation* concept identifier. For example, our rules add a synonym of "collagen binding activation" to the concept "GO:0033343 - positive regulation of collagen binding". Because of the stemmer, many spans of "collagen binding activity" are grounded to GO:0033343, which is incorrect. "activation"⇒"activity" makes up a majority of the errors, but we also find the text span "importance of glycine" grounded to "GO:0036233 - glycine import" due to the rule generated synonym "import of glycine". These errors could be removed by not including the stemmer in the dictionary lookup or employing a stemmer that handles these words in a more linguistically sensitive manner. It is unclear what other effects that would have on the other concepts identified. We plan on exploring the impact of different parameter combinations with these new synonym generation rules.

The second most common type of error, making up 25% (95 out of 377), is due to synonyms being generated at differing levels of specificity; this can occur during information loss or recursively using narrow/broad defined synonyms within the ontology. Currently our rules treat all synonyms for a concept the same (this can be changed quite easily) and when these other types of related synonyms are incorporated through the recursive syntactic rules they can introduce varying levels of specificity. For example, the text spans in the large corpus "anti-ryanodine receptor" and "inhibition of ryanodine receptors" are identified to be of concept "GO:0060315 - negative regulation of ryanodine-sensitive calcium-release channel activity". We believe that these errors are due to incorporating current synonyms of different specificities, i.e. broad or narrow synonyms within the Gene Ontology. All of these mentions are *related* to the concept identifier they are normalized to, but not an *exact* synonym. Many of these errors can be judged as partially correct; the use of a hierarchical precision and hierarchical recall metric in a comparison to a gold standard would give such partial credit (Verspoor et al., 2006; Clark and Radivojac, 2013; Bada et al., 2014).

The least common type of error seen at only 15% (57 out of 377) is due to incorrect punctuation being incorporated into the text span. These are most likely due to our tokenizer performing poorly in the parsing of particular situations. Some errors appear to come from information derived from tables. These could be removed through a post-processing filter – e.g. any mentions with semicolons or colons, unmatched parenthesis or quotes could be removed as these are often tokenized incorrectly. The more difficult types of punctuation to filter are those containing commas. For example, the concept "GO:2001170 - negative regulation of ATP biosynthetic process" is recognized within the sentence "These include cessation of **ATP synthesis, inhibition** of respiration, and a drop in $\Delta\Psi$." (Garlid et al., 2009); it is evident to a human that this is incorrect, but without the comma the span "ATP synthesis inhibition" appears to be correct. Since punctuation other than sentence boundaries is ignored during matching, such sentence can result in false positive matches.

To help reduce these errors we implemented two simple strategies. 1) Removed all text-spans that mentions containing unmatched parenthesis, semicolons, or colons and 2) removed the specific rule within the *regulation of* rule that generates the *X activation* synonyms; we refer to this new set of mentions produced as the *corrected* set. In total there were ∼850,000 erroneous mentions removed through these two observations. Additionally, we performed manual validation of the same concepts randomly sampled from the *corrected* mentions in the "With rules" column from Table 3.12. After manual re-validation we found that accuracy increased from 0.74 to 0.82 for these 217 concepts and increased from 0.83 to 0.88 when aggregated over all 342 concepts manually examined. Through error analysis we have shown that the accuracy of our rules can be improved using only very simple techniques, however, we believe we can achieve much higher accuracy through future work by incorporating syntactic parses along with more detailed analysis and refinement of the current rules.

3.4.7 The impact of supercomputing on concept recognition tasks

We ran the our concept recognition pipeline over the large full text collection on the Pando supercomputer located at the University of Colorado, Boulder campus. It has 60 – 64 core systems with 512GB each along with 4 – 48 core systems with 1TB ram each, for a total of 4,032 compute nodes. We utilized a quarter of the machine and ran our pipeline over 1,000 directories with 1,000 full text documents in each. We were able to produce GO annotations for all one million documents in around 10 minutes. Granted, no components are particularly complicated. They consist of a sentence splitter, tokenizer, stemmer/lemmatizer, followed by dictionary lookup, but we have performed similar tasks on a large memory machine, with 32 cores and the complete task has taken 3-4 weeks. Given that Pubmed consists of over 24 million publications, if it was possible to obtain all documents and performance is linear to the number of documents, we could recognize GO concepts from the entirety of the biomedical literature in around 4 hours. More complex and time consuming tasks, such as relation extraction, will take longer but will still be on the order of days or weeks utilizing the power of a supercomputer, since these tasks are "embarrassingly parallel".

3.4.8 Generalization to other biomedical ontologies

The methodology presented here, of breaking down complex concepts into their most basic parts, generating synonym for the parts, then recursively combining to form synonyms of the original concept is one that can generalize to many other ontologies. The Gene Ontology contains very complex and lengthy worded concepts; the rules required to implement compositional synonyms in other ontologies might not need as many syntactic and derivational rules as we present here.

A great example of an ontology that could have its synonyms extended by this methodology is the Human Phenotype Ontology (HPO) (Robinson et al., 2008). For example, there is a high level HPO term that corresponds to "phenotypic abnormality". There are just over 1,000 terms (~10% of all HPO concepts) that are descendants of "phenotypic abnormality" that can be decomposed into: "abnormality of [the] *other concept*" (e.g. HP:0000818 - abnormality of endocrine system). Not only can we add syntactic rules to reorder words, semantic synonyms of "abnormality", such as "malformation" or "deformity", can be added to express the concepts in similar ways. There are many other concepts that could benefit from recursively generating synonyms as the HPO appears to have compositional characteristics as well. There could also be subsets of rules depending on the context; recognizing concepts in doctor's notes or EMR notes will be expressed differently than those within the biomedical literature.

A great example of an ontology that could have its synonyms extended by this methodology is the Human Phenotype Ontology (HPO) (Robinson et al., 2008). For example, there is a high level HPO term that corresponds to "phenotypic abnormality". There are just over 1,000 terms (∼10% of all HPO concepts) that are descendants of "phenotypic abnormality" that can be decomposed into: "abnormality of [the] *other concept*". Not only can we add syntactic rules to reorder words, semantic synonyms of "abnormality", such as "malformation" or "deformity", can be added to express the concepts in similar ways. There are many other concepts that could benefit from recursively generating synonyms as the compositional nature of underlies HPO as well. There could also be subsets of rules depending on the context; recognizing concepts in doctor's notes or EMR notes will be expressed differently than those within the biomedical literature.

3.5 Conclusions

In this work, we present a set of simple language generation rules to automatically generate synonyms for concepts in the Gene Ontology. These rules take into account the compositional nature of GO terms along with manually created syntactic and derivational variants derived from discussions with biologists, ontologists, and through analyzing Gene Ontology concepts as they are expressed within the literature. The 18 hand-crafted rules automatically generate over ∼1.5 million new synonyms for ∼66% of all concepts within the Gene Ontology. The approach overgenerates synonyms, but we find that many such synonyms do not appear within biomedical text, thus not hindering performance.

We argue that current synonyms in structured ontologies are insufficient for text-mining due to the vast degree of variability of expression within natural language; our methods do not propose to solve this problem, but make a step in the right direction. This claim is supported through the examination of specific examples of concept variation in biomedical text and an empirical evaluation of the overlap of current GO synonyms and their expression in the CRAFT corpus.

We evaluate our synonym generation rules two both intrinsically and extrinsically. Utilizing the CRAFT corpus for intrinsic evaluation, we evaluate three different sources of automatically generated synonyms 1) external ontology mappings, 2) recursive syntactic rules and 3) derivational variant rules. External mappings introduce many false positives and are currently not recommended for use. The recursive syntactic rules added many new synonyms but did not significantly affect performance. Using a combination of recursive syntactic rules and derivational variant rules we saw an increase in F-measure performance of 0.14, mostly due to greatly increased recall. This illustrates the importance of derivational variants for capturing natural expression.

Our rules were extrinsically evaluated on a large collection of one million full text documents. The rules aid in the recognition of ∼2,000 more unique concepts and increase the frequency in which all concepts are identified by 41% over the baseline, using only current information contained within the Gene Ontology. Specifically, the synonyms generated aid in the recognize of more complex and informative concepts. Manual validation of random samples conclude accuracy is not as high as desirable (74%). An error analysis produced concrete next steps to increase the accuracy; simply removing one generation sub-rule, and filtering mentions with unmatched punctuation, increases accuracy of a random sample of 217 newly recognized concepts (∼1,400 mentions) to 82%. Overall, manual analysis of 342 concepts (∼2,700 mentions) leads to an accuracy of 88%. We find that our rules increase the ability to recognize concepts from the Gene Ontology by incorporating natural language variation.

Even though we chose a specific dictionary based-system, ConceptMapper, to evaluate our rules, the synonyms can also be useful for many other applications. Any other dictionary based system can supplement its dictionary with the generated synonyms. Additionally, any machine learning or statistical based methods will be able to utilize the synonyms we generate to try to normalize the span of text identified as a specific entity type to an ontological identifier; this will provide a richer feature representation for target concepts. In addition, we provide examples of how these rules could generalize to other biomedical ontologies and discuss the impact of supercomputing on scaling this work.

Not only have our rules proven to be helpful for recognition of GO concepts, but there are also other applications separate from the evaluated task. They could be used to identify inconsistencies within the current Gene Ontology synonyms. Concepts that share similar patterns, i.e. *regulation of X*, should all contain synonyms that correspond to a certain syntactic pattern. While performing this work we identified a few concepts that should contain synonyms but do not (Verspoor et al., 2009). Additionally, a certain conservative subset of our rules could easily be incorporated into TermGenie (Dietze et al., 2014), a web application that automatically generates new ontology terms. Our rules would be help to generate synonyms of the automatically generated concepts.

Not only are these rules presented and evaluated in this chapter, but we also apply and evaluate their impact on the automated function prediction task in Chapter VI.

CHAPTER IV

PHARMACOGENOMIC AND DISEASE GENE DISCOVERY FROM TEXT[4]

4.1 Introduction

This chapter marks a transition in my dissertation from focus on the task of concept recognition to the application of text-mining for biomedical predictions. In this chapter, I focus on the ability to predict pharmacogenes – genes where variants could affect drug efficacy and metabolism or disease related genes. To address the question of what information should be mined from the literature, I begin by exploring the use curated GO annotations along with simple features mined from the literature, such as, Gene Ontology concepts, bigrams, and collocations. I re-evaluate the original predictions and show that 6 of the top 10 hypothesized pharmacogenes in May 2013 now have curated support within PharmGKB, indicating that literature features are useful for making biologically insightful predictions.

4.2 Background

One of the most important problems in the genomic era is identifying variants in genes that affect response to pharmaceutical drugs. Variability in drug response poses problems for both clinicians and patients (Evans and Relling, 1999). Variants in disease pathogenesis can also play a major factor in drug efficacy (Poirier et al., 1995; Kuivenhoven et al., 1998). However, before variants within genes can be examined efficiently for their effect on drug response, genes interacting with drugs or causal disease genes must be identified. Both of these tasks are open research questions.

Databases such as DrugBank (Wishart et al., 2006) and The Therapeutic Target DB (Chen et al., 2002) contain information about gene-drug interactions, but only The Pharmacogenomics Knowledgebase (PharmGKB)(Hewett et al., 2002) contains information about how variation in human genetics leads to variation in drug response and drug pathways. Gene-disease variants and relationships are contained in Online Mendelian Inheritance in

[4] The work presented in this chapter is republished with permission from: *Combining heterogenous data for prediction of disease related and pharmacogenes* In Pacific Symposium on Biocomputing 19:328-339, 2014.

Man (OMIM) (Hamosh et al., 2005), the genetic association database (Becker et al., 2004), and the GWAS catalog (Hindorff et al., 2009). Curated databases are important resources, but they all suffer from the same problem: they are incomplete (Baumgartner et al., 2007b). One approach to this problem is the development of computational methods to aid in database curation; this type of systems needs high accuracy and should extract information from text and provide evidence to a curator. A different approach would be a system to provide hypotheses based upon current knowledge and would direct biologist possibly where to focus their next efforts. We present here a method that takes advantage of the large amount of information in the biomedical literature to make predictions over all sets of genes.

Having a classifier that is able to predict as-yet-uncurated pharmacogenes would allow researchers to focus on identifying the variability within the genes that could affect drug response or disease, and thus, shorten the time until information about these variants is useful in a clinical setting. (We use the term "pharmacogene" to refer to any gene such that a variant has been seen to affect drug response or is implicated in a disease.) Computational methods have been developed to predict the potential relevance of a gene to a query drug (Hansen et al., 2009). Other computational methods have been developed to identify genetic causes underlying disorders through gene prioritization, but many of these are designed to work on small sets of disease-specific genes (Aerts et al., 2006; Vanunu et al., 2010; Hutz et al., 2008; Chen et al., 2009; Tranchevent et al., 2008; Gonzalez et al., 2007). The method which is closest to the one that we present here is described in Costa *et al.*(Costa et al., 2010); they create separate classifiers to predict morbidity-associated and druggable genes on a genome-wide scale. A majority of these methods use sequence-based features, network topology, and other features from curated databases while only a few use information from literature (Aerts et al., 2006; Tranchevent et al., 2008; Gonzalez et al., 2007).

In the work presented here, the goal is to predict pharmacogenes at genome-wide scale using a combination of features from curated databases and features mined from the biomedical literature. We evaluate a number of hypotheses:

1. There is a set of GO concepts that are enriched when comparing the functions of important pharmacogenes and the rest of the human genome and by examining this set of enriched GO concepts, a classifier can be created to provide hypotheses regarding further genes in which variants could be of importance.

2. Text-mined features will increase performance when combined with features from curated databases.

4.3 Methods

4.3.1 Pharmacogenes

By *pharmacogene*, we mean any gene such that a variant of that gene has been seen to affect drug response or such that variants have been implicated in disease. PharmGKB contains over 26,000 genes, with only a few having annotations that signify their importance in disease or drug response. For the experiments reported here, only those genes in which a variant exists in the PharmGKB relationship database, specifically gene-disease or gene-drug relationships, are considered to be gold-standard pharmacogenes. By this definition, 1,124 genes meet the criteria for classification as pharmacogenes and are positively labeled training instances; these make up <5% of all genes in PharmGKB. PharmGKB is constantly being updated, so a snapshot of PharmGKB on May 2, 2013 was taken and is used as the gold standard.

4.3.2 Background genes

The rest of the 25,110 genes in PharmGKB, which do not contain disease or drug relationships, are considered to be background genes and will be used as negatively labeled training instances. We acknowledge the fact that PharmGKB is incomplete and that a missing annotation is not indicative of a gene not being involved in disease or drug relationships, but the fact that they have not been discovered or curated yet. (This is an obvious motivation for the work reported here.) Two data sets were created from the background genes. One consists of all 25,110 genes. This is referred to as the unbalanced set. The second consists of 1,124 background genes that have similar numbers of publications as the known pharmacogenes. This is referred to as the balanced set. That is, the two sets differ in whether or not they result in a balanced set of positive and negative exemplars.

4.3.3 Functional annotations from curated databases

Links within PharmGKB were used to obtain Entrez Gene (EG) identifiers for both pharmacogenes and background genes. To extract all Gene Ontology (GO) (Consortium, 2001) annotated functions associated with these genes, the NIH's gene2go file was used. Only curated evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, and ISS) were used, in order to ensure high-quality annotations. This dataset will be referred to as the curated dataset. It contains many EGID to GO ID mappings obtained solely from curated GO annotations.

4.3.4 Functional annotations from biomedical literature

Entez Gene IDs and the NIH's gene2pubmed file were used to relate genes to documents of which they are the primary subject. By using the gene2pubmed file, we assume that all information retrieved from the article is associated with the gene that is the primary subject. Note that this is not always true and could introduce noise.

The 26,234 genes are mapped to 379,978 unique PubMed/MEDLINE articles. From these ∼380,000 articles, two different textual datasets were created, one consisting only of abstracts and the other containing full text. The abstract dataset consists of all abstracts from all articles. For ∼26,000 articles, we were only able to download XML or plain text, because PMC articles are available in any format, with some, such as PDF, not being suitable for natural language processing. The ∼26,000 full-text articles constitute our full-text dataset. All full-text documents come from the PubMed Open Access Subset.

To extract gene functions (GO concepts) from these corpora, ConceptMapper, a dictionary-based concept recognizer (Tanenblatt et al., 2010), was used with parameters tuned for each branch of the Gene Ontology (Molecular Function, Biological Process, and Cellular Component), as seen in the previous chapter (Funk et al., 2014a). Descriptive statistics of the documents and the functional annotations retrieved from them and from the curated database are shown in Table 4.1.

Table 4.1: Summary of gene-document and gene-annotation associations. The number of genes within each dataset along with the mean number of biomedical literature documents associated with each **set of genes** and mean number of GO annotations per gene. (+) denotes that this set of genes is the positive labeled set while (–) denotes the negative training sets. The row labelled "Total Numbers" gives the count, not means, of documents and GO annotations.

		Mean # Docs		**Mean # GO Annotations**		
	# Genes	**Abstracts**	**Full-text**	**GOA curated**	**NLP abstracts**	**NLP full-text**
All genes	26,234	35.5	3.1	8.8	80.1	122.0
Known pharmacogenes (+)	1,124	215.2	15.5	16.3	227.5	220.7
All background genes (–)	25,110	26.7	2.5	8.2	72.8	128.7
Small background genes (–)	1,124	211.1	17.1	20.4	310.0	298.9
Total Numbers	26,234	379,978	25,987	112,356	1,891,566	1,951,982

4.3.5 Enrichment of Gene Ontology concepts

FatiGO (Al-Shahrour et al., 2004) was used to test whether there are functional concepts that are enriched when pharmacogenes are compared to background genes. FatiGO is a tool that uses Fisher's exact test to extract over- or under-represented GO concepts from two lists of genes and provides a list of enriched GO concepts and their respective p-values as output. The p-values are corrected for multiple testing as described in Ge *et al.*(Ge et al., 2003). The gene lists and all three sets of annotations—curated, and text-mined– were provided to FatiGO as custom annotations. Fisher's exact test was conducted between GO concepts annotated to pharmacogenes and those annotated to background genes for all three sets of Gene Ontology concepts (curated, mined from abstracts, and mined from full text).

4.3.6 Binary classification

All classifiers were implemented in the Weka toolkit, version 3.6.9. Three different baselines were used: OneR, a one node decision tree; Naive Bayes; and randomly assigning class labels. Against these, we compared three systems: Random Forests and two different Support Vector Machine implementations. Random Forests provide fast decision-tree training. Support Vector Machines (SVM) are currently the most popular classifier. The built-in classifiers for OneR (weka.classifiers.rules.OneR), Naive Byes (weka.classifiers.bayes.NaiveBayes), Random Forest (weka.classifiers.trees.RandomForest), and Support Vector Machine (weka.classifiers.functions.SMO) were used with default parameters. LibSVM (weka.classifiers.functions.LibSVM) was used with all but one default parameter. By default LibSVM maximizes accuracy; with the unbalanced dataset, this is not optimal, so weights of 90.0 and 10.0 were assigned to the pharmacogene and background classes, respectively. When using LibSVM with the balanced dataset, equal weights were given to both classes. All numbers reported are from five-fold cross-validation.

Table 4.2: Machine learning features per dataset. A breakdown of the type and number of unique features for each dataset.

Dataset	**# Genes**	**# Features**	**Type**
GOA curated	12,704	39,329	Curated GO annotations from the GOA database.
NLP abstract	23,849	39,329	GO annotations recognized from MEDLINE abstracts.
NLP full-text	15,168	39,329	GO annotations recognized from full-text journal articles.
Abstract GO + Bigrams	23,849	858,472	GO annotations and bigrams from MEDLINE abstracts.
Full-text GO + Bigrams	15,168	906,935	GO annotations and bigrams from full-text journal articles.
Combined GO + Bigrams	23,867	1,189,175	Curated and NLP GO annotations and all bigrams.
Abstract GO + Collocations	23,849	346,878	GO annotations and collocations from MEDLINE abstracts.
Full-text GO + Collocations	15,168	54,951	GO annotations and collocations from full-text journal articles.
Combined GO + Collocations	23,867	349,243	Curated and NLP GO annotations and all collocations.

4.3.7 Features derived from natural language processing

Additional features were extracted from the abstract and full-text document collections using natural language processing. (This is in addition to the automatically extracted Gene Ontology annotations, which are also produced by natural language processing.) These features were word bigrams and collocations. Collocations, or sets of words that co-occur more often than expected, have not been commonly used in text classification, but provide a better reflection of the semantics of a text than bigrams. Both bigrams and collocations were extracted using the Natural Language Tool Kit (NLTK)(Bird, 2006). Any bigram or collocation where one of the tokens only contained punctuation was removed. Additionally, only those features that appear in three or more documents were retained. Six different NLP-derived feature sets were created by combining the three datasets (abstract, full-text, curated + abstract + full-text) along with the two different types of surface linguistic features (bigrams and collocations); these feature sets were tested and trained on both the balanced and unbalanced datasets.

4.3.8 Machine learning input

A breakdown of the kind and number of features used in each dataset can be seen in Table 4.2.

4.3.9 Evaluation metrics

The performance of our classifier was assessed by estimating precision (P), recall (R), and F-measure (F). The area under the receiving operator characteristic curve (AROC) is reported, as it allows for comparison against other classifiers, but with a word of caution interpreting the unbalanced dataset: inflated AROCs have been seen when working with skewed class distributions (Kaymak et al., 2012). All scores were determined by taking the average of 5-fold cross-validation for all datasets.

4.4 Results and discussion

4.4.1 Enriched Gene Ontology concepts

To assess the viability of a machine learner separating background and pharmacogenes, we first determine whether functional differences between the pharmacogenes and background genes exist. At least one curated or text-mined functional annotation was retrieved for 23,647 out of 26,236 total genes (90% of all genes in PharmGKB). The details of obtaining the annotations are given in Sections 4.3.3 and 4.3.4. The gene sets and their annotations were passed to FatiGO, a web tool that extracts over- and under-represented GO concepts from two lists of genes, and a list of enriched GO concepts and probabilities was returned as output. Examining the output from FatiGO, we found that, depending on the dataset, between 800-4000 GO concepts were enriched, consistent with our hypothesis that there are enriched pharmacogenetic functions. The top 10 enriched GO concepts for Molecular Function and Biological Process can be seen in Tables 4.3 and 4.4, respectively. These lists were obtained by comparing the annotations from all pharmacogenes to all background genes. To ensure that bias was not introduced solely because there is a large difference in the number of genes and the number of annotations between the two sets, another comparison was done between all pharmacogenes and the set of 1,124 background genes with equal representation in the biomedical literature. The enriched GO concepts returned are similar the concepts returned when comparing against all background genes, and therefore we can conclude that no bias is introduced. Because there are many statistically enriched GO concepts returned for each dataset, we can conclude that there are functional differences between the set of pharmacogenes and background genes and provide a biological basis for the machine learner to be able to distinguish between the two sets.

Many of the enriched GO concepts can be categorized as playing a role in pharmacodynamics (PD) or pharmacokinetics (PK). Pharmacodynamics is the study of the activity of a drug in the body, e.g. its binding and effect on the body. Examples of PD concepts are “integral to plasma membrane” (GO:0005887), “drug binding” (GO:0008144), and “positive regulation of protein phosphatase type 2B activity” (GO:0032514)—they are either associated with receptors that drugs bind to, or refer to the possible effect that a drug has on the body. Pharmacokinetics is the study of drug absorption, distribution, metabolism, and excretion. Examples of PK concepts are “xenobiotic metabolic process” (GO:0006805), “small molecule metabolic process” (GO:0044281), and “active transmembrane transporter activity” (GO:0022804)—they refer to metabolism of a molecule or are involved in the metabolism or transportation of a molecule.

There are interesting differences when examining the top enriched concepts between the different datasets (curated, abstracts, and full text). Impressionistically, curated annotations seem to be more specific, while NLP annotations appear to be more general (especially evident when examining Biological Processes, Table 4.4). This may be the case because there are limitations to the depth in GO that concept recognizers can identify; a large gap exists between how near-terminal concepts are stated in the ontology and their expression in free text.

4.4.2 Classification of pharmacogenes

Having established that the functions of pharmacogenes are different from background genes, the next step is to test the ability of machine learning to differentiate between them. Our goal is to predict at genome-wide scale pharmacogenes that are not currently known in PharmGKB to have drug or disease relationships. We approach the problem as binary classification, where the classifier separates pharmacogenes from the rest of the genes.

Table 4.3: Top 10 enriched GO concepts from the Molecular Function hierarchy. The enriched GO concepts from the Molecular Function branch of Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

GOA curated		
Concept ID	**Concept name**	**Adj. P-value**
GO:0005515	protein binding	$< 1.0 \times 10^{-8}$
GO:0019899	enzyme binding	$< 1.0 \times 10^{-8}$
GO:0042803	protein homodimerization activity	$< 1.0 \times 10^{-8}$
GO:0046982	protein heterodimerization activity	$< 1.0 \times 10^{-8}$
GO:0004497	monooxygenase activity	$< 1.0 \times 10^{-8}$
GO:0005245	voltage-gated calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0020037	heme binding	$< 1.0 \times 10^{-8}$
GO:0004713	protein tyrosine kinase activity	$< 1.0 \times 10^{-8}$
GO:0004674	protein serine/threonine kinase activity	$< 1.0 \times 10^{-8}$
GO:0003677	DNA binding	$< 1.0 \times 10^{-8}$

NLP abstracts		
Concept ID	**Concept name**	**Adj. P-value**
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005322	low-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005321	high-density lipoprotein	$< 1.0 \times 10^{-8}$
GO:0005320	apoplioprotein	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0005041	low-density lipoprotein receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0004697	protein kinase C activity	$< 1.0 \times 10^{-8}$
GO:0045289	luciferin monooxygenase activity	$< 1.0 \times 10^{-8}$

NLP full-text		
Concept ID	**Concept name**	**Adj. P-value**
GO:0042031	angiotensin-converting enzyme inhibitor activity	$< 1.0 \times 10^{-8}$
GO:0005262	calcium channel activity	$< 1.0 \times 10^{-8}$
GO:0016088	insulin	$< 1.0 \times 10^{-8}$
GO:0022804	active transmembrane transporter activity	$< 1.0 \times 10^{-8}$
GO:0005179	hormone activity	$< 1.0 \times 10^{-8}$
GO:0004872	receptor activity	$< 1.0 \times 10^{-8}$
GO:0005215	transporter activity	$< 1.0 \times 10^{-8}$
GO:0016791	phosphatase activity	$< 1.0 \times 10^{-8}$
GO:0008083	growth factor activity	$< 1.0 \times 10^{-8}$
GO:0004601	peroxidase activity	$< 1.0 \times 10^{-8}$

4.4.3 Classification using Gene Ontology concepts

To see how well known pharmacogenes can be classified through their functional annotation similarity, five classifiers were created using the manually curated and text-mined functional annotations on both the unbalanced and balanced datasets. Baselines for comparison against are a one-node decision tree (OneR), Naive Bayes, and randomly assigning class labels. Performance of all classifiers and baselines can be seen in Table 4.5. A breakdown of features used for each dataset can be seen in Table 4.2 and a summary of functional annotations is seen in Table 4.1.

The results are shown in Table 4.5. A clear effect of balance versus imbalance in the data is evident. F-measure increases between 0.29 and 0.53 when using a balanced training set. Examining performance across unbalanced training sets, we notice that Naive Bayes

Table 4.4: Top 10 enriched GO concepts from the Biological Process hierarchy. The enriched GO concepts from the Biological Process branch of the Gene Ontology obtained when comparing pharmacogenes versus all background genes using FatiGO.

GOA curated		
Concept ID	**Concept name**	**Adj. P-value**
GO:0044281	small molecule metabolic process	$< 1.0 \times 10^{-8}$
GO:0007596	blood coagulation	$< 1.0 \times 10^{-8}$
GO:0030168	platelet activation	$< 1.0 \times 10^{-8}$
GO:0006805	xenobiotic metabolic process	$< 1.0 \times 10^{-8}$
GO:0048011	neurotrophin TRK receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007268	synaptic transmission	$< 1.0 \times 10^{-8}$
GO:0008543	fibroblast growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0007173	epidermal growth factor receptor signaling pathway	$< 1.0 \times 10^{-8}$
GO:0045087	innate immune response	$< 1.0 \times 10^{-8}$
GO:0055085	transmembrane transport	$< 1.0 \times 10^{-8}$

NLP abstracts		
Concept ID	**Concept name**	**Adj. P-value**
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0006629	lipid metabolic process	$< 1.0 \times 10^{-8}$
GO:0007610	behavior	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0014823	response to activity	$< 1.0 \times 10^{-8}$
GO:0006280	mutagenesis	$< 1.0 \times 10^{-8}$
GO:0042638	exogen	$< 1.0 \times 10^{-8}$

NLP full-text		
Concept ID	**Concept name**	**Adj. P-value**
GO:0009626	plant-type hypersensitive response	$< 1.0 \times 10^{-8}$
GO:0007568	aging	$< 1.0 \times 10^{-8}$
GO:0016311	dephosphorylation	$< 1.0 \times 10^{-8}$
GO:0032514	positive regulation of protein phosphatase type 2B activity	$< 1.0 \times 10^{-8}$
GO:0008152	metabolic process	$< 1.0 \times 10^{-8}$
GO:0009405	pathogenesis	$< 1.0 \times 10^{-8}$
GO:0042592	homeostatic process	$< 1.0 \times 10^{-8}$
GO:0046960	sensitization	$< 1.0 \times 10^{-8}$
GO:0006810	transport	$< 1.0 \times 10^{-8}$
GO:0050817	coagulation	$< 1.0 \times 10^{-8}$

produces the highest recall (0.68) but the lowest precision (0.17), whereas Random Forest produces highest precision (0.69) but lowest recall (0.11). The same trends do not hold for the balanced training sets. On both training sets, it is the SVM-based classifiers that balance precision and recall and produce the highest F-measures. The highest F-measures of 0.81 and 0.78, are produced by LibSVM and SMO, respectively, on the balanced NLP abstract annotations. Naive Bayes and Random Forrest perform poorly in comparison to the SVM classifiers, but better than a single-node decision tree or random assignment; OneR performs slightly better than random assignment.

For a majority of the classifiers, GO annotations from literature produce the best performance—surprisingly, text-mined annotations seem to be better features than those from curated datasets. One explanation could be that more information is encoded in text-mined annotations than just gene function. From this set of experiments, we can conclude that using only Gene Ontology concepts, we are able classify pharmacogenes on the balanced training set but it remains unclear, because of poor performance, whether it is sufficient to use only GO concepts with an unbalanced training set. We can also conclude that LibSVM should be used for the next set of experiments because it is best performing and was the fastest to train (training time not shown).

4.4.4 Classification using GO concepts and literature features

To test the hypothesis that features derived from surface linguistic features can increase performance over conceptual features alone, we trained classifiers with two additional feature types: bigrams and collocations. Bigrams consist of every sequence of two adjacent words in a document and are commonly used in text classification. Collocations are a subset of bigrams, containing words that co-occur more frequently than expected. They are a better representation of the semantics of a text than bigrams alone. The methods for extracting these features are described above in Section 4.3.7. Adding bigrams and collocations introduces up to 30x more features than functional annotations alone (Table 4.2).

The performance of LibSVM with GO annotations and bigrams/collocations on both training sets can be seen in Table 4.6. Baselines are the same.

Table 4.5: Classification using Gene Ontology concepts. Five-fold cross validation performance of five binary classifiers when providing Gene Ontology concepts as features. Results from both unbalanced and balanced training sets are shown. The highest F-measure is bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

	GOA curated	**NLP abstracts**	**NLP full-text**
Classifier	P/R/F	P/R/F	P/R/F
		Unbalanced Training	
Random	0.05/0.50/0.09	0.07/0.50/0.12	0.05/0.50/0.09
OneR	0.57/0.01/0.03	0.56/0.17/0.25	0.80/0.10/0.18
Naive Bayes	0.17/0.60/0.26	0.17/0.68/0.27	0.17/0.59/0.26
Random Forest	0.53/0.17/0.25	0.69/0.12/0.21	0.58/0.11/0.18
SMO	0.43/0.31/0.36	0.39/0.41/0.40	0.37/0.34/0.35
LibSVM	0.29/0.55/**0.38**	0.41/0.58/**0.48**	0.37/0.52/**0.42**
		Balanced Training	
Random	0.50/0.50/0.50	0.50/0.50/0.50	0/50/0.50/0.50
OneR	0.71/0.41/0.52	0.68/0.51/0.59	0.73/0.48/0.56
Naive Bayes	0.65/0.72/0.68	0.75/0.70/0.72	0.67/0.70/0.68
Random Forest	0.63/0.71/0.67	0.72/0.77/0.74	0.67/0.73/0.69
SMO	0.64/0.66/0.65	0.79/0.77/0.78	0.70/0.73/0.72
LibSVM	0.71/0.71/**0.71**	0.83/0.80/**0.81**	0.76/0.79/**0.78**

On the unbalanced training set, the maximum F-measure seen is 0.57, obtained by using text-mined functional annotations and bigrams extracted from abstracts. By using bigrams in addition to GO annotations, precision is increased by 0.17 while recall is decreased by 0.02, resulting in an increase in F-measure of 0.09 (Table 4.5 versus Table 4.6). On the balanced training set, the maximum F-measure seen is 0.81, also obtained by using text-mined functional annotations and bigrams from abstracts. With the addition of bigrams, both precision and recall are increased by 0.06 and 0.03,respectively, resulting in an increase in F-measure of 0.06 (comparing Table 4.5 to Table 4.6).

Table 4.6: Classification with GO concepts and natural language processing. Five-fold cross-validation performance of LibSVM when combining Gene Ontology concepts and literature-based features. Both the balanced and unbalanced training results are shown. The highest F-measure and AROC are bolded. The baselines provided are OneR (one-node decision tree), Naive Bayes, and randomly assigning classes (median of 5 random assignments).

	Abstract GO + Bigrams		**Full-Text GO + Bigrams**		**Combined GO + Bigrams**	
Classifier	**P/R/F**	**AUC**	**P/R/F**	**AUC**	**P/R/F**	**AUC**
Unbalanced Training						
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.58/0.56/**0.57**	**0.771**	0.50/0.46/0.48	0.711	0.50/0.54/0.52	0.756
Balanced Training						
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.75/0.59/0.66	0.696	0.71/0.53/0.61	0.663	0.79/0.50/0.61	0.685
LibSVM	0.89/0.83/**0.86**	**0.860**	0.79/0.82/0.80	0.807	0.86/0.83/0.85	0.848

	Abstract GO + Collocations		**Full-Text GO + Collocations**		**Combined GO + Collocations**	
Classifier	**P/R/F**	**AUC**	**P/R/F**	**AUC**	**P/R/F**	**AUC**
Unbalanced Training						
Random	0.07/0.50/0.12	0.501	0.05/0.50/0.09	0.501	0.05/0.50/0.09	0.499
LibSVM	0.54/0.56/**0.55**	**0.767**	0.41/0.52/0.46	0.730	0.47/0.56/0.51	0.763
Balanced Training						
Random	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500	0.50/0.50/0.50	0.500
OneR	0.78/0.46/0.58	0.664	0.67/0.64/0.66	0.675	0.75/0.59/0.66	0.698
LibSVM	0.87/0.82/**0.85**	**0.850**	0.77/0.80/0.78	0.786	0.85/0.81/0.83	0.833

4.4.4.1 Comparison with other methods

As mentioned in the introduction, there are very few methods against which our method can be compared. Most gene-disease or gene prioritization methods are designed to work on small sets of disease-specific genes (Aerts et al., 2006; Vanunu et al., 2010; Hutz et al., 2008), while our method was designed to predict pharmacogenes on a genome-wide scale. To obtain a completely fair analysis all systems would have to be trained and run over the same input data then predictions or prioritization ranking would need to be manually compared and validation. Instead, we summarize a few systems and report the performance numbers within the original publication.

One method, Garten *et al.* (Garten et al., 2010), utilizes text mining to extract drug-gene relationships from the biomedical literature, also using PharmGKB as a gold standard, with an AUC of 0.701. The closest methods to ours do not predict pharmacogenes as defined here, but only predict disease genes. CIPHER (Wu et al., 2008) predicts human disease genes with precision of ∼0.10 using protein-protein interaction networks and gene-phenotype associations. PROSPECTR (Adie et al., 2005) uses 23 sequence-based features and predicts disease genes from OMIM with precision = 0.62 and recall = 0.70 with an AUC of 0.70.The most directly comparable method, presented in Costa *et al.* (Costa et al., 2010), utilizes topological features of gene interaction networks to predict both morbidity genes (P=0.66, R=0.65, AUC=0.72) and druggable genes (P=0.75, R=0.78, AUC=0.82). While the majority of other methods utilize sequence-based features, protein interactions, and other genomic networks, our method requires only Gene Ontology annotations and simple bigrams/collocations extracted from biomedical literature. Precision and recall for our classifier trained on the unbalanced dataset with GO annotations and bigrams from abstracts are slightly lower than both PROSPECTR and the method presented in Costa *et al.*, our AUC (0.771) is higher than all but the predicted druggable genes from Costa *et al.* Performance on the balanced training set using GO concepts and bigrams extracted from abstracts (F=0.86, AUC=0.860) are higher than any of the other methods presented here.

4.4.4.2 Limitations

There are two major limitations of our work. The first is that we grouped together all pharmacogenes, while it may have been more useful to differentiate between disease-associated and drug-response-associated variant and even further subdivide genes that act through PD or PK. We hypothesize that these three subsets of genes will all have enriched GO concepts that are different from the other groups of pharmacogenes, by our definition. The other limitation is that we don't provide a ranking, but rather just a binary classification, but provide external systems to rank the predicted genes.

Table 4.7: Top 10 predicted pharmacogenes. Top 10 pharmacogenes predicted by all combined classifiers and ranked by functional similarity to the known pharmacogenes. All information from PharmGKB and OMIM is presented along with the class that was predicted by Costa *et al.*(Costa et al., 2010) (Morbid: mutations that cause human diseases, Druggable: protein-coding genes whose modulation by small molecules elicits phenotypic effects).

EG ID	Symbol	PharmGKB Annotations	OMIM Phenotype	Costa *et al.*(Costa et al., 2010) predicted
2903	*GRIN2A*	None	Epilepsy with neurodevelopment defects	Druggable
7361	*UGT1A*	None	None	Not tested
2897	*GRIK1*	None	None	Druggable
1128	*CHRM1*	None	None	Druggable
1131	*CHRM3*	Member of Proton Pump Inhibitor Pathway	Eagle-Barrett syndrome	Druggable
3115	*HLA-DPB1*	None	Beryllium disease	Morbid/Druggable
6571	*SLC18A2*	Member of Nicotine, Selective Serotonin Reuptake Inhibitor, and Sympathetic Nerve Pathway	None	Morbid/Druggable
477	*ATP1A2*	None	Alternating hemiplegia of childhood, Migraine (familial basilar and familial hemiplegic)	Morbid/Druggable
3643	*INSR*	Member of Anti-diabetic Drug Potassium Channel Inhibitors and Anti-diabetic Drug Repaglinide Pathways	Diabetes mellitus, Hyperinsulinemic hypoglycemia, Leprechaunism, Rabson-Mendenhall syndrome	Morbid/Druggable
2905	*GRIN2C*	None	None	Druggable

4.4.5 Prediction of pharmacogenes

Now that classifiers have been created and evaluated, we can analyze the predicted pharmacogenes. 141 genes were predicted to be pharmacogenes by all six unbalanced datasets seen in Table 4.6. Predictions from unbalanced models were analyzed because the models produced through balanced training were unknowingly weighted for recall. For example, the balanced model trained on abstract GO and bigrams produces a recall of 0.99 and precision of 0.10 when the classifier is applied to all genes in PharmGKB; this is not informative and further work and error analysis will be conducted to examine why this is.

The top 10 predicted genes, ranked by functional similarity (as calculated by ToppGene) to the known pharmacogenes, along with all known information from PharmGKB and Online Mendelian Inheritance in Man (OMIM)(Hamosh et al., 2005), and if/what the gene was predicted to be by Costa *et al.* can be seen in Table 4.7. We first notice that there are no gene-disease or gene-drug relationships in PharmGKB for these predicted genes, but a few of them participate in curated pathways. We expand our search to see if other databases have drug or disease information about them. OMIM provides insight into genetic variation and phenotypes; half of the predicted genes have a variant that plays a role in a mutant phenotype. We also looked up our predicted genes in the results from a previous study on predicting morbid and druggable genes, and 90% (9 out of 10) of our predicted pharmacogenes were also predicted to be morbid (variations cause hereditary human diseases) or druggable (Costa et al., 2010).

To assess the hypothesized pharmacogenes further, PubMed and STITCH (Kuhn et al., 2008) were used to find any known drug or disease associations not in PharmGKB or OMIM. The top-ranked gene, *GRIN2A*, seems to play a part in schizophrenia and autism spectrum disorders (Tarabeux et al., 2011) along with binding to memantine, a class of Alzheimer's medication blocking glutamate receptors. Interestingly, *UGT1A* is unable to be found in STITCH or OMIM, but an article from May 2013 introduces a specific polymorphism that suggests that it is an important determinant of acetaminophen glucuronidation and could affect an individual's risk for acetaminophen-induced liver injury (Freytsis et al., 2013). It is also known to be linked to irinotecan toxicity. We also find genetic variations in *GRIK1* have been linked to schizophrenia (Hirata et al., 2012) and down syndrome (Ghosh et al., 2012). Even only examining the top three predicted pharmacogenes, there is evidence in other databases and literature that suggests these should be further examined by the PharmGKB curators for possible annotation.

4.4.6 Re-analysis with new knowledge

The version of PharmGKB used in this work was from May 2013, since performing the original work, almost 2 years have passed and new curated annotations have accrued. This allows us to re-evaluate our predictions in light of this new knowledge. We present this re-analysis, using the February 2015 version of PharmGKB in Table 4.8. We find that 6 out of our top 10 predictions have at least one specific polymorphism that appears to affect drug efficacy or a corresponding disease, many have more than one. Granted, given enough time, variants within all genes will most likely have variants that affect drugs or diseases. Interestingly, for *UGT1A* utilizing the PubMed searches for validation, we were able to

Table 4.8: Top 10 predicted pharmacogenes re-analyzed taking into account new PharmGKB annotations. Top 10 pharmacogenes predicted by all combined classifiers and ranked by functional similarity to the known pharmacogenes. These annotations have been added since May 2013.

EG ID	**Symbol**	**Polymorphisms**	**Drugs**	**Disease**
2903	*GRIN2A*	None	None	None
7361	*UGT1A*	rs1042640, rs10929303, rs8330	acetoaminophen, ritonavir, atazanavir	Acute liver failure, HIV
2897	*GRIK1*	rs2832407	topiramate	Alcohol-related diseases
1128	*CHRM1*	None	None	None
1131	*CHRM3*	rs2155870	None	None Postoperative vomiting and nausea
3115	*HLA-DPB1*	rs1042136, rs1042151, rs3097671	aspirin	aspirin-induced asthma
6571	*SLC18A2*	rs1420, rs363224, rs363390, rs929493	antipsychotics, citalopram	None
477	*ATP1A2*	None	None	None
3643	*INSR*	None	None	None
2905	*GRIN2C*	rs8092654	exemestane, anastrozole	None

pinpoint the article that lead to at least one of the now curated annotations – the interaction with aspirin and acute liver failure. This re-evaluation shows that a pipeline consisting of hypothesis made through use of the biomedical literature and supporting evidence mined through a NLP pipeline designed could be provided to curators of databases to prioritize their efforts to specific genes and articles.

4.5 Conclusions

One of the surprising findings of this study was that features extracted from abstracts performed better than features extracted from full text. We believe this could be due to fact that abstracts are more concise and features obtained from them will more related to the proteins linked to the article; full text documents contain many other spurious sections and most likely will refer to many other proteins introducing noise. We experiment in the next chapter with limiting spans of mentions. Also, since full text was available for a smaller number of genes, the comparison may not be as meaningful or appropriate.

The fact that features derived from text-mined functional annotations outperformed manually curated annotations was a surprise. In this work, we did not evaluate the correctness of text-mined functional annotations. Therefore, the performance of the text-mined functional annotation features is the only indication of how well we are able to recognize Gene Ontology concepts. In Chapter II we found that baseline recognition (what was used in this work) was underwhelming. Based on the fact that they performed higher than the manually curated Gene Ontology concepts, it appears that the performance of the ConceptMapper approach was at minimum, good enough for this task. Incorporating the rules developed in Chapter III would most likely improve the ability to distinguish between the two sets; we explore the impact of the rules presented in Chapter III for a different task in Chapter VI.

In this work we identified a set of functions enriched in known pharmacogenes. This list could be used to rank genes predicted by our classifier, but also has usefulness beyond the work presented here. The list could prove useful in literature-based discovery by providing linkages to identify gene-drug or gene-disease relationships from disparate literature sources.

We also present a classifier that is able to predict pharmacogenes at a genome wide scale (F=0.86, AUC=0.860). The top 10 hypothesized pharmacogenes predicted by our classifier are presented; 50% contain allelic variations in OMIM and 90% were previously predicted but remain unannotated in PhamGKB. Additionally, using other sources at least the top three genes predicted are known to bind a drug or to be associated with a disease. Other methods attempting similar problems, utilize sequence based features and genomic networks; only a few incorporate literature features. Our method, on the other hand, uses mainly features mined from the biomedical literature along with functional annotations from databases.

We re-analyzed the top 10 hypothesized pharmacogenes in light of the knowledge gained in the 2 years since the original worked was completed and found that 60% had at least one variant now associated with a drug or disease. This supports the original findings, that text mined features from the biomedical literature can serve as useful features for predictions. We briefly explore one advantage that literature features have, their ability to not only serve as input features, but to offer supporting evidence or validation of the predictions. We dive much deeper and offer more evidence on this topic in Chapter VI.

CHAPTER V

PROTEIN FUNCTION PREDICTION USING LITERATURE FEATURES[5,6,7]

5.1 Introduction

With the cost of high-throughput methods decreasing every year, many gene or gene product sequences are deposited in databases with no known function. Manual annotation is expensive and lagging behind; the only feasible methods that are able to keep up with the growth in databases is through computational methods. Additional discussion of the problem and brief summary of common methods is presented in Section 1.3. There have been very few methods on incorporating the vast amount of information contained within the biomedical literature.

Most computational methods use features derived from sequence, structure or protein interaction databases (Radivojac et al., 2013); very few take advantage of the wealth of unstructured information contained in the biomedical literature. Because little work has been conducted using the literature for function prediction, it is not clear what type of text-derived information will be useful for this task or the best way to incorporate it. In this chapter I discuss the use of literature features for automated protein function prediction. In both this chapter and the following we evaluate literature features in the context of the machine learning framework, GOstruct (Sokolov and Ben-Hur, 2010). To address the question "what features mined from the literature are useful?", I build upon the features discussed in Chapter IV and explore more complex features, specifically co-mentions, mined from the biomedical literature. I explore the different ways to combine the multiple types of literature features and how to best integrate them with more commonly used sequence- or network-based features. Both these questions are evaluated through the use of our par-

The work presented in this chapter is republished with permission from:

[5] *Combining heterogeneous data sources for accurate functional annotation of proteins* BMC bioinformatics 14.Suppl 3 (2013): S10.

[6] *Improving automated protein function prediction by integrating natural language processing and machine learning* Poster at 9th Annual Rocky Mountain Bioinformatics Conference 2011, F1000Posters 2012, 3: 70 http://f1000.com/posters/browse/summary/1089821

[7] *Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct* Journal of Biomedical Semantics (*in press*).

ticipation in community challenges that focused on prediction of protein function, Critical Assessment of Function Annotation (CAFA). I describe both our methodology for each task and highlight the overall conclusions for the use of literature for function prediction.

5.1.1 GOstruct

Most machine learning function prediction methods combine the output of many binary classifiers answering the question “does this protein have function X?”, which results in training thousands of classifiers and then rectifying the many predictions within the Gene Ontology hierarchy. In contrast, GOstruct is a support vector machine learning framework that models predictions as a hierarchical multi-label classification task using a single classifier; when training and predicting it takes into account the entire GO structure and is able to predict the entire set of GO classes at once. For a more technical presentation, please refer to the original publication, Sokolov *et al* (Sokolov and Ben-Hur, 2010).

5.1.2 Co-mentions

We use the term *co-mention* to define the co-occurrence of any two entities within a predefined span of text. In this case we explore two different types of co-mentions: protein-protein co-mentions and protein-GO term co-mentions. We briefly experimented with GO-GO co-mentions but found they were not helpful; it could be the case that they would be useful in other contexts but not for input into function predication algorithms. We also used two different spans to identify co-mentions within this work, sentence and paragraph. For the case of abstracts, we equated the whole abstract as a paragraph and when dealing will full text documents we utilized the traditional notion of a paragraph. We explore using these because 1) they are very quick to extract from the literature and 2) are able to act as a proxy to actual relationships, without the cost of parsing. This allows us to extract this type of feature from large amounts of literature (10+ million abstracts, 1+ million full text) very quickly. As discussed in the Section 1.3, there are only a few other methods using literature for function prediction; the scale at which we mine the literature is one thing that sets this work apart.

5.2 Background

Literature mining has been shown to have substantial promise in the context of automated function prediction, although there has been limited exploration to date (Verspoor, 2014). The literature is a potentially important resource for this task, as it is well known that the published literature is the most current repository of biological knowledge and curation of information into structured resources has not kept up with the explosion in publication (Baumgartner et al., 2007a). A few teams from the first Critical Assessment of Functional Annotation (CAFA) experiments (Radivojac et al., 2013) used text-based features to support prediction of Gene Ontology (GO) functional annotations (The Gene Ontology Consortium, 2000).

Wong and Shatkay (Wong and Shatkay, 2013) was the only team in CAFA that used exclusively literature-derived features for function prediction. They utilized a k-nearest neighbor classifier with each protein related to a set of predetermined characteristic terms. In order to have enough training data for each functional class, they condensed information from all terms to those GO terms in the second level of the hierarchy, which results in only predicting 34 terms out of the thousands in the Molecular Function and Biological Process sub-ontologies. Recently, there has been more in-depth analysis into how to use text-based features to represent proteins from the literature without relying on manually annotated data or information extraction algorithms (Shatkay et al., 2014). This work explored using abstracts along with unigram/bigram feature representation of proteins.

Another team, Björne and Salakoski (Björne and Salakoski, 2011), utilized events, specifically molecular interactions, extracted from biomedical literature along with other types of biological information from databases; they focused on predicting the 385 most common GO terms.

In this work, we explore a variety of text-mined features, and different ways of combining these features, in order to understand better the most effective way to use literature features for protein function prediction.

5.3 CAFA 1

In 2011, the first Critical Assessment of Function Annotation (CAFA) was held to evaluate the accuracy of current function prediction methods (Radivojac et al., 2013). To

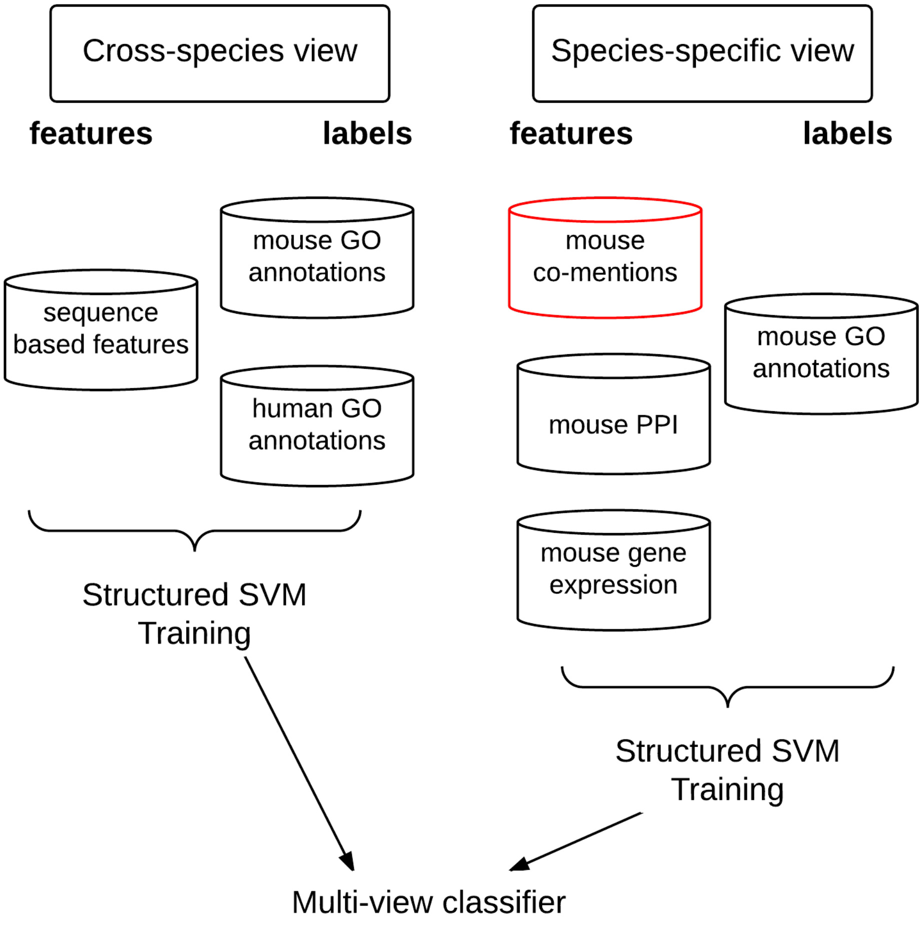


Figure 5.1: Multi-view approach to training. Data is separated into two different view: a cross-species view that contains features computed from sequence and a species-specific view which contains, in this figure, only features derived from mouse co-mentions, PPI, and gene expression. Both views are summed into the multi-view classifier. The red highlighted box are mined from the literature.

start, participants were provided ∼44,000 SwissProt proteins, from 11 different species, that contained no experimentally annotated function. A period of ∼6 months was provided to train and submit predictions of Molecular Function and Biological Process for all possible target proteins. After predictions were locked, experimental annotations were allowed to accrue for 11 months; the 866 proteins that had experimentally validated functions were the evaluation targets.

5.3.1 Features

An overview of the experimental setup used for making predictions along with the types of features used can be seen in Figure 5.1. GOstruct has two different views of data that it combines, the species-specific and cross-species. My contributions are the features mined from the literature using concept recognition (red box in Figure 5.1). I describe in detail the literature mined features extracted and for complete context, briefly cover the other types of sequence-based and cross-species features.

5.3.1.1 Literature features

Literature features were incorporated within the species-specific view; for this first experiment we only focused on mining features related to mouse proteins (MGI IDs). We extracted two different types of literature features, protein-protein co-mentions (PPC) and protein-GO co-mentions (PGC) within both sentences and documents. These features were extracted from a set of 11.7 million PubMed abstracts, all Medline abstracts on 9/8/2011 that had title and body text. The abstracts were fed into a natural language processing pipeline based on the BioNLP UIMA resources http://bionlp-uima.sourceforge.net/ which consists of the following steps: 1) splitting the abstracts into sentences 2) protein name tagging using the LingPipe named entity recognizer http://alias-i.com/lingpipe with the CRAFT model (Verspoor et al., 2012) 3) Gene Ontology term recognition via dictionary lookup (using ConceptMapper as described in Chapter II) and 4) extraction of the two types of co-mentions at the abstract and sentence level. Protein names were mapped to mouse MGI IDs using a MGI name dictionary lookup. Assuming only mouse references allowed us to avoid the full gene normalization problem (Morgan and Hirschmann, 2007) and fit in well with the other data sources of the species-specific classifier.

Counts of features extracted from all Medline can be seen in Table 5.1. We find that for both types of sentence co-mentions there is ∼10x increase when comparing unique verses total and for document co-mentions there is ∼100x increase. Having 100's of millions of co-mentions from such a large literature collection there will be some noise introduced. We hypothesize that good co-mention signal will rise from the noise. The most common co-mentions will likely have been documented. An extreme example of a common co-mention is *interleukin 6*, which is mentioned 426,031 times in conjunction with "interleukin-6 receptor binding”; this represents no new knowledge. The least common co-mentions, those mentioned very few times, either represent noise or brand new knowledge that is most likely not curated and accounted for from other sources. For training, the co-mentions were provided to GOstruct as frequency data – each protein is characterized by a vector that provides the number of times it co-occurs with each protein or GO term.

Table 5.1: Counts of mouse literature features extracted for CAFA 1. Counts of different types of co-mentions extracted from 11.7 million abstracts using an MGI and GO dictionary lookup. Document refers to co-mentions within the entire abstract, regardless of the number of paragraphs.

Features	Unique proteins	Unique GO terms	Unique co-mentions	Total co-mentions
PPC-sentence	3,588	–	211,543	4,174,302
PPC-document	3,654	–	350,417	37,145,104
PGC-sentence	3,721	9,345	696,141	18,713,932
PGC-document	3,738	11,591	1,392,033	180,142,251

5.3.1.2 Other features

Besides the species-specific literature features already discussed, there are multiple other types of biological features that are commonly used for function prediction[8]. I briefly touch on them to provide full context of the information provided for prediction of function and to compare the literature mined features against.

Cross-species features

These features are calculated from the protein sequence alone and are therefore not tied directly to a species. These features are where homology and the ‘transfer of annotation’ is encoded.

1. BLAST hits – Proteins are represented in terms of its sequence similarity to other proteins using the BLAST score against a database of annotated proteins (Altschul et al., 1990).

2. Localization signals – Sub cellular localization can offer insight into function as some processes are specific to certain compartments (Rost et al., 2003). WoLF PSORT (Horton et al., 2007) is used to identify these signals from sequence.

[8] Credit for other features and performing evaluation goes to students from Asa Ben-Hur’s lab: Artem Sokolov and Kiley Graim.

3. Transmembrane predictions – Certain functions such as "cell adhesion" or "transport of ions" tend to be associated with transmembrane proteins. The TMHMM program (Krogh et al., 2001) is used to establish how many possible transmembrane domains are in the protein.

4. Low complexity regions – These types of regions are abundant with proteins and have an effect on the protein's function (Coletta et al., 2010). A sliding window of 20 amino acids with the lowest diversity of amino acid sequence is used as a feature.

Species-specific features

Certain type of data that you cannot obtain for all species or it is not clear, for example, how frog PPI could directly apply to human, so there is a species-specific view. For these experiments, we create only mouse specific features.

1. Protein-protein interactions – *M. musculus* PPI were extracted from the STRING database (Szklarczyk et al., 2011).

2. Gene expression – Similarity of expression for ∼15,000 microarray experiments were provided by PILGRIM (Greene and Troyanskaya, 2011).

5.3.2 Preliminary evaluation of useful literature features

To test which literature features we should use for the competition, we started by performing an evaluation of two types of co-mention features extracted from the literature using a small mouse dataset of ∼3,500 proteins. Performance is presented as AUC (mean per GO term) and can be seen in Table 5.2 from 5-fold cross-validation. The baseline (0.782) presented is only using BLAST and PPI; using only literature features, protein-protein and protein-GO co-mentions from abstracts, produces comparable, but slightly reduced performance (0.774). The only protein-protein co-mention that improve performance are those within a sentence, even though the increase could be determined negligible. Document protein-protein co-mentions reduce performance when combined with any other features; this is most likely due to noise introduced and them being a very poor approximation to a traditional interaction. On the other hand, we find that protein-GO term co-mentions improve performance when combined with all other features. Using the document co-mentions improve performance over the sentence co-mentions. In this experiment, the sentence co-mentions are a subset of the document co-mentions and we believe more textual context can be encoded within the data when using document co-mentions over sentence. The best performance is when the document protein-GO co-mentions are combined with both BLAST and PPI. These literature co-mentions complement the other types of data and preliminary results show their usefulness for function prediction. Based upon this preliminary experiment, we chose to use the protein-GO term co-mentions from the entire abstract as the literature features for the first CAFA predictions.

Table 5.2: Performance on combination set of ~3,500 mouse proteins. Combination of AUC is the mean per GO term. PPC is protein-protein co-mentions and PGC is protein-GO co-mentions and the span of the co-mention is represented by either sentence or document (entire abstract). The difference from baseline is also presented.

Set of features	AUC	Δ Baseline
Mouse-BLAST + Mouse-PPI	0.782	–
PPC-document + PGC-document	0.774	-0.008
Mouse-BLAST + Mouse-PPI + PPC-sentence	0.783	+0.001
Mouse-BLAST + Mouse-PPI + PPC-document	0.779	-0.003
Mouse-BLAST + Mouse-PPI + PPC-sentence + PPC-document	0.773	-0.009
Mouse-BLAST + Mouse-PPI + PGC-sentence	0.807	+0.025
Mouse-BLAST + Mouse-PPI + PGC-document	**0.814**	+0.032
Mouse-BLAST + Mouse-PPI + PGC-sentence + PGC-document	**0.813**	+0.031
Mouse-BLAST + Mouse-PPI + PPC-sentence + PGC-sentence	0.799	+0.017
Mouse-BLAST + Mouse-PPI + PPC-document + PGC-document	0.799	+0.017
Mouse-BLAST + Mouse-PPI + PPC-sentence + PPC-document + PGC-sentence + PGC-document	0.789	+0.016

5.3.3 Impact of individual features

Our method was one of the top performers in the first CAFA with an F-max of ~0.57 for all targets for the Molecular Function branch (Radivojac et al., 2013). Unfortunately, due to timing issues, we were unable to incorporate literature features for the official predictions, but after the fact we performed extensive evaluation to explore the impact that each feature had using the methodology but also included literature features (Sokolov et al., 2013b). To assess the contribution of each source of data to predict function, we compared the performance of models trained on individual features and combinations of them. These results are presented in Table 5.3.

Our first observation is that BLAST data accounts for the largest contribution to the predictive power of the cross-species SVM, although the additional sequence-based features provide an increase in performance. From the species-specific view, the PPI features yield the highest accuracy, and outperforms all sequence-based predictors in biological function and cellular component namespaces. Furthermore, these features are complementary to the co-mention features, as demonstrated by the strong increase in performance over either feature set by itself when using the combination of the two. A classifier based solely on gene expression data did not fare well by itself. Nevertheless, inclusion of gene expression data provides a marginal increase in performance. We imagine that if we had time to incorporate the features into the submitted predictions that our performance on the ∼150 mouse proteins in the evaluation set would be improved.

Table 5.3: Classifier performance in predicting GO terms using individual sources of data and some of their combinations using only data from mouse. Reported performance is AUC calculated through 5-fold cross validation. P20R represents precision at recall 20%. BLAST refers to a classifier trained on BLAST scores only; the Sequence entry uses all the sequence-based features. In addition to classifiers trained on PPI, co-mention and expression individually, we also provide results using PPI and co-mention and the combination of all three.

	AUC			P@R20		
Set of features	MF	BP	CC	MF	BP	CC
BLAST	0.77	0.61	0.69	0.40	0.13	0.25
Sequence	0.83	0.65	0.76	0.41	0.14	0.26
PPI	0.78	0.80	0.81	0.33	0.25	0.43
Protein-GO co-mention	0.78	0.75	0.79	0.24	0.17	0.33
Gene Expression	0.58	0.64	0.62	0.04	0.06	0.10
PPI + co-mention	0.85	0.82	0.85	0.43	0.29	0.45
PPI + co-mention + expression	0.86	0.83	0.86	0.42	0.29	0.46

5.3.4 Manual validation of incorrect predictions

A manual analysis of incorrect predictions using literature features was performed to examine what information GOstruct used to make the prediction and to show how useful literature features are beyond prediction; they can be used for validation of predictions. Analysis of the top 25 false positives from the molecular function namespace is presented in Table 5.4.

Three main conclusions can be drawn from the analysis. First, predictions made are more accurate than the evaluation estimated; our system identified biologically correct annotations that were not yet available in the gold standard. The gold standard used for evaluation was from Feb 2011. When evaluated against the contents of SwissProt from April 2012, 16 out of the top 25 predictions are supported. Second, our NLP pipeline is able to extract pertinent information for function prediction. Even individual sentences can contain evidence of multiple GO annotations. For example, a sentence extracted by our pipeline from PMID:19414597,"LKB1, a master kinase that controls at least 13 downstream protein kinases including the AMP-activated protein kinase (AMPK), resides mainly in the nucleus.", describes both the function and the subcellular localization of the protein LKB1. Finally, even though the sentences extracted provide useful information, more sophisticated methods to extract information from them will need to be developed. Because we are using simple co-occurrence of protein and GO-terms, extracted associations are not always correct. For example, our pipeline associated peptidase activity with TIMP-2 on the basis of the following sentence: "The 72-kDa protease activity has been found to be inhibited by tissue inhibitor of metalloprotease-2 (TIMP-2), indicating that the protease is the matrix metalloprotease-2 (MMP-2)". Clearly, TIMP-2 does not actually have peptidase activity, but inhibits it. This incorrect association, and others like it, possibly mislead GOstruct predictions. Such errors will be addressed in future work by incorporating the semantic role of the protein in regards to the described function.

Table 5.4: **Analysis of the top 25 false positive predictions made by GOstruct.** Analysis of the top 25 false positive predictions made by GOstruct. We present the best supporting sentence for the function of each protein, the document source, and the most recent known annotation.

Protein	**GOstruct prediction / current annotation**	**Best supporting sentence**	**Pubmed ID**	**GO term(s) in sentence**
MGI:103293	GO:0016787 hydrolase activity	We recently demonstrated that human protein tyrosine phosphatase (PTP) L1, a large cytoplasmic phosphatase also known as PTPBAS/PTPN13/PTP-1E, is a negative regulator of IGF-1R/IRS-1/Akt pathway in breast cancer cells.	19782949	GO:0004722
MGI:103305	GO:0016787 hydrolase activity / N/A	N/A	N/A	N/A

Continued on next page

Protein	GOstruct prediction / current annotation	Best supporting sentence	Pubmed ID	GO term(s) in sentence
MGI:104597	GO:0016740 transferase activity / N/A	Using this assay system, chloramphenicol acetyltransferase activity directed by the cTNT promoter/upstream region was between two and three orders of magnitude higher in cardiac or skeletal muscle cells than in fibroblast cells, indicating that cis elements responsible for cell-specific expression reside in this region of the cTNT gene.	3047142	GO:0008811, GO:0016407
MGI:104744	GO:0022857 transmembrane transporter activity / GO:0005242 inward rectifier potassium channel activity	Many Andersen syndrome cases have been associated with loss-of-function mutations in the inward rectifier K(+) channel Kir2.1 encoded by KCNJ2.	18690034	GO:0015267
MGI:104744	GO:0022892 substrate-specific transporter activity / GO:0005242 inward rectifier potassium channel activity	IRK1, but not GIRK1/GIRK4 channels, showed a marked specificity toward phosphates in the 4,5 head group positions.	10593888	GO:0015267
MGI:105926	GO:0005515 protein binding	Based on our results together with previous work showing that Rin1 interacts with signal transducing adapter molecule to facilitate the degradation of EGFR, we hypothesize that the selective association of Rab5A and Rin1 contributes to the dominance of Rab5A in EGFR trafficking	19723633	GO:0005488
MGI:105938	GO:0005515 protein binding / GO:0030742 GTP-dependent protein binding	To validate this method, the binding of EEA-1 was confirmed and several novel Rab5-binding proteins were also identified by 2-dimensional electrophoresis and liquid chromatographymass spectrometry/mass spectrometry (LC-MS/MS).	19526728	GO:0017091, GO:0005488
MGI:107548	GO:0005515 protein binding / N/A	In vitro binding assays revealed that TRAF5 associates with the cytoplasmic tail of CD40, but not with the cytoplasmic tail of tumor receptor factor receptor type 2, which associates with TRAF2.	8790348	GO:0005515 GO:0003818

Continued on next page

Protein	GOstruct prediction / current annotation	Best supporting sentence	Pubmed ID	GO term(s) in sentence
MGI:1316660	GO:0005515 protein binding / N/A	Members of the voltage-gated calcium channel y subunit gene family (Cacng), have been rapidly discovered since the discovery of the identification of the mouse gamma2 gene (Cacng2) and its association with the stargazer mutant mouse line.	15000525	GO:0015267, GO:0005262
MGI:1341870	GO:0016301 kinase activity	LKB1, a master kinase that controls at least 13 downstream protein kinases including the AMP-activated protein kinase (AMPK), resides mainly in the nucleus.	19414597	GO:0050405
MGI:1341870	GO:0016740 transferase activity	LKB1 can phosphorylate the Thr174 of BRSK2, increasing its activity ¿50- fold.	16870137	GO:0016310
MGI:1341870	GO:0016772 transferring phosphorus containing groups	LKB1 tumour suppressor protein kinase phosphorylates and activates protein kinases belonging to the AMP activated kinase (AMPK) subfamily	15733851	GO:0004674
MGI:1343087	GO:0016740 transferase activity	PKCzeta thus functions as an adaptor, associating with a staurosporine insensitive PDK2 enzyme that catalyzes the phosphorylation of S472 of PKBgamma.	12162751	GO:0004697, GO:0004740
MGI:1343087	GO:0016772 transferring phosphorus containing groups / GO:0004740 pyruvate dehydrogenase kinase activity	PKCzeta thus functions as an adaptor, associating with a staurosporine insensitive PDK2 enzyme that catalyzes the phosphorylation of S472 of PKBgamma.	12162751	GO:0004697, GO:0004740
MGI:1926334	GO:0016787 hydrolase activity / GO:0004722 protein serine/threonine phosphatase activity	The protein B-50 is dephosphorylated in rat cortical synaptic plasma membranes (SPM) by protein phosphatase type 1 and 2A (PP-1 and PP-2A)-like activities.	1319470	GO:0004722
MGI:1926334	GO:0016788 hydrolase activity, acting on ester bonds / GO:0004722 protein serine/threonine phosphatase activity	The protein B-50 is dephosphorylated in rat cortical synaptic plasma membranes (SPM) by protein phosphatase type 1 and 2A (PP-1 and PP-2A)-like activities.	1319470	GO:0004722

Continued on next page

Protein	GOstruct prediction / current annotation	Best supporting sentence	Pubmed ID	GO term(s) in sentence
MGI:2140494	GO:0016787 hydrolase activity / N/A	Nuclear inhibitor of protein phosphatase-1 (NIPP1; 351 residues) is a nuclear RNA-binding protein that also contains in its central domain two contiguous sites of interaction with the catalytic subunit of protein phosphatase-1 (PP1(C)).	11104670	GO:0016791, GO:0003723
MGI:2140494	GO:0016788 hydrolase activity, acting on ester bonds / N/A	Nuclear inhibitor of protein phosphatase-1 (NIPP1; 351 residues) is a nuclear RNA-binding protein that also contains in its central domain two contiguous sites of interaction with the catalytic subunit of protein phosphatase-1 (PP1(C)).	11104670	GO:0016791, GO:0003723
MGI:2180854	GO:0005515 protein binding / N/A	We report here that RFXAP, a subunit of the DNA-binding RFX complex, also binds BRG1 and therefore provides a mechanism by which MHC class II gene chromatin can be remodeled in the absence of CIITA.	15781111	GO:0005515, GO:0003677, GO:0017091
MGI:2385847	GO:0005515 protein binding	In contrast with other MOs, this conformational switch is coupled with the opening of a channel to the active site, suggestive of a protein substrate.	16275925	GO:0005515, GO:0015267
MGI:96785	GO:0005515 protein binding	Here, using the yeast one-hybrid system and electrophoretic mobility shift assay, we report that Lhx2, a LIMhomeodomain protein, binds to the homeodomain site in the mouse M71 OR promoter region.	15173589	GO:0005515, GO:0017091
MGI:97531	GO:0005515 protein binding	Many proteins bind to the activated platelet derived growth factor receptor (PDGF-R) either directly or by means of adapter molecules.	8619809	GO:0005515
MGI:97809	GO:0016787 hydrolase activity	We conclude that VE-PTP is a Tie- 2 specific phosphatase expressed in ECs, and VE-PTP phosphatase activity serves to specifically modulate Angiopoietin/Tie-2 function.	10557082	GO:0004722, GO:0004725, GO:0016791
MGI:98753	GO:0008233 peptidase activity / N/A	The 72-kDa protease activity has been found to be inhibited by tissue inhibitor of metalloprotease-2 (TIMP-2), indicating that the protease is the matrix metalloprotease-2 (MMP-2).	12102173	GO:0008233, GO:0004222

Continued on next page

Protein	GOstruct prediction / current annotation	Best supporting sentence	Pubmed ID	GO term(s) in sentence
MGI:98753	GO:0016787 hydrolase activity / N/A	In the comparison of normal and cloned samples, a total of 41 spots were identified as differentially expressed proteins, of which 25 spots were upregulated proteins such as TIMP-2, glutamate-ammonia, and esterase 10, while 16 spots were down-regulated proteins such as PBEF and annexin A1.	20684987	GO:0004091

5.4 CAFA 2

The second CAFA was held in 2014 and follows a similar setup as the first. The prediction task has expanded, systems were required to predict all branches of GO (MF, BP, and CC) and predictions for human proteins could also be made to the Human Phenotype Ontology (HPO). Participants were provided ∼100,000 possible targets from 27 different species. Unlike the first competition there could be two different types of annotations evaluated:

1. For proteins that had no experimental annotations before, evaluation would be performed the same as the first CAFA.
2. Proteins can already have experimental annotations and could accrue new experimental annotations, this is called the re-annotation task.

5.4.1 Experimental setup

We use similar approaches to those we used in the first CAFA, we highlight those changes and their impact on function prediction in the following sections. To account for the large increase in number of target species, we first expanded the protein dictionary to cover the 27 species covered by the 100,000 target proteins. To increase the ability to identify proteins in text, synonyms for proteins were added from UniProt (Consortium et al., 2008) and BioThesaurus version 0.7 (Liu et al., 2006). We also made a few changes to the way co-mentions were extracted based upon lessons learned from previous experiments.

Unlike the first CAFA, where literature features were mined for only *M. musculus*, for the second iteration literature features were incorporated into 8 different species-specific views (*A. thaliana*, *H. sapiens*, *M. musculus*, *R. norvegicus*, *S. cerevisiea*, *S. pombe*, *D. melanogaster*, and *E. coli*). In human and yeast ∼75% of all proteins had at least one co-mention associated with them.

Previously, we've shown that literature improve performance on the prediction of *M. musculus* proteins. In the following sections we expand to other species and explore the impact of literature features on their ability to predict function of proteins from both *H. sapiens* (∼20,000 proteins) and *S. cerevisiea*(∼6,000 proteins). Performance is reported from 5-fold cross-validation using precision, recall, F-max, and AUC as evaluation metrics; these are computed in a protein-centric manner as described in Radivojac *et al* (Radivojac et al., 2013).

5.4.1.1 Changes made to co-mentions

In the first CAFA, we used document (whole abstract) level protein-GO term co-mentions. We experimented with sentence co-mentions but saw reduced performance when compared to document level (Section 5.3.2). We believe that sentence co-mentions should contain useful information, but since they were a complete subset of the document co-mentions, their full potential was not realized. For the second CAFA, we considered two separate spans: sentence and non-sentence. Sentence co-mentions are two entities of interest seen within a single sentence while non-sentence co-mentions are those that are mentioned within the same paragraph/abstract, but not within the same sentence. The number of co-mentions extracted for human and yeast proteins can be seen in Table 5.5. We also expanded the literature collection that we extracted these co-mentions from. We mined ∼13.5 million abstracts available from Medline along with ∼600 thousand full-text articles from the PubMed Open Access Collection (PMCOA). Comparing these numbers to those from CAFA1, utilizing the improved protein dictionary and a newer version of GO, we are able to identify many more proteins and GO concepts.

Table 5.5: Counts of co-mentions extracted from both Medline and PMCOA for the second CAFA.

Human				
Span	Unique Proteins	Unique GO Terms	Unique Co-mentions	Total Co-mentions
sentence	12,826	14,102	1,473,579	25,765,168
non-sentence	13,459	17,231	3,070,466	147,524,964
combined	13,492	17,424	3,222,619	173,289,862
Yeast				
Span	Unique Proteins	Unique GO Terms	Unique Co-mentions	Total Co-mentions
sentence	5,016	9,471	317,715	2,945,833
non-sentence	5,148	12,582	715,363	18,142,448
combined	5,160	12,819	748,427	21,088,281

5.4.1.2 Exploring how to combine co-mention features

We mined co-mentions from two different text spans and explore four different ways to use them.

1. only using sentence co-mentions
2. only using non-sentence co-mentions
3. combining counts from sentence and non-sentence co-mentions into one feature set in the input representation
4. using two separate feature sets for sentence and non-sentence co-mentions

The spans were explained in more detail above, under the *Changes made to co-mentions* section.

The performance of these four different strategies for combining the co-mention features for the enhanced dictionary can be seen in Figure 5.2. Each branch of GO is predicted and evaluated separately, but the way to combine features is the same for all branches. Using the two types of co-mentions as two separate feature sets provide the best performance on all branches of GO (see green shapes in Figure 5.2). These two types of co-mentions encode different but complementary information and the classifier is able to build a better model by considering them separately.

Interestingly, non-sentence co-mentions perform better than sentence co-mentions. This goes against intuition, as co-mentions within a sentence boundary act as a proxy to a relationship between the protein and its function. However, it was seen in Bada et al. (Bada et al., 2013) that often function annotations do not occur within a sentence boundary with

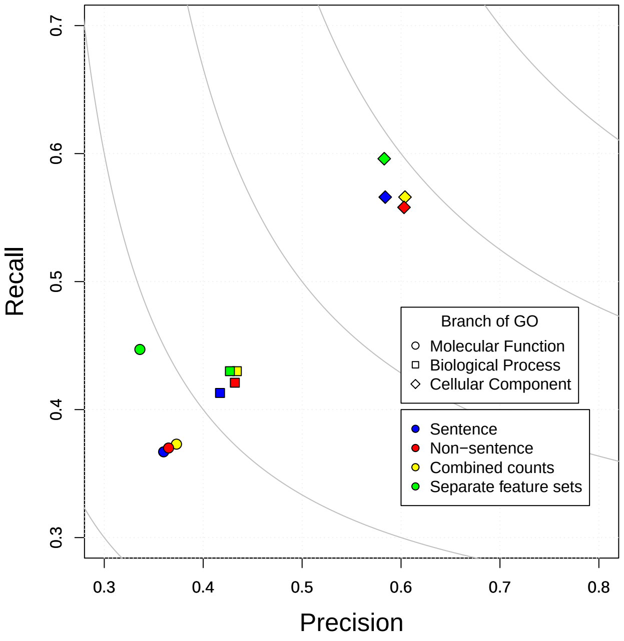


Figure 5.2: Precision, recall, and F-max performance of four different co-mention feature sets on function prediction. Better performance is to the upper-right and the grey iso bars represent balance between precision and recall. Diamonds – Cellular Component, Circle – Biological Process, Square – Molecular Function.

the corresponding protein. While coreference resolution may be required to correctly resolve such relationships, capturing function concepts in close proximity to a protein appears to be a useful approximation. This could be the reason why non-sentence co-mentions perform better. Based upon these results, from now on and in Chapter VI, when we say "co-mention features" we are referring to using both sentence and non-sentence as separate feature sets but within the same classifier.

5.4.1.3 Changes in other biological features

The same types of cross-species and species-specific features were used as outlined in Section 5.3.1.2; all data has been updated to the most recent datasets to include new knowledge gained over the past years. The most notable change was incorporation of more for PPI data, not only STRING, but BioGRID (Stark et al., 2006) and GeneMANIA (Warde-Farley et al., 2010) were combined to form species-specific network for the 8 species mentioned above.[9]

[9] Credit for other features and performing evaluation goes to Indika Kahanda from Asa Ben-Hur's lab.

5.4.2 Contribution from individual sources of data

To understand the effectiveness of each source of data we train and test classifiers created from each feature set alone along with the combination of all feature sets; this comparison on all three GO namespaces can be seen in Figure 5.3. The results show that literature data is almost as effective as homology and network-based data alone. For yeast, we find that on both Biological Process and Cellular Component, co-mentions alone outperforms transmembrane/localization and homology and falls slightly behind network information. We notice that despite high performance on yeast proteins, co-mentions are not as effective on human proteins. For human, literature outperforms transmembrane/localization features on all namespaces while F-max is 0.05-0.15 below both network and homology features. For all GO namespaces and both species, the combination of all features yields improved performance over any feature alone. This shows that all types of input data is useful and complementary to one another.

5.5 Conclusions

In this chapter, we've briefly described the work behind participation in two community challenges along with presenting further exploration of the impact that literature mined features have on function prediction. Over the course of the experiments we've refined the extraction and utilization of co-mentions. Examining performance on the ability to predict human, mouse, and yeast proteins, we've concluded that protein-GO term co-mentions are the most useful type of co-mentions (compared to protein-protein or GO-GO co-mentions). In addition, we compared the value of varying the span of text where the co-mentions occurs in: within a sentence ("sentence co-mention") and across a sentence boundary ("non-sentence co-mention"). Interestingly, we found that sentence and non-sentence co-mentions are equally useful, and that they are best used in conjunction as separate feature sets within a single classifier. Overall, we've shown that literature is a very informative feature for function predictions and continued work to develop more sophisticated methods for extracting protein-GO relations are required. While literature can be useful on its own, its real usefulness comes when combining with other features.

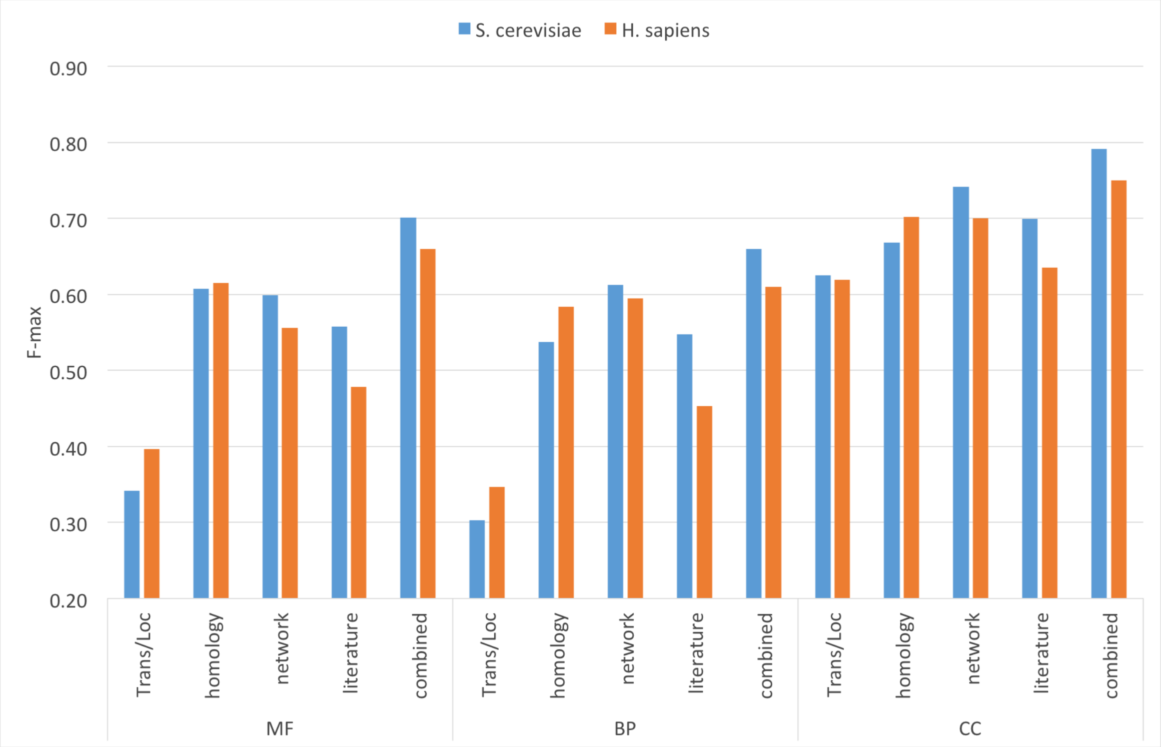


Figure 5.3: Performance of individual sources of data on *S. cerevisia* and *H. sapiens*. Transmembrane/localization signal are from TMHMM and WoLF PSORT, respectively. Homology is calculated from BLAST, and network aggregates interaction data from STRING, BioGRID, and GeneMANIA. Literature is the combination of sentence and non-sentence protein-GO term co-mentions from a large collection of literature. Combined is from all features combined.

We benchmarked the ability to recognize concepts in Chapter II and found recognition of GO concepts to be lacking. While it is clear from previous research that exact term matching is inadequate for good recall of Gene Ontology terms in text (Verspoor et al., 2003), it is also clear that accurately recognizing Gene Ontology terms is a challenging problem not only due to linguistic variation (Rice et al., 2005a) but due to variability in term informativeness in the context of the GO itself (Couto et al., 2005). We test the impact that improving GO recognition, presented in Chapter III, has on function prediction in Chapter VI. Our conservative exact-match approach to recognizing GO terms is highly precise, and its low coverage is likely offset by the large document collection we have considered in this work.

One thing that sets this work apart is that our literature collection is orders of magnitude larger than previous collections (for instance, another CAFA participant, Wong *et al* (Wong and Shatkay, 2013) uses 68,337 abstracts for training and the BioCreative data (Blaschke et al., 2005) consisted of 30,000 (full text) documents). Our use of direct protein mentions within a document to relate proteins to GO terms, and aggregated across the corpus as a whole, also differentiates this work from previous efforts that use externally provided protein-text links (like was done in Chapter IV). In BioCreative, the test data consisted of protein-document pairs in the input and most systems considered only the information within the document(s) provided for a protein rather than any document in the collection that might mention the protein; Wong *et al* (Wong and Shatkay, 2013) associates proteins to text via curated protein-document links in UniProt. This means our methods consider many more implied relationships than other methods. Additionally, we find that assuming all protein mentions are species independent did not hinder performance. This is possibly, again, due to the large literature collection. Another possible explanation is that proteins with the same name have similar or at least related functions in all organisms.

Lastly, the usefulness of co-mentions beyond prediction, for validation or evidence finding, is explored. We extract supporting sentences from the extracted co-mentions for the top 25 predictions made and show that many have supporting evidence within the literature. We explore this idea further in the next chapter.

CHAPTER VI

IMPACT OF CONCEPT RECOGNITION ON PROTEIN FUNCTION PREDICTION[10]

6.1 Introduction

As mentioned in the previous chapter, characterizing the functions of proteins is an important task in bioinformatics today. In recent years, many computational methods to predict protein function have been developed to help understand functions without performing costly experiments. In the last chapter I explored the usefulness of both protein-protein and protein-GO term co-mentions extracted from large literature collections on their ability to predict protein function. In this work, we introduce and explore another scalable literature feature – a bag-of-words model. In Chapter III, I implemented GO synonym generation rules to help increase the recall of GO concept recognition. In this chapter, I test the hypothesis that with the ability to better extract GO concepts from the literature would lead to more informative predictions. As mentioned in the last chapter, I also provide many examples of how extracted literature features can be helpful beyond prediction – for verification or validation and present a pipeline that could possibly help speed up the rate of functional curation.

6.2 Background

The work we presented in the first CAFA (Sokolov et al., 2013a) (Chapter V) is on a different scale from these previous efforts, and integrates information relevant for predicting protein function from a range of sources. We utilize as much of the biomedical literature as possible and are able to make predictions for the entire Gene Ontology, thanks to a structured output support vector machine (SVM) approach called GOstruct (Sokolov and Ben-Hur, 2010). We found in that previous work that features extracted from the literature alone approach performance of many commonly used features from non-literature sources, such as protein-protein interactions derived from a curated resource. However, we used only

[10] The work presented in this chapter is republished with permission from: *Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct* Journal of Biomedical Semantics 6.1 (2015): 9.

concept co-occurrence features – focusing on simple, scalable features – leaving open many questions about the best strategy for representing the literature for the task of automated protein function prediction. In this work we explore another scalable feature commonly used for natural language processing tasks, bag-of-words.

We have extended our workshop paper (Funk et al., 2014b) by refining enhanced GO synonym generation rules, performing more extensive analysis of the data at the functional class level, and extending validation through manual curation using a “medium-throughput” curation pipeline. As in the last chapter, we explore these questions in the context of the structured output SVM model, GOstruct.

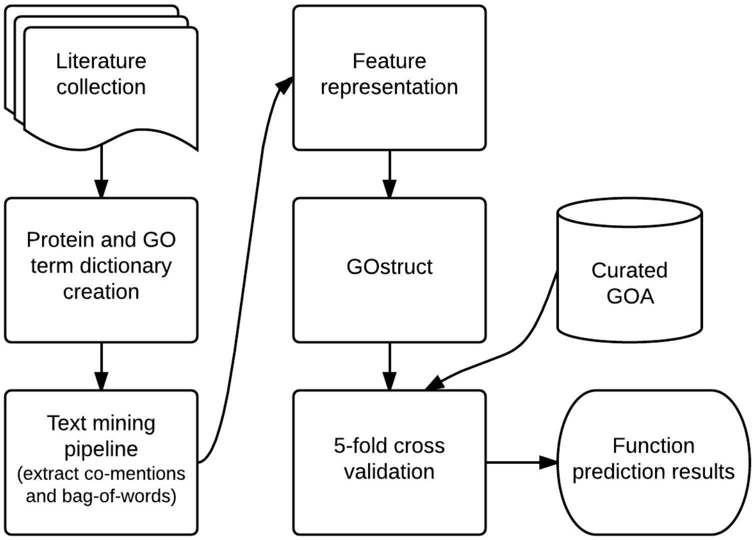

Literature collection
Feature representation
Protein and GO term dictionary creation
GOstruct
Curated GOA
Text mining pipeline (extract co-mentions and bag-of-words)
5-fold cross validation
Function prediction results

Figure 6.1: Overview of the experimental setup used for function prediction.

6.3 Methods

An overview of our experimental setup can be seen in Figure 6.1 with more specific details about each process following.

6.3.1 Data

We extracted text features from two different literature sources: (1) 13,530,032 abstracts available from Medline on October 23, 2013 with both a title and abstract text and

(2) 595,010 full-text articles from the PubMed Open Access Collection (PMCOA) downloaded on November 6, 2013. These literature collections were processed identically and features obtained from both were combined. Gold standard Gene Ontology annotations for both human and yeast genes were obtained from the Gene Ontology Annotation (GOA) data sets (Camon et al., 2004). Only annotations derived experimentally were considered (evidence codes EXP, IDA, IPI, IMP, IGI, IEP, TAS). Furthermore, the term Protein Binding (GO:0005515) was removed due to its broadness and overabundance of annotations. The human gold standard set consists of over 13,400 proteins annotated with over 11,000 functional classes while the yeast gold standard set consists of over 4,500 proteins annotated with over 6,500 functional classes. Even though the gold standard sets are large, only proteins where there is enough training data will produce predictions. Additionally, to produce meaningful area under the curve (AUC) scores only GO terms with at least 10 annotations in the gold standard are considered as possible prediction targets; this corresponds to 509 Molecular Function classes, 2,088 Biological Process classes, and 345 Cellular Component classes.

6.3.2 Literature features

Co-mentions are mentions of both a specific protein and concept from the Gene Ontology that co-occur with a specified span of text; they represent a simple knowledge-directed approach to represent the information contained within the biomedical literature. Through experiments conducted in the previous chapter, we know that protein-GO term co-mentions extracted at both the sentence and non-sentence level provide complementary types of data when used as separate features within the same classifier. Here, we explore the combination of these co-mentions with a very simple and commonly used set of features – a bag-of-words model. This is another representation of biomedical information is to relate proteins to words mentioned in the surrounding context; this is a knowledge-free approach because we are not grounding what we relate to proteins into some ontology, but only strings.

6.3.2.1 Text-mining pipeline

A pipeline was created to automatically extract the two different types of literature features using Apache UIMA version 2.4 (IBM, 2009). Whole abstracts were provided as input and full-text documents were provided one paragraph at a time. The pipeline consists of splitting the input documents into sentences, tokenization, and protein entity detection through LingPipe trained on CRAFT (Verspoor et al., 2012), followed by mapping of protein mentions to UniProt identifiers through a protein dictionary. Then, Gene Ontology (GO) terms are recognized through dictionaries provided to ConceptMapper (Tanenblatt et al., 2010). Finally, counts of GO terms associated with proteins, and sentences containing proteins, are output. A modified pipeline to extract proteins, GO terms, or any entity from an ontology file from text is available at http://bionlp.sourceforge.net/nlp-pipelines/. Details of the individual steps are provided below.

6.3.2.2 Protein mention extraction

The protein dictionary consists of over 100,000 protein targets from 27 different species, all protein targets from the CAFA2 competition (http://biofunctionprediction.org). To increase the ability to identify proteins in text, synonyms for proteins were added from UniProt (Consortium et al., 2008) and BioThesaurus version 0.7 (Liu et al., 2006).

6.3.2.3 Gene Ontology term extraction

The best performing dictionary-based system and parameter combination for GO term recognition identified in previous work was used (Funk et al., 2014a). ConceptMapper (CM) is highly configurable dictionary lookup system that is a native UIMA component. CM is highly configurable through the use of many parameters.

Two different dictionaries were provided to CM to extract Gene Ontology mentions from text: original and enhanced. Both dictionaries are based on GO from 2013-11-13. The original directly utilizes GO terms and synonyms, with the exception that the word "activity" was removed from the end of ontology terms. The enhanced dictionary augments the original dictionary with additional synonyms for many GO concepts. This is work that is presented in Chapter III. Rules were manually created by examining variation between ontology terms and the annotated examples in a natural language corpus. This enhanced dictionary improved GO recognition F-measure performance on CRAFT corpus (Bada et al., 2012; Verspoor et al., 2012) by 0.1 (from 0.49 to 0.59), through application of term transformation rules to generate synonyms.

A simple rule deals with the many GO terms of the form “*X* metabolic process”, which we have observed often do not occur literally in published texts. For example, for term GO:0043705, “cyanophycin metabolic process” synonyms of “cyanophycin metabolism” and “metabolism of cyanophycin” are generated. It is also noted that most of the terms in GO are nominals, so it is important to generate part of speech variants. There are also many ”positive regulation of *X*” terms; not only will we generate synonyms of “positive regulation of” such as “stimulation” and “pro”, but if there exist inflectional and derivational variants of *X* we can also substitute that in. For example, “apoptotic stimulation” and “pro-apoptotic” are added for “positive regulation of apoptosis” (GO:0043065). The version of the enhanced dictionary differs from the dictionary originally used for CAFA2, as described in (Funk et al., 2014b).

Table 6.1: Statistics of co-mentions extracted from both Medline and PMCOA using the different dictionaries for identifying GO terms.

Human					
Dictionary	Span	Unique Proteins	Unique GO Terms	Unique Co-mentions	Total Co-mentions
Original	sentence	12,826	14,102	1,473,579	25,765,168
	non-sentence	13,459	17,231	3,070,466	147,524,964
	combined	13,492	17,424	3,222,619	173,289,862
Enhanced	sentence	12,998	15,415	1,839,360	33,199,284
	non-sentence	13,513	18,713	3,725,450	196,761,554
	combined	13,536	18,920	3,897,951	229,960,838

Yeast					
Dictionary	Span	Unique Proteins	Unique GO Terms	Unique Co-mentions	Total Co-mentions
Original	sentence	5,016	9,471	317,715	2,945,833
	non-sentence	5,148	12,582	715,363	18,142,448
	combined	5,160	12,819	748,427	21,088,281
Enhanced	sentence	5,063	12,877	414,322	3,853,994
	non-sentence	5,160	13,769	901,123	23,986,761
	combined	5,167	14,018	939,743	27,840,755

6.3.2.4 Co-mentions

Co-mentions are based on co-occurrences of entity and ontology concepts identified in the literature text. We introduced them in more detail in Chapter V. This approach represents a targeted knowledge-based approach to feature extraction. The co-mentions we use here consist of a protein and Gene Ontology term that co-occur anywhere together in a specified span. While this approach does not capture relations as specific as an event extraction strategy (Björne and Salakoski, 2011), it is more targeted to the protein function prediction context as it directly looks for the GO concepts of the target prediction space. It also has higher recall since it doesn't require an explicit connection to be detected between the protein and the function term.

The number of co-mentions extracted for human and yeast proteins using both dictionaries can be seen in Table 6.1. For human proteins, the enhanced dictionary identifies 1,500 more GO terms than the original dictionary, which, leads to a 35% increase in the number of co-mentions identified (∼56 million more). Similar increases are seen with yeast proteins.

6.3.2.5 Bag-of-words

Bag-of-words (BoW) features are commonly used in many text classification tasks. They represent a knowledge-free approach to feature extraction. For these experiments, proteins are associated to words from sentences in which they were mentioned. All words were lowercased and stop words were removed, but no type of stemming or lemmatization was applied.

6.3.2.6 Feature representation

The extracted literature information is provided to the machine learning framework as sets of features. Each protein is represented as a list of terms, either Gene Ontology or words, along with the number of times the term co-occurs with that protein in all of the biomedical literature. An example entry from the co-mention features is as follows: "Q9ZPY7, co_GO:0003675=6, co_GO:0005623=2, co_GO:0009986=2, co_GO:0016020=2. . . ". We utilize a sparse feature representation and only explicitly state the non-zero features for both co-mentions and BoW.

6.3.3 Experimental setup

We evaluate the performance of literature features using the structured output SVM approach GOstruct (Sokolov and Ben-Hur, 2010). GOstruct models the problem of predicting GO terms as a hierarchical multi-label classification task using a single classifier. As input, we provide GOstruct with different sets of literature features for each protein, as described above, along with the gold standard GO term associations of that protein, used for training. From these feature sets, GOstruct learns patterns associating the literature features to the known functional labels for all proteins in the training set. Given a set of co-occurring terms for a single protein, a full set of relevant Gene Ontology terms can be predicted. In these experiments, we use no additional resource beyond the literature to represent proteins.

GOstruct provides confidence scores for each prediction; therefore, all results presented in this paper are based upon the highest F-measure over all sets of confidence scores, F-max (Radivojac et al., 2013). Precision, recall, and F-max are reported based on evaluation using 5-fold cross validation. To take into account the structure of the Gene Ontology, all gold standard annotations and predictions are expanded via the 'true path rule' to the root node of GO. The 'true path rule' states that 'the pathway from a child term all the way up to its top-level parent(s) must always be true'. We then compare the expanded set of terms. (This choice of comparison impacts the interpretations of our results, which is discussed further below.) All experiments were conducted on both yeast and human.

Note that the 'true path rule' is only utilized during the evaluation of features through machine learning system (as discussed in *Impact of evaluation metric on performance*). All numbers reported about the performance and predictions made by the machine learning system have the rule applied, while numbers strictly referring to counts of co-mentions mined from the literature do not.

6.3.4 Gene Ontology term information content

To better explore and understand the predictions that our system is making we'd like to know how specific or informative a specific function is; our goal is to perform well on the highly informative terms. We calculate an annotation-based information content(IC) for each GO term based upon the gold standard annotations. We utilize the information content formula outlined in Resnik *et al.* (Resnik, 1995) and applied directly to GO in Mazandu *et al.* (Mazandu and Mulder, 2013). The IC of a term is given by

$$IC(x) = -ln(p(x)), \tag{6.1}$$

where $p(x)$ is the relative frequency of term x in the GOA gold standard dataset, obtained from the frequency $f(x)$ representing the number $\mathscr{A}(x)$ of proteins annotated with x, considering the 'true path rule'. The frequency $f(x)$ is given by

$$f(x) = \begin{cases} \mathscr{A}(x) & \text{if } x \text{ is a leaf} \\ \mathscr{A}(x) + \sum_{z \in \mathscr{C}_h(x)} \mathscr{A}(z) & \text{otherwise,} \end{cases} \tag{6.2}$$

where a leaf is term with no children and $\mathscr{C}_h(x)$ is the set of GO terms that have x as a parent.

$$p(x) = \frac{f(x)}{f(R)}, \tag{6.3}$$

where $f(R)$ is the count of annotations corresponding to the root R and all of it's children; the three sub-ontologies (MF, BP, CC) were calculated separately because they share different root nodes.

The larger the term information content score is, the more informative the term is. From the GOA annotations, the range of scores computed is 0-10.89. The root nodes of the ontologies have an information content of 0. A very broad function that many gene/gene products share will have a low IC content. For instance, "GO:0005488 - binding" has an IC score of 0.20. There highest IC concept, "GO:0016862 - intramolecular oxidoreductase activity, interconverting keto- and enol-groups", has a score of 10.89.

6.3.5 Human evaluation of co-mentions

To support evaluation of the accuracy of the co-mention features, we sampled a number of them and asked a human assessor to rate each one as "good" (True Positive) or "bad" (False Positive), i.e., whether or not it captures a valid relationship. To assess accuracy of co-mentions as a whole, 1,500 sentence co-mentions were randomly sampled from the 33.2 million co-mentions for annotation. Additionally, three smaller subsets of co-mentions of specific functional classes, totaling about 500 co-mentions, were selected for annotation to assess accuracy of sentence co-mentions for specific functional classes. In total, there were around 3,000 full sentences annotated.

To enable fast annotation of this rating, we developed an approach that allows for "medium-throughput" manual annotation of co-mentions, about 60-100 per hour. The sentence co-mentions are transformed to brat rapid annotation tool (http://brat.nlplab.org/) format. The annotator views both the identified protein and functional concept in differing colors within the context of the entire sentence. The annotator is only required to connect them with a single relationship, either "Good-Comention" or "Bad-Comention". The annotator was instructed to view the labeled protein and GO concept as correct and to only annotate "Good-Comention" when there exists a relationship between the specified entities. While a relationship may exist between the annotated GO category and another exact mention of the labeled protein, that would be considered incorrect for the purposes of this annotation, i.e., it is a decision relative to individual mentions of the protein in a specific textual context. We utilized these annotations to assess quality of a random set of co-mentions and also to label subsets of co-mentions containing particular functional concepts.

6.4 Results and discussion

In Chapter V, Section 5.4.1.2 we performed evaluation of the best ways to combine the co-mention features. From now on, when we say "co-mention features" we are referring to using both sentence and non-sentence as separate feature sets but within the same classifier.

To establish a baseline we utilized the co-mentions themselves as a classifier; the co-mentions are used as the final predictions of the system. We performed evaluations using both original and enhanced co-mentions. Results from combining counts between sentence and non-sentence co-mentions are presented in Table 6.2. The baseline leads to very low precision for all branches but we do see impressive levels of recall. This signifies that information from the literature is able to capture relevant biological information, but because we identify many different co-mentions the false positive rate is fairly high.

We utilized our "medium-throughput" human annotation pipeline and curated 1,500 randomly sampled sentence co-mentions; we found that ∼30% (441 out of 1,500) appeared to correctly relate the labeled protein with the labeled function. From these results it seems that sentence co-mentions contain a high false positive rate, most likely due to many mentions of proteins or GO concepts within a single sentence. Methods for filtering sentences that contain ambiguous mentions, due to both ambiguous protein names and many annotations within sentences containing complex syntactic structure, are still to be explored. Additionally, more complicated relationship or event detection would reduce the number of false positives seen and provide the classifier with higher quality sentence co-mentions, but significantly reduce the total number of identified co-mentions. It is unclear which method would be preferred for function prediction features.

Table 6.2: Overall performance of literature features on human proteins. Precision, Recall are micro-averaged across all proteins and F-max is a protein-centric metric. Baseline corresponds to using only the co-mentions mined from the literature as a classifier. Macro-AUC is the average AUC per GO category. "Co-mentions + BoW" utilizes original co-mentions and BoW features within a single classifier.

Molecular Function				
Features	F-max	Precision	Recall	macro-AUC
Baseline (Original)	0.094	0.055	0.327	0.680
Baseline (Enhanced)	0.064	0.036	0.322	0.701
Co-mentions (Original)	0.386	0.302	**0.533**	0.769
Co-mentions (Enhanced)	0.377	0.336	0.447	0.764
BoW	0.394	**0.376**	0.414	0.768
Co-mentions + BoW	**0.408**	0.354	0.491	**0.790**

Biological Process				
Features	F-max	Precision	Recall	macro-AUC
Baseline (Original)	0.134	0.091	0.249	0.610
Baseline (Enhanced)	0.155	0.103	0.311	0.611
Co-mentions (Original)	0.424	0.426	0.422	0.750
Co-mentions (Enhanced)	0.429	0.427	0.430	0.752
BoW	**0.461**	**0.467**	0.455	0.768
Co-mentions + BoW	0.459	0.426	**0.510**	**0.779**

Cellular Component				
Features	F-max	Precision	Recall	macro-AUC
Baseline (Original)	0.086	0.050	0.305	0.640
Baseline (Enhanced)	0.073	0.041	0.317	0.642
Co-mentions (Original)	0.587	0.590	0.585	0.744
Co-mentions (Enhanced)	0.589	0.583	0.596	0.753
BoW	**0.608**	**0.594**	**0.624**	0.755
Co-mentions + BoW	0.607	0.592	0.622	**0.773**

6.4.1 Performance on human proteins

We report performance of all four feature sets on human proteins in Table 6.2. Comparing the performance of the co-mention features, we find that the original co-mention features produce the better performance on Molecular Function (MF), while the enhanced co-mentions perform slightly better on both Biological Process (BP) and Cellular Component (CC). The most surprising result is that bag of words performed as well as it did, considering the complexity of the Gene Ontology with its many thousands of terms. Many text classification tasks utilize BoW and achieve very good performance while some have tried to recognize functional classes from text with BoW models with poorer results (Mao et al., 2014; Jacob et al., 2013). Their applicability to function prediction has only begun to be studied in this work and Wong *et al.*(Shatkay et al., 2014). One explanation for their performance could be due to their higher utilization of the biomedical literature; co-mentions only capture information when both a protein and GO term are recognized together while BoW only relies on a protein to be recognized. In other words, the knowledge-based co-mentions are limited by the performance of automatic GO concept recognition, a challenging task in itself (Funk et al., 2014a), while the BoW features have no such limitation. In support of that, we note that on average, there are 2,375 non-zero BoW features per protein, whereas there are an average of 135 sentence and 250 non-sentence non-zero co-mention features per protein. The results reported here are for human proteins.

Overall, best performance for all branches of the Gene Ontology is seen when using both co-mentions and the bag-of-words features. This suggests that all types of features provide complementary information. In view of this observation, we explored an alternative to using the features in combination to train a single classifier, which is to train separate classifiers and combine their scores. This approach gave similar results to those reported here (data not shown). It can be difficult to understand the impact of each type of feature solely by looking at the overall performance, since it is obtained by averaging across all proteins; we dive deeper in the following sections and provide examples that indicate that using co-mentions produces higher recall than precision.

Another observation to make is that performance for all three branches of GO as measured using the macro-AUC is very similar, indicating that the three sub-ontologies are equally difficult to predict from the literature. The differences in performance as measured by F-max, which is a protein-centric measure, are likely the result of the differences in the distribution of terms across the different levels in the three sub-ontologies. The similar performance across the sub-ontologies is in contrast to what is observed when using other types of data: MF accuracy is typically much higher than BP accuracy, especially when using sequence data (Sokolov et al., 2013b; Radivojac et al., 2013), with the exception of network data such as protein-protein interactions that yields better performance in BP.

Table 6.3: Description of the gold standard human annotations and predictions made by GOstruct from each type of feature. All numbers are counts based on the predictions broken down by sub-ontology; these counts have the 'true path rule' applied.

	Molecular Function	Biological Process	Cellular Component
Feature type	# Predictions	# Predictions	# Predictions
Gold standard	36,349	264,631	79,631
Original	102,486	268,068	76,513
Enhanced	64,919	276,734	81,094
BoW	40,499	268,114	77,753
Combined	62,039	386,267	78,475

6.4.2 Exploring differences between original and enhanced co-mentions

Examining Table 6.1, we see that the enhanced dictionary finds ~35% (~56 million) more unique co-mentions, makes about 32,000 fewer predictions (Table 6.3) and performs slightly better at the function prediction task (Table 6.2). To elucidate the differences that GO term recognition plays in the function prediction task, co-mention features and predictions were examined for individual proteins.

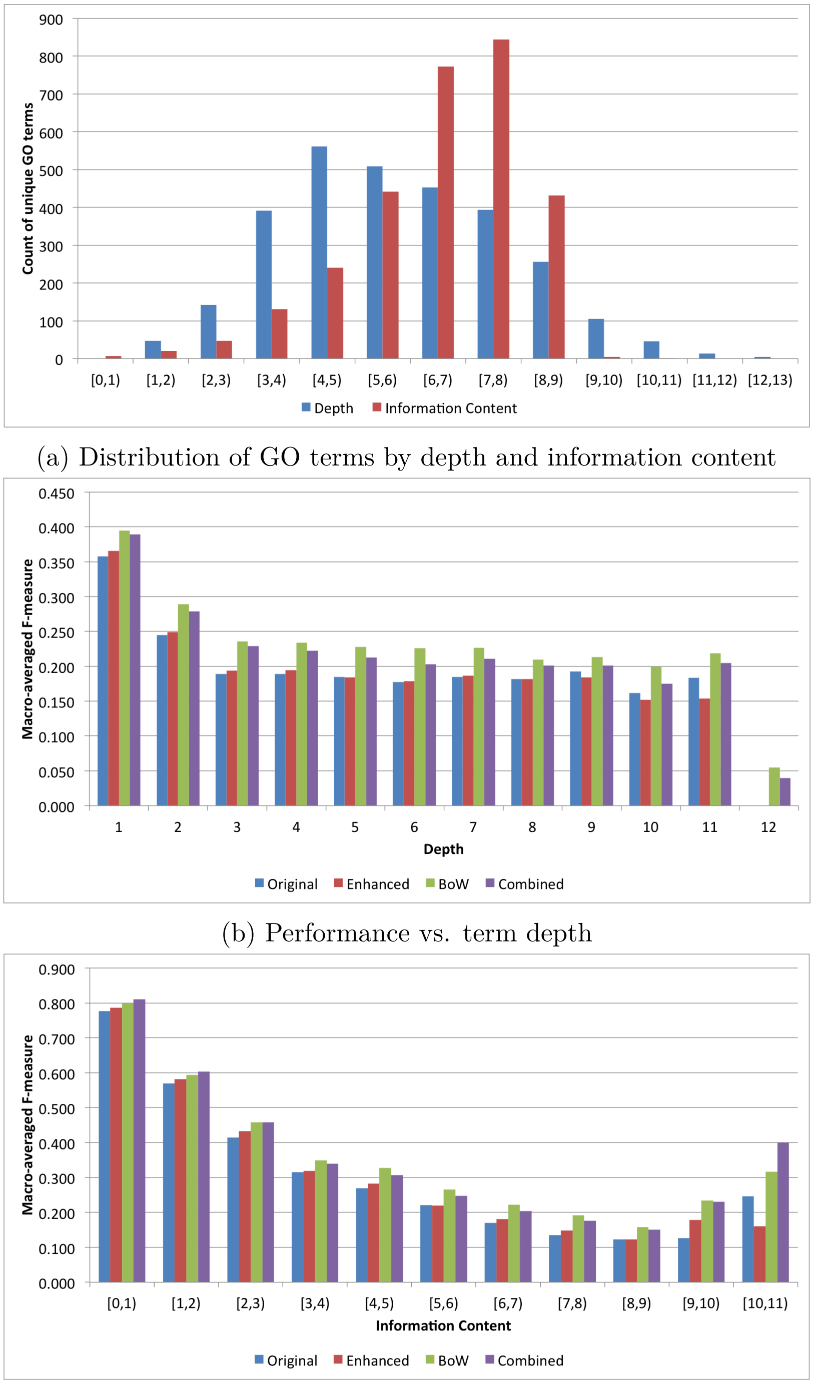
Examining individual predictions it appears that many of the predictions made from enhanced co-mention features are more specific than both the original dictionary and the gold standard annotations; this is also supported by further evidence presented in the functional analysis in the *Functional class analysis* and *Analysis of individual Biological Process and Molecular Function classes* sections. For example, in GOstruct predictions using the original dictionary, DIS3 (Q9Y2L1) is (correctly) annotated with rRNA processing (GO:0006364). Using co-mentions from the enhanced dictionary, the protein is predicted to be involved with maturation of 5.8S rRNA (GO:0000460), a direct child of rRNA processing. There are 10 more unique sentence and 31 more unique non-sentence GO term co-mentions provided as features by the enhanced dictionary. Some of the co-mentions identified by the enhanced and not by the original dictionary refer to "mRNA cleavage", "cell fate determination", and "dsRNA fragmentation". Even though none of these co-mentions directly correspond to the more specific function predicted by GOstruct, it could be that the machine learner is utilizing this extra information to make more specific predictions. Interestingly, the human DIS3 protein is not currently known to be involved with the more specific process, but the yeast DIS3 protein is. We did not attempt to normalize proteins to specific species because that is a separate problem in itself. It is probable that if we normalized protein mentions to specific species or implemented a cross-species evaluation utilizing homology the results of the enhanced dictionary would show improved performance.

We expected to see a bigger increase in performance because we are able to recognize more specific GO terms utilizing the enhanced dictionary. One possible reason that we don't is due to increased ambiguity in the dictionary. In the enhanced dictionary, for example, a synonym of "implantation" is added to the term "GO:0007566 - embryo implantation". While a majority of the time this synonym correctly refers to that GO term, there are cases such as "...tumor cell implantation" for which an incorrect co-mention will be added to the feature representation. These contextually incorrect features could limit the usefulness of those GO terms and result in noisier features. One way to address this may be to create a separate feature set of only co-mentions based on synonyms so the machine learner could differentiate or weight them differently; this could help improve performance using the enhanced dictionary co-mentions.

6.4.3 Functional class analysis

We now move to an analysis of functional classes to assess how well different parts of GO are predicted by different feature sets (Figure 6.2). We use two separate metrics, depth within the GO hierarchy and information content (IC) of the GO term derived from our gold standard annotations. Because the GO is a graph with multiple inheritance and depth can be a fuzzy concept (Joslyn et al., 2004), we define depth as the length of the shortest path from the root to the term in the GO hierarchy. We calculate an annotation-based information content(IC) for each GO term based on the gold standard annotations using the IC statistic described in Resnik *et al.*(Resnik, 1995).

Figure 6.2(a) shows the distribution of counts of GO terms within the gold standard and predictions by both depth and information content, Figure 6.2(b) shows the macro-averaged performance (F-measure) for each feature set by depth, and Figure 6.2(c) shows the macro-averaged performance for each feature set binned by GO term information content. Examining 6.2(a) we find that terms appear to be normally distributed with mean depth of 4. Looking at information content, we find that over two-thirds of the terms have an information content score between 6 and 8, indicating that a majority of terms within the



(a) Distribution of GO terms by depth and information content

(b) Performance vs. term depth

(c) Performance vs. term information content

Figure 6.2: Functional class analysis of all GO term annotations and predictions. A) Distribution of the depth and information content of GO term annotations. As IC values are real numbers, they are binned, and each bar represents a range, e.g. '[1,2)' includes all depth 1 terms and IC between 1 and 2 (not including 2). B) Macro-averaged F-measure performance broken down by GO term depth. C) Macro-averaged F-measure performance binned by GO term information content.

gold standard set are annotated very few times. Overall, for all sets of features, performance of concepts decreases as the depth and information content increases; it is intuitive that terms that are more broad, and less informative, would be easier to predict than terms that are specific and more informative.

Examining performance by depth (Figure 6.2(b)) we see a decrease in performance between depths 1-3, after which performance levels off. As a function of information content we obtain a more detailed picture, with a much larger decrease in performance with increased term specificity; all features are able to predict low information content, less interesting terms, such as “binding” (IC=0.20) or “biological regulation” (IC=0.66) with high confidence (F-measure > 0.8). Performance drops to its lowest for terms that have information content between 7 and 9 indicating there still remains much work to be done to accurately predict these specific and informative terms. Interestingly, there is an increase in performance for the most specific terms, especially using the BoW and combined representations; however, there are very few such terms as seen in Figure (6.2(a)), representing very few proteins, so it’s not clear if this is a real trend. Finally, we observe that for both depth and IC analysis the knowledge-free BoW features usually outperform the knowledge-based co-mentions and that the enhanced co-mentions usually produce slightly better performance than the original co-mentions.

6.4.4 Analysis of Biological Process and Molecular Function classes

To further explore the impact of the different features on predictions, we examined the best (Tables 6.4 & 6.5) and worst (Table 6.6) Biological Process and Molecular Function categories.

Examining the top concepts predicted, it is reinforced that the enhanced co-mentions are able to make more informative predictions, in addition to increasing recall without a loss in precision when compared to the original co-mentions. All 12 of the top terms predicted by the original co-mentions have an information content < 2 as opposed to only 7 terms from the enhanced co-mentions. We can compare the performance on specific functional classes. For example, “GO:0007076 - mitotic chromosome condensation” is the second highest predicted GO term by the enhanced co-mentions (F=0.769) while it is ranked 581 for the original co-

Table 6.4: **Top Biological Process and Molecular Function classes predicted by co-mention features.**

Original Co-Mentions

GO ID	Name	# Predictions	Precision	Recall	F-measure	Depth	IC
GO:0009987	cellular process	6,164	0.812	0.875	0.842	1	0.66
GO:0044699	single-organism process	4,849	0.743	0.765	0.754	1	0.96
GO:0044763	single-organism cellular process	4,295	0.681	0.714	0.697	2	1.20
GO:0008152	metabolic process	3,893	0.644	0.726	0.682	1	1.22
GO:0065007	biological regulation	3,615	0.691	0.629	0.658	1	0.90
GO:0071704	organic substance metabolic process	3,489	0.611	0.677	0.643	2	1.42
GO:0050789	regulation of biological process	3,350	0.668	0.601	0.633	2	0.97
GO:0044238	primary metabolic process	3,337	0.593	0.655	0.623	2	1.56
GO:0044237	cellular metabolic process	3,268	0.590	0.644	0.616	2	1.49
GO:0050794	regulation of cellular process	3,156	0.648	0.583	0.614	3	1.11
GO:0050896	response to stimulus	2,968	0.606	0.590	0.597	1	1.62
GO:0043170	macromolecule metabolic process	2,640	0.548	0.618	0.581	3	1.77

Enhanced Co-Mentions

GO ID	Name	# Predictions	Precision	Recall	F-measure	Depth	IC
GO:0009987	cellular process	6,223	0.816	0.887	0.850	1	0.66
GO:0007076	mitotic chromosome condensation	6	0.833	0.714	0.769	4	8.58
GO:0006323	DNA packaging	6	0.833	0.714	0.769	3	7.81
GO:0044699	single-organism process	4,957	0.744	0.783	0.763	1	0.96
GO:0044763	single-organism cellular process	4,423	0.682	0.736	0.708	2	1.20
GO:0008152	metabolic process	3,887	0.643	0.723	0.681	1	1.22
GO:0065007	biological regulation	3,701	0.683	0.636	0.659	1	0.90
GO:0050789	regulation of biological process	3,453	0.662	0.613	0.637	2	0.97
GO:0071704	organic substance metabolic process	3,491	0.605	0.670	0.636	2	1.42
GO:0043252	sodium-independent organic anion transport	11	0.636	0.583	0.608	7	8.50
GO:0000398	mRNA splicing, via spliceosome	140	0.492	0.697	0.577	10	5.88
GO:0006607	NLS-bearing protein import into nucleus	15	0.533	0.571	0.551	6	8.50

Table 6.5: **Top Biological Process and Molecular Function classes predicted by BoWs.**

Bag-of-words							
GO ID	Name	# Predictions	Precision	Recall	F-measure	Depth	IC
GO:0009987	cellular process	6,005	0.820	0.869	0.844	1	0.66
GO:0044699	single-organism process	4,940	0.754	0.799	0.776	1	0.96
GO:0044763	single-organism cellular process	4,449	0.696	0.764	0.728	2	1.20
GO:0043252	sodium-independent organic anion transport	8	0.875	0.583	0.700	7	8.50
GO:0065007	biological regulation	3,865	0.698	0.686	0.692	1	0.90
GO:0008152	metabolic process	3,870	0.647	0.733	0.688	1	1.22
GO:0050789	regulation of biological process	3,597	0.680	0.663	0.671	2	0.97
GO:0006479	protein methylation	13	0.615	0.727	0.666	8	6.52
GO:0051568	histone H3-K4 methylation	13	0.615	0.727	0.666	11	7.94
GO:0007076	mitotic chromosome condensation	5	0.800	0.571	0.666	4	8.58
GO:0050794	regulation of cellular process	3,440	0.657	0.651	0.654	3	1.11
GO:0006497	protein lipidation	9	0.889	0.500	0.640	7	6.79

Co-Mentions + Bag-of-words							
GO ID	Name	# Predictions	Precision	Recall	F-measure	Depth	IC
GO:0009987	cellular process	6,420	0.813	0.913	0.860	1	0.66
GO:0044699	single-organism process	5,338	0.736	0.834	0.782	1	0.96
GO:0044763	single-organism cellular process	4,862	0.674	0.800	0.731	2	1.20
GO:0065007	biological regulation	4,445	0.669	0.749	0.707	1	0.90
GO:0008152	metabolic process	4,252	0.638	0.785	0.704	1	1.22
GO:0050789	regulation of biological process	4,199	0.650	0.733	0.689	2	0.97
GO:0050794	regulation of cellular process	4,046	0.626	0.723	0.671	3	1.11
GO:0043252	sodium-independent organic anion transport	15	0.600	0.750	0.667	7	8.50
GO:0071704	organic substance metabolic process	3,883	0.602	0.743	0.665	2	1.42
GO:0043170	macromolecule metabolic process	3,007	0.540	0.694	0.607	3	1.77
GO:0051716	cellular response to stimulus	3,176	0.520	0.674	0.587	3	1.89
GO:0006386	termination of RNA polymerase III transcription	12	0.583	0.583	0.583	7	8.18

mentions (F=0.526). Granted, there will always be specific cases where one performs better than the other; from these and previous analyses, we find that the enhanced co-mentions are able to predict more informative terms for more proteins than the original co-mention features (Figure 6.2 and Tables 6.4 & 6.5). This shows that improving GO term recognition leads to an improvement in the specificity of function prediction.

Considering the top concepts predicted by the BoW features, we see a pattern similar to the enhanced co-mentions. Five out of the top twelve concepts predicted have an information content score greater than 6; these informative terms are different between the two feature sets. For the top functions predicted by all features the combined classifier of co-mentions and BoW produces more predictions, leading to higher recall and better F-measure. Even though some of the top terms predicted are informative and interesting we still strive for better performance on the most informative terms.

We also analyze the most difficult functional classes to predict, results can be seen in Table 6.6. Between all features we find some similar terms are difficult to predict; "localization" and "electron carrier activity" are in the worst five from all feature sets. It is interesting to note that the information content of these difficult to predict terms lies around the median range for all predicted terms. We might have expected that the most difficult terms to predict would be those most informative terms (IC around 10). We believe that these terms are difficult to predict because the ontological term names are made up of common words that will be seen many times in the biomedical literature, even when not related to protein function. This ambiguity likely results in a high number of features corresponding to these terms which results in poor predictive performance. There is still further work needed to address these shortcomings of literature mined features.

Table 6.6: Most difficult Biological Process and Molecular Function classes. IC represents information content of term.

Original Co-Mentions

GO ID	Name	# Predictions	Precision	Recall	F-measure	IC
GO:0051179	localization	28	0.107	0.054	0.072	5.70
GO:0016247	channel regulator activity	115	0.043	0.208	0.071	6.53
GO:0009055	electron carrier activity	108	0.03	0.111	0.055	6.94
GO:0007067	mitosis	23	0.043	0.031	0.036	7.54
GO:0042056	chemoattractant activity	53	0.018	0.067	0.029	7.56

Enhanced Co-Mentions

GO ID	Name	# Predictions	Precision	Recall	F-measure	IC
GO:0009055	electron carrier activity	102	0.090	0.138	0.109	6.94
GO:0051179	localization	42	0.071	0.055	0.061	5.70
GO:0019838	growth factor binding	44	0.021	0.035	0.027	5.99
GO:0070888	E-box binding	99	0.010	0.066	0.019	7.49
GO:0030545	receptor regulator activity	152	0.007	0.020	0.010	7.63

Bag-of-words

GO ID	Name	# Predictions	Precision	Recall	F-measure	IC
GO:0051179	localization	18	0.277	0.090	0.137	5.70
GO:0009055	electron carrier activity	29	0.103	0.083	0.092	6.94
GO:0016042	lipid catabolic process	26	0.076	0.054	0.063	5.80
GO:0015992	proton transport	15	0.066	0.047	0.055	7.29
GO:0005516	calmodulin binding	14	0.071	0.033	0.045	7.25

Co-Mentions + Bag-of-words

GO ID	Name	# Predictions	Precision	Recall	F-measure	IC
GO:0051179	localization	61	0.100	0.109	0.104	5.70
GO:0009055	electron carrier activity	62	0.079	0.138	0.101	6.94
GO:0030545	receptor regulator activity	63	0.064	0.080	0.071	7.63
GO:0042056	chemoattractant activity	24	0.041	0.066	0.051	7.56
GO:0040007	growth	27	0.030	0.066	0.047	7.33

6.4.5 Manual analysis of predictions

6.4.5.1 Manual analysis of individual predictions

We know that GO annotations are incomplete and therefore some predictions that are classified as false positives could be actually correct. The prediction may even be supported by an existing publication, however due to the slow process of curation they are not yet in a database. We manually examined false positive predictions that contain sentence level co-mentions of the protein and predicted function to identify a few examples of predictions that look correct but are counted as incorrect:

- Protein GCNT1 (Q02742) was predicted to be involved with carbohydrate metabolic process (GO:0006959). In PMID:23646466 (Ze-Min et al., 2013) we find "Genes related to **carbohydrate metabolism** include PPP1R3C, B3GNT1, and **GCNT1**. . . ".

- Protein CERS2 (Q96G23) was predicted to play a role in ceramide biosynthetic process (GO:0046513). In PMID:22144673 (Tidhar et al., 2012) we see ". . . **CerS2**, which uses C22-CoA for **ceramide synthesis**. . . ".

These are just two examples taken from the co-mentions, but there are most likely more, which could mean that the true performance of the system is underestimated. Through these examples we show how the input features can be used not only for prediction, but also for validation. This is not possible when using features that are not mined from the biomedical literature and illustrate their importance.

6.4.5.2 Manual analysis of functional classes

In the previous section we explored individual co-mentions that could serve as validation for an incorrect GOstruct prediction. In addition to this one-off analysis, we can label subsets of co-mentions pertaining to particular functional concepts for validation on a medium-throughput scale. To identify functional classes for additional exploration, all GO concepts were examined for three criteria: 1) their involvement in numerous co-mentions with human proteins 2) numerous predictions made with an overall average performance and 3) confidence in the ability to extract the concept from text. The concepts chosen for further annotation were GO:0009966 – "regulation of signal transduction", GO:0022857 – "transmembrane transporter", and GO:0008144 - "drug binding". For each of these classes all human co-mentions were manually examined.

We identified 204 co-mentions between a human protein and "GO:0008144 - drug binding" (IC=6.63). Out of 204 co-mentions, 112 appeared to correctly related the concept with the protein (precision of 0.554). 61 unique proteins were linked to the correct 112 co-mentions. Of these, only 4 contained annotations of "drug binding" in GOA, while the other 57 are not currently known to be involved with "drug binding". When we examined the predictions made by GOstruct for these proteins, unfortunately, none of them were predicted as "drug binding". After further examination of the co-mentions, most appear to be from structure papers and refer to drug binding pockets within specific residues or domains of the proteins. It is unlikely that the specific drug could be identified from the context of the sentence and many refer to a proposed binding site with no experimental data for support.

The concept "GO:0022857 - transmembrane transporter" (IC=4.17) co-occurred with a human protein 181 different times. 69 co-mentions appeared to correctly relate the concept with the labeled protein (precision of 0.381). A total of 32 proteins could be annotated with this concept; out of the 32 only 6 are not already annotated with "transmembrane transporter" in GOA. When we examine the predictions made from the enhanced features, only 1 out of the 6 proteins are predicted to be involved with "transmembrane transporter".

There were a total of 134 human co-mentions containing "GO:0009966 – regulation of signal transduction" (IC=3.30). 73 out of 134 co-mentions appeared to correctly relate the concept with the protein (precision of 0.543). A total of 58 proteins could be annotated based upon these co-mentions. 21 proteins already contain annotations conceptually related to "regulation of signal transduction", while the other 37 proteins do not contain annotations related to "regulation of signal transduction"; the later could represent true but uncurated functions. When we examine the predictions made by GOstruct using the enhanced co-mention features, 9 out of those 37 proteins were predicted to be involved with "regulation of signal transduction".

When a random subset of 1,500 human co-mentions were labeled it was found that ∼30% (441 out of 1,500) correctly related the labeled protein and GO term. By annotating co-mentions of specific functional concepts we see that these categories have a higher proportion of correct co-mentions than the random sample from all co-mention; there will also be some categories where performance of co-mentions is quite low. This information can be used in multiple different ways. If we are more confident that certain categories related to function can be extracted from co-mentions, we can use this information to inform the classifier by encoding the information into the input features. Additionally, we show the importance and ability of co-mentions to not only be used as input features, but also for validation and enhancing the machine learning results. We show that many of the predictions made by our system could possibly be correct, but just not curated in the gold standard annotations.

6.4.5.3 Impact of evaluation metric on performance

In our initial experiments, we required predictions and gold standard annotations to match exactly (data not shown), but we found, through manual examination of predictions, that many false positives are very close (in terms of ontological distance) to the gold standard annotations. This type of evaluation measures the ability of a system to predict functions exactly, at the correct specificity in the hierarchy, but it doesn't accurately represent the overall performance of the system. It is preferable to score predictions that are close to gold standard annotations higher than a far distant prediction. We are aware of more sophisticated methods to calculate precision and recall that take into account conceptual overlap for hierarchical classification scenarios (Verspoor et al., 2006; Clark and Radivojac, 2013). For the results reported in Table 6.2, to take into account the hierarchy of the Gene Ontology, we expanded both the predictions and annotations via the 'true path rule' to the root. By doing this, we see a large increase in both precision and recall of all features; this increase in performance suggests that many of the predictions made are close to the actual annotations and performance is better than previously thought. A downside of our chosen comparison method is that many false positives could be introduced via an incorrect prediction that is of a very specific functional class. This could possibly explain why co-mentions from the enhanced dictionary display a decrease in performance; a single, very specific, incorrect prediction introduces many false positives.

6.5 Conclusions

In this work we explored the use of protein-related features derived from the published biomedical literature to support protein function prediction. We evaluated two different types of literature features, ontology concept co-mentions and bag-of-words, and analyzed their impact on the function prediction task. Both types of features provided similar levels of performance. The advantage of the bag-of-words approach is its simplicity. The additional effort required to identify GO term mentions in text pays off by offering users the ability to validate predictions by viewing the specific literature context from which an association is derived, as demonstrated in our experiments.

Combining co-mentions and bag-of-words data provided only a marginal advantage, and in future work we will explore ways to obtain better performance from these features together. We also show that increasing the ability to recognize GO terms from biomedical text leads to more informative functional predictions. Additionally, the literature data we used provides performance that is on par with other sources of data such as network and sequence and has the advantage of being easy to verify on the basis of the text, as seen in the previous chapter. We believe that when we incorporate the synonym generation rules and bag-of-words features with the other biological features compared in the previous chapter that predictions would also be more informative.

Our experiment in medium-throughput manual inspection of protein-GO term co-mentions suggests that this strategy can be used as a way of speeding up the process of curation of protein function. The literature contains millions of co-mentions, and a human-in-the-loop system based on the detected co-mentions prioritized by GOstruct can be a highly effective method to dramatically speed up the rate at which proteins are currently annotated.

CHAPTER VII

CONCLUDING REMARKS AND FUTURE DIRECTIONS

The focus of this dissertation is the importance of biomedical concept recognition and application of recognizing concepts from ontologies for biomedical prediction. I begin by presenting an in-depth evaluation of concept recognition systems (Chapter II) and follow with improving performance of concepts from the Gene Ontology through hand-crafted synonym generation rules (Chapter III). I then switch focus and present two applications where features derived from mining concepts from a large collection of the biomedical literature are effective at making informative predictions. The two specific problems discussed are pharmacogene prediction (Chapter IV) and automated function prediction (Chapter V). I conclude with an analysis of the impact that improving concept recognition can have on function prediction (Chapter VI). In the following sections I review the most significant conclusions from the work and outline future directions.

7.1 Concept recognition

Ontologies have become a great enabling technology in modern bioinformatics. They aid in linking large scale genomic data for database curation and are useful for many biomedical natural language processing tasks, where they can be used as terminology and semantic constraints on entities and events. In Chapter II, I perform a rigorous linguistic evaluation, grounding concepts to both ontological identifier and exact span of text, of three dictionary-based concept recognition systems on their ability to recognize concept from eight biomedical ontologies against a fully annotated gold standard corpus. A full exploration of parameter space for each of system is also presented. We test multiple hypothesis and conclude the following:

1. Recognition performance is associated with the linguistic characteristics on an ontology and varies widely depending on the concept recognition system and parameters used for dictionary lookup.

2. Surprisingly, default parameters are not the best, or even close, set for most ontologies; illustrating the importance of parameter tuning for specific applications.

3. The heuristic methods implemented by some systems are helpful to identify complex multi-token concepts, but more sophisticated methods are needed to achieve better performance.

4. Morphological processing is an important task to incorporate into concept recognition. Most of the best parameters incorporated stemming or lemmatization – this helps to reduce the variability between the information contained within the ontology and the expression of concepts in text.

The Gene Ontology represents the standard nomenclature when discussing processes and functions that gene or gene products participate in. As seen in Chapter II, the ability to recognize these concepts is lacking because they are more complex than other ontologies – both in length and syntactic structure. The goal of the work presented in Chapter III is to address points improve performance on GO concept recognition by expanding on points 3 and 4 above. We take advantage of the compositional nature within the Gene Ontology and hand-craft 18 syntactic decompositional and derivational variant generation rules. These rules generate new synonyms for two-thirds of concepts with GO. Through intrinsic and extrinsic evaluation, we show these generated synonyms help to close the gap between ontological concept and their expression in natural language.

7.2 Biomedical discovery

The rest of my dissertation (Chapters IV-VI) focuses on the utility of Gene Ontology concept recognition for biomedical discovery. There has been little work incorporating the vast amount of knowledge contained in the biomedical literature into common computational prediction methods. I explore the following questions concerning literature features within two specific prediction tasks, pharmacogene predictions and protein function prediction:

1. What information should be mined from the literature?

2. How should it be combined with other of data, both literature and sequenced-based?

Another theme that runs throughout these chapters is the usefulness of text-mined features not only for prediction but for validation. We show that these automatic NLP pipelines could aid in speeding up manual curation.

7.2.1 Pharmacogene prediction

In Chapter IV, I present a classifier based upon a set of enriched functions that known pharmacogenes share to predict possible new pharmacogenes at a genome-wide scale. We find that there is a set of enriched GO concepts – mostly related to pharmacodynamics and pharmacokinetics. A multitude of features were explored along with combinations of them through a variety of machine learning algorithms: curated GO concepts from GOA, text-mined GO concepts, and bigrams from both abstract and full-text documents related to proteins. We find that using text-mined GO concepts and bigrams from abstracts is best able to separate known pharmacogenes from background genes. Using our classifier, there were a total of 141 hypothesized uncurated pharmacogenes. In light of new knowledge obtained since the original work was conducted, 6 out of the top 10 predicted pharmacogenes now have annotations curated in PharmGKB. Not only did our classifier predict these genes, but in the original work we provided literature support that now serves as evidence.

7.2.2 Protein function prediction

Chapters V and VI contains an evaluation of a variety of literature features effectiveness on predicting protein function within the machine learning framework, GOstruct. Most of the presented work is framed within the context of our participation in two community challenges. Our work focuses on mining simple scalable features, with the main construct being the co-mention, a co-occurrence of two entities in a predefined span of text. We find that the most effective type of co-mention is a protein and GO concept – a proxy to a relationship between the protein and its function. Another simple scalable feature text feature that contributes to performance is a bag-of-words model that captures the context around the protein mention. This work is set apart not only by the type of features mined but also for the large size of the literature collection. Because we use such a large collection of literature our features can consider many more implied relationships and signals will rise from the noise. By comparing literature features versus commonly used sequence- and network-based features, we find that these literature features approach the usefulness and are complementary to commonly used biologically features – best performance is always seen from the combination of all features.

We return to the concept recognition task in Chapter VI and evaluate the rules introduced in Chapter III within the function prediction task. We find that improving concept recognition leads to more informative predictions. Additionally, a "medium-throughput" pipeline is introduced to manually inspect co-mentions extracted from the literature. We suggest that co-mentions manual inspection of prioritized co-mentions could speed up the rate at which proteins are currently annotated.

7.3 Future directions

Research is never complete. In this section I describe the possible extensions for work presented here.

7.3.1 Chapter III - improving Gene Ontology performance

As seen in distribution of new synonyms and through the examples presented, our rules heavily favor the Biological Process branch, specifically the "regulation of" terms. Further rules can and will be developed to apply to the many other types of composition seen within GO. For this chapter, we chose to focus on the smallest subset that would have the most direct impact. This involves converting more *Obol* grammars to recursive syntactic rules and enumerating possible ways to express concepts. For the immediate next steps, it makes sense to focus on specific areas/type of concepts depending on the user or intended application.

There is previous work in creating lexical elementary synonym sets from the ontology itself (Hamon and Grabar, 2008). These could be incorporated once we decompose the concepts . Another place to explore for new synonyms would be the definition field within the Gene Ontology. This field is manually curated and contains beneficial information that could offer alternative ways and wording to express the concept.

Now that we are able to automatically generate synonyms we'd like other people to be able to use them for text-mining. It is our desire to submit the "good" synonyms identified within the text to the Gene Ontology Consortium for curation into the ontology. There could possibly be a "text-mining" synonym category added or we can deposit them, for the time being, within a larger application such as Freebase (Bollacker et al., 2008).

7.3.2 Chapter IV - pharmacogene prediction

Because our method only predicts only genes and not individual variants, that is a natural progression. This could be done through the use of text mining the literature about the predicted genes and extracting mentions of individual variants, specific mutations, along with mentions of disease or drugs. Another approach would be to combine the types of features many other methods use, network topology and sequence features, with the literature features explored here.

Another improvement that could be made would be to separate the prediction of disease and drug response genes. The drug response genes could even be subdivided into those that are display pharmacodynamics or pharmacokinetics. This could be done with the data we have now, but by creating new positive and negative training sets.

7.3.3 Chapter V and VI - function prediction

This work marks only the beginning of incorporating text mining for protein function prediction. There are always other more sophisticated or semantic features to explore, but based upon these results, there are some natural next steps. Overall, we've shown that literature is a very informative feature for function predictions and continued work to develop more sophisticated methods for extracting protein-GO relations are required. This includes incorporating negation along with the semantic role of the protein identified.

The first would be to incorporate larger spans for a bag-of-words model due to the surprising performance of the non-sentence co-mentions. By including words from surrounding sentences, or an entire paragraph, more context would be encoded and the model might result in better predictions.

Secondly, we found that an enhanced dictionary produced more individual co-mentions and fewer predictions, resulting in slightly increased performance. We explored several possible explanations as to why there is not a greater impact. It could be due to a large number of competing co-mentions that prevent good patterns from emerging or the possibility of introducing noise through ambiguous protein mentions. A filter or classifier that could identify a "good" co-mention would be providing much higher quality co-mentions as input, which would in turn likely lead to better predictions. Another way to potentially improve performance is to separate co-mentions found from synonyms from the original co-mentions, thereby allowing the classifier to provide them with different weights.

REFERENCES

Euan A Adie, Richard R Adams, Kathryn L Evans, David J Porteous, and Ben S Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC bioinformatics*, 6(1):55, 2005.

Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, Bert Coessens, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, et al. Gene prioritization through genomic data fusion. *Nature biotechnology*, 24(5): 537–544, 2006.

Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

Fátima Al-Shahrour, Ramón Díaz-Uriarte, and Joaquín Dopazo. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990. ISSN 0022-2836.

Sophia Ananiadou and John McNaught. *Text mining for biology and biomedicine*. Citeseer, 2006.

Sophia Ananiadou, Sampo Pyysalo, Jun'ichi Tsujii, and Douglas B Kell. Event extraction for systems biology by text mining the literature. *Trends in biotechnology*, 28(7):381–390, 2010.

A. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proc AMIA 2001*, pages 17–21, 2001. URL http://citeseer.ist.psu.edu/aronson01effective.html. Last accessed 2015-04-15.

Sofia J Athenikos and Hyoil Han. Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1):1–24, 2010.

Multiple authors. Poster session i, october 26, 2001 11:30 am–2:00 pm: Hepatobiliary/transplant. *Journal of Pediatric Gastroenterology and Nutrition*, 33: 358–370, 2001. URL http://journals.lww.com/jpgn/Fulltext/2001/09000/POSTER_SESSION_I_FRIDAY,_OCTOBER_26,_2001_11_30.33.aspx. Last accessed 2015-04-15.

ROBERT Bacallao and LEON G Fine. Molecular events in the organization of renal tubular epithelium: from nephrogenesis to regeneration. *American Journal of Physiology-Renal Physiology*, 257(6):F913–F924, 1989.

Michael Bada, Lawrence E Hunter, Miriam Eckert, and Martha Palmer. An overview of the craft concept annotation guidelines. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 207–211. Association for Computational Linguistics, 2010.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner Jr., Kevin Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. Concept annotation in the craft corpus. *BMC Bioinformatics*, 13(161), 2012.

Michael Bada, Dmitry Sitnikov, Judith A. Blake, and Lawrence E. Hunter. Occurrence of gene ontology, protein ontology, and ncbi taxonomy concepts in text toward automatic gene ontology annotation of genes and gene products. Berlin, Germany, 2013. BioLink – an ISMB Special Interest Group. URL http://biolinksig.org/proceedings/2013/biolinksig2013_Bada_etal.pdf. Last accessed 2015-04-15.

Michael Bada, William Baumgartner Jr, Christopher Funk, Lawrence Hunter, and Karin Verspoor. Semantic precision and recall for concept annotation of text. In *Proceedings of the BioOntologies SIG at ISMB'14*, 2014.

Eva Banik, Eric Kow, and Vinay K Chaudhri. User-controlled, robust natural language generation from an evolving knowledge base. In *ENLG*, volume 2013, page 14th, 2013.

J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biol*, 6(2), 2005. ISSN 1465-6914. Retrieved file cell.obo (versison 1.24 25:05:2007 09:56) from http://obofoundry.org/cgi-bin/detail.cgi?cell on June 14, 2007.

William Baumgartner, K. Cohen, Lynne Fox, George Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinf*, 23(13):i41–i48, 2007a. URL http://Bioinf.oxfordjournals.org/content/23/13/i41.abstract. Last accessed 2015-04-15.

William A. Baumgartner, K. Bretonnel Cohen, Lynne Fox, George K. Acquaah-Mensah, and Lawrence Hunter. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*, 23:i41–i48, 2007b.

William A Baumgartner Jr, Zhiyong Lu, Helen L Johnson, J Gregory Caporaso, Jesse Paquette, Anna Lindemann, Elizabeth K White, Olga Medvedeva, K Bretonnel Cohen, and Lawrence Hunter. An integrated approach to concept recognition in biomedical text. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume 23, pages 257–271. Centro Nacional de Investigaciones Oncologicas (CNIO) Madrid, Spain, 2007.

Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. The genetic association database. *Nature genetics*, 36(5):431–432, 2004.

Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch. Support vector machines and kernels for computational biology. *PLoS computational biology*, 4(10):e1000173, 2008.

Open biomedical ontologies. Open biomedical ontologies (obo). http://www.obofoundry.org/. Last accessed 2015-04-15.

Steven Bird. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

Jari Björne and Tapio Salakoski. A machine learning model and evaluation of text mining for protein function prediction. In *Automated Function Prediction Featuring a Critical Assessment of Function Annotations (AFP/CAFA) 2011*, pages 7–8, Vienna, Austria, 2011. Automated Function Prediction – an ISMB Special Interest Group. URL http://iddo-friedberg.net/afp-cafa-2011-booklet.pdf. Last accessed 2015-04-15.

Jari Björne, Filip Ginter, Sampo Pyysalo, Jun'ichi Tsujii, and Tapio Salakoski. Complex event extraction at pubmed scale. *Bioinformatics*, 26(12):i382–i390, 2010.

Christian Blaschke, Eduardo A. Leon, Martin Krallinger, and Alfonso Valencia. Evaluation of BioCreative assessment of task 2. *BMC Bioinformatics*, 6 Suppl 1, 2005.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.

N Bouayad-Agha, G Casamayor, and L Wanner. Natural language generation and semantic web technologies. *Semantic Web Journal*, 2012.

C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. Data-driven ontology evaluation, 2004.

Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The gene ontology annotation (goa) database: sharing knowledge in uniprot with gene ontology. *Nucleic acids research*, 32(suppl 1):D262–D266, 2004.

David Campos, Sérgio Matos, and José L Oliveira. A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):281, 2013.

J. Gregory Caporaso, William A. Baumgartner, David A. Randolph, K. Bretonnel Cohen, and Lawrence Hunter. Mutationfinder: A high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23:1862–1865, 2007.

Jt Chang and H. Schutze. *Abbreviations in biomedical text*, pages 99–119. Artech House, 2006.

Vinay K Chaudhri, Britte Cheng, Adam Overtholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. Inquire biology: A textbook that answers questions. *AI Magazine*, 34(3):55–72, 2013.

Jing Chen, Bruce J Aronow, and Anil G Jegga. Disease candidate gene identification and prioritization using protein interaction networks. *BMC bioinformatics*, 10(1):73, 2009.

Xin Chen, Zhi Liang Ji, and Yu Zong Chen. Ttd: therapeutic target database. *Nucleic acids research*, 30(1):412–415, 2002.

Jung-Hsien Chiang and Hsu-Chun Yu. Extracting functional annotations of proteins based on hybrid text mining approaches. In *Proc BioCreAtIvE Challenge Evaluation Workshop*. Citeseer, 2004.

Wenlei Mao Q. Zou Chu Wesley W, Zhenyu Liu. Kmex: A knowledge-based digital library for retrieving scenario-specific medical text documents. In *Biomedical Information Technology*. Elsevier, 2007.

Wyatt T Clark and Predrag Radivojac. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics*, 29(13):i53–i61, 2013.

Aaron M Cohen and William R Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9), 2008.

Kevin Bretonnel Cohen and Dina Demner-Fushman. *Biomedical natural language processing*, volume 11. John Benjamins Publishing Company, 2014.

Alain Coletta, John W Pinney, David YW Solís, James Marsh, Steve R Pettifer, and Teresa K Attwood. Low-complexity regions within protein sequences have position-dependent roles. *BMC systems biology*, 4(1):43, 2010.

The Gene Ontology Consortium. Creating the Gene Ontology resource: design and implementation. *Genome Research*, 11:1425–1433, 2001.

UniProt Consortium et al. The universal protein resource (uniprot). *Nucleic acids research*, 36(suppl 1):D190–D195, 2008.

Pedro R Costa, Marcio L Acencio, and Ney Lemke. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC genomics*, 11(Suppl 5):S9, 2010.

Francisco M Couto, Mário J Silva, and Pedro M Coutinho. Finding genomic ontology terms in text using evidence content. *BMC bioinformatics*, 6(Suppl 1):S21, 2005.

Robert Dale, Hermann Moisl, and Harold Somers. *Handbook of natural language processing*. CRC Press, 2000.

John Francis Davies, Marko Grobelnik, and Dunja Mladenic. *Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies*. Springer Science & Business Media, 2008.

Kirill Degtyarenko. Chemical vocabularies and ontologies for bioinformatics. In *Proc 2003 Itnl Chem Info Conf*, 2003.

David S DeLuca, Elena Beisswanger, Joachim Wermter, Peter A Horn, Udo Hahn, and Rainer Blasczyk. Mahco: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining. *Bioinformatics*, 25(16):2064–2070, 2009.

Joshua C Denny, Jeffrey D Smithers, Randolph A Miller, and Anderson Spickard. Understanding medical school curriculum content using knowledgemap. *Journal of the American Medical Informatics Association*, 10(4):351–362, 2003.

Joshua C Denny, Randolph A Miller Anderson Spickard III, Jonathan Schildcrout, Dawood Darbar, S Trent Rosenbloom, and Josh F Peterson. Identifying umls concepts from ecg impressions using knowledgemap. In *AMIA Annual Symposium Proceedings*, volume 2005, page 196. American Medical Informatics Association, 2005.

Heiko Dietze, Tanya Z Berardini, Rebecca E Foulger, David P Hill, Jane Lomax, David Osumi-Sutherland, Paola Roncaglia, and Christopher J Mungall. Termgenie–a web-application for pattern-based ontology class generation. *Journal of Biomedical Semantics*, 5(1):48, 2014.

A. Doms and M. Schroeder. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33:783–786, 2005.

Ian Donaldson, Joel Martin, Berry De Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, et al. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC bioinformatics*, 4(1):11, 2003.

F. Ehrler, A. GeissbÔæÉÔ°hler, A. Jimeno, and P. Ruch. Data-poor categorization and passage retrieval for gene ontology annotation in swiss-prot. *BMC Bioinformatics*, 6 Suppl 1, 2005. ISSN 1471-2105.

K. Eilbeck, S. E. Lewis, C. J. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biol*, 6(5), 2005.

William E Evans and Mary V Relling. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, 286(5439):487–491, 1999.

C. Fellbaum. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, Cambridge, Massachusetts, May 1998a. URL http://www.amazon.ca/exec/obidos/redirect?tag=citeulike04-20{&}path=ASIN/026206197X. Last accessed 2015-04-15.

D. Ferrucci and A. Lally. Building an example application with the unstructured information management architecture. *IBM Systems Journal*, 43(3):455–475, July 2004. ISSN 0018-8670.

Lynne M Fox, Leslie A Williams, Lawrence Hunter, and Christophe Roeder. Negotiating a text mining license for faculty researchers. *Information Technology and Libraries*, 33(3):5–21, 2014.

Marina Freytsis, Xueding Wang, Inga Peter, Chantal Guillemette, Suwagmani Hazarika, Su X Duan, David J Greenblatt, William M Lee, et al. The udp-glucuronosyltransferase (ugt) 1a polymorphism c. 2042c¿ g (rs8330) is associated with increased human liver acetaminophen glucuronidation, increased ugt1a exon 5a/5b splice variant mrna ratio, and decreased risk of unintentional acetaminophen-induced acute liver failure. *Journal of Pharmacology and Experimental Therapeutics*, 345(2): 297–307, 2013.

Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Cohen, Lawrence Hunter, and Karin Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):59, 2014a. ISSN 1471-2105. URL http://www.biomedcentral.com/1471-2105/15/59. Last accessed 2015-04-15.

Christopher Funk, Indika Kahanda, Asa Ben-Hur, and Karin Verspoor. Evaluating a variety of text-mined features for automatic protein function prediction. In *Proceedings of the BioOntologies SIG at ISMB'14*, 2014b.

Christopher S Funk, Lawrence E Hunter, and K Bretonnel Cohen. Combining heterogenous data for prediction of disease related and pharmacogenes. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 328–339. World Scientific, 2014c.

Christopher S Funk, Indika Kahanda, Asa Ben-Hur, and Karin M Verspoor. Evaluating a variety of text-mined features for automatic protein function prediction with gostruct. *Journal of Biomedical Semantics*, 6(1):9, 2015.

Keith D Garlid, Alexandre DT Costa, Casey L Quinlan, Sandrine V Pierre, and Pierre Dos Santos. Cardioprotective signaling to mitochondria. *Journal of molecular and cellular cardiology*, 46(6):858–866, 2009.

YAEL Garten, NICHOLAS P Tatonetti, and RUSS B Altman. In *Pac Symp Biocomput*, volume 305. World Scientific, 2010.

Hui Ge, Albertha JM Walhout, and Marc Vidal. Integrating 'omic'information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10):551–560, 2003.

Debarati Ghosh, Sailesh Gochhait, Disha Banerjee, Anindita Chatterjee, Swagata Sinha, and Krishnadas Nandagopal. Snapshot assay in quantitative detection of allelic nondisjunction in down syndrome. *Genetic Testing and Molecular Biomarkers*, 16 (10):1226–1235, 2012.

Julien Gobeill, Emilie Pasche, Dina Vishnyakova, and Patrick Ruch. Bitem/sibtex group proceedings for biocreative iv, track 4. In *Proceedings of the 4th BioCreative Challenge Evaluation Workshop*, volume 1, pages 108–113, 2013a.

Julien Gobeill, Emilie Pasche, Dina Vishnyakova, and Patrick Ruch. Managing the data deluge: data-driven go category assignment improves while complexity of functional annotation increases. *Database*, 2013:bat041, 2013b.

H. Goichi, M. Akio, and M. Shuji. Cartilage differentiation regulating gene, October 23 2003. URL https://www.google.com/patents/WO2003087375A1?cl=en. WO Patent App. PCT/JP2003/004,802.

Maria Beatriz Goncalves, Emma-Jane Williams, Ping Yip, Rafael J Yáñez-Muñoz, Gareth Williams, and Patrick Doherty. The cox-2 inhibitors, meloxicam and nimesulide, suppress neurogenesis in the adult mouse brain. *British journal of pharmacology*, 159(5):1118–1125, 2010.

Graciela Gonzalez, Juan C Uribe, Luis Tari, Colleen Brophy, and Chitta Baral. Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures. In *Pac Symp Biocomput*, volume 12, pages 28–39, 2007.

Nancy Green. Genie: an intelligent system for writing genetic counseling patient letters. In *AMIA Annual Symposium Proceedings*, volume 2005, page 969. American Medical Informatics Association, 2005.

Casey S Greene and Olga G Troyanskaya. Pilgrm: an interactive data-driven discovery platform for expert biologists. *Nucleic acids research*, 39(suppl 2):W368–W374, 2011.

Tudor Groza and Karin Verspoor. Assessing the impact of case sensitivity and term information gain on biomedical concept recognition. *PloS one*, 10(3):e0119091, 2015.

Thierry Hamon and Natalia Grabar. Acquisition of elementary synonym relations from biological structured terminology. In *Computational Linguistics and Intelligent Text Processing*, pages 40–51. Springer, 2008.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.

David Hancock, Norman Morrison, Giles Velarde, and Dawn Field. Terminizer–assisting mark-up of text using ontological terms, 2009.

Niclas Tue Hansen, Søren Brunak, and RB Altman. Generating genome-scale candidate gene lists for pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 86(2): 183–189, 2009.

Claudia Herr, Christoph S Clemen, Gisela Lehnert, Rüdiger Kutschkow, Susanne M Picker, Birgit S Gathof, Carlotta Zamparelli, Michael Schleicher, and Angelika A Noegel. Function, expression and localization of annexin a7 in platelets and red blood cells: Insights derived from an annexin a7 mutant mouse. *BMC biochemistry*, 4(1):8, 2003.

Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. Pharmgkb: the pharmacogenetics knowledge base. *Nucleic acids research*, 30(1):163–165, 2002.

David P Hill, Judith A Blake, Joel E Richardson, and Martin Ringwald. Extension and integration of the gene ontology (go): combining go vocabularies with external vocabularies. *Genome research*, 12(12):1982–1991, 2002.

Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

Yuko Hirata, Clement C Zai, Renan P Souza, Jeffrey A Lieberman, Herbert Y Meltzer, and James L Kennedy. Association study of grik1 gene polymorphisms in schizophrenia: case–control and family-based studies. *Human Psychopharmacology: Clinical and Experimental*, 27(4):345–351, 2012.

Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of BioCreAtIvE: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6, 2005.

Paul Horton, Keun-Joon Park, Takeshi Obayashi, Naoya Fujita, Hajime Harada, CJ Adams-Collier, and Kenta Nakai. Wolf psort: protein localization predictor. *Nucleic acids research*, 35(suppl 2):W585–W587, 2007.

Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47–50, 2008.

Lawrence Hunter, Zhiyong Lu, James Firby, William A Baumgartner, Helen L Johnson, Philip V Ogren, and K Bretonnel Cohen. 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*, 9(1):78, 2008.

Rachael P Huntley, Midori A Harris, Yasmin Alam-Faruque, Judith A Blake, Seth Carbon, Heiko Dietze, Emily C Dimmer, Rebecca E Foulger, David P Hill, Varsha K Khodiyar, et al. A method for increasing expressivity of gene ontology annotations using a compositional approach. *BMC bioinformatics*, 15(1):155, 2014.

Janna E Hutz, Aldi T Kraja, Howard L McLeod, and Michael A Province. Candid: a flexible method for prioritizing candidate genes for complex human traits. *Genetic epidemiology*, 32(8):779–790, 2008.

IBM. UIMA Java framework. http://uima-framework.sourceforge.net/, 2009.

Kristoffer Illergård, David H Ardell, and Arne Elofsson. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009.

Christoph Jacob, Philippe Thomas, and Leser Ulf. Comprehensive benchmark of gene ontology concept recognition tools. In *Proceedings of BioLINK Special Interest Group*, pages 20–26, July 2013.

A Jimeno. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(Suppl 3):S3, 2008. URL http://dx.doi.org/10.1186/1471-2105-9-S3-S3. Last accessed 2015-04-15.

Clement Jonquet, Nigam H Shah, and Mark A Musen. The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56, 2009.

Cliff A. Joslyn, Susan M. Mniszewski, Andy Fulmer, and Gary Heaton. The gene ontology categorizer. *Bioinformatics*, 20(suppl 1):i169–i177, 2004. URL http://bioinformatics.oxfordjournals.org/content/20/suppl_1/i169.abstract. Last accessed 2015-04-15.

Ning Kang, Bharat Singh, Zubair Afzal, Erik M van Mulligen, and Jan A Kors. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 2012.

George Karakatsiotis, Dimitrios Galanis, and Ion Androutsopoulos. Naturalowl: Generating texts from owl ontologies in protégé and in second life. In *System demonstration, 18th European Conference on Artificial Intelligence*, 2008.

V Karkaletsis, A Valarakos, and CD Spyropoulos. Populating ontologies in biomedicine and presenting their content using multilingual generation. *Programme Committee*, page 51, 2006.

Courtney M Karner, Amrita Das, Zhendong Ma, Michelle Self, Chuo Chen, Lawrence Lum, Guillermo Oliver, and Thomas J Carroll. Canonical wnt9b signaling balances progenitor cell expansion and differentiation during kidney development. *Development*, 138(7):1247–1257, 2011.

Uzay Kaymak, Arie Ben-David, and Rob Potharst. The auk: A simple alternative to the auc. *Engineering Applications of Artificial Intelligence*, 25(5):1082–1089, 2012.

Purvesh Khatri and Sorin Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005. doi: 10.1093/bioinformatics/bti565. URL http://bioinformatics.oxfordjournals.org/content/21/18/3587.abstract. Last accessed 2015-04-15.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl. 1):180–182, 2003.

Jin-Dong Kim, Sampo Pyysalo, Tomoko Ohta, Robert Bossy, Ngan Nguyen, and Jun'ichi Tsujii. Overview of bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 1–6. Association for Computational Linguistics, 2011.

Asako Koike, Yoshiki Niwa, and Toshihisa Takagi. Automatic extraction of gene/protein biological functions from biomedical text. *Bioinformatics*, 21(7):1227–1236, April 2005. ISSN 1367-4803.

Martin Krallinger, Maria Padron, and Alfonso Valencia. A sentence sliding window approach to extract protein annotations from biomedical articles. *BMC Bioinformatics*, 6 Suppl. 1, 2005.

Anders Krogh, BjoÈrn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of molecular biology*, 305(3):567–580, 2001.

Michael Kuhn, Christian von Mering, Monica Campillos, Lars Juhl Jensen, and Peer Bork. Stitch: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(suppl 1):D684–D688, 2008.

Jan Albert Kuivenhoven, J Wouter Jukema, Aeilko H Zwinderman, Peter de Knijff, Ruth McPherson, Albert VG Bruschke, Kong I Lie, and John JP Kastelein. The role of a common variant of the cholesteryl ester transfer protein gene in the progression of coronary atherosclerosis. *New England Journal of Medicine*, 338(2):86–93, 1998.

Robert Leaman, Graciela Gonzalez, et al. Banner: an executable survey of advances in biomedical named entity recognition. In *Pacific Symposium on Biocomputing*, volume 13, pages 652–663. Citeseer, 2008.

Haibin Liu, Ravikumar Komandur, and Karin Verspoor. From graphs to events: A subgraph matching approach for information extraction from biomedical text. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 164–172. Association for Computational Linguistics, 2011.

Haibin Liu, Tom Christiansen, William A. Baumgartner Jr., and Karin Verspoor. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of Biomedical Semantics*, 3(3), 212.

Hongfang Liu, Zhang-Zhi Hu, Jian Zhang, and Cathy Wu. Biothesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics*, 22(1):103–105, 2006.

Yaniv Loewenstein, Domenico Raimondo, Oliver C Redfern, James Watson, Dmitrij Frishman, Michal Linial, Christine Orengo, Janet Thornton, Anna Tramontano, et al. Protein function annotation by homology-based inference. *Genome Biol*, 10(2):207, 2009.

Yuqing Mao, Kimberly Van Auken, Donghui Li, Cecilia N Arighi, Peter McQuilton, G Thomas Hayman, Susan Tweedie, Mary L Schaeffer, Stanley JF Laulederkind, Shur-Jen Wang, et al. Overview of the gene ontology task at biocreative iv. *Database*, 2014: bau086, 2014.

Gaston K Mazandu and Nicola J Mulder. Information content-based gene ontology semantic similarity approaches: toward a unified framework theory. *BioMed research international*, 2013, 2013.

A. T. McCray, A. C. Browne, and O. Bodenreider. The lexical properties of the gene ontology. *Proc AMIA Symp*, pages 504–508, 2002. 1531-605X.

Ruslan Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2005.

N Miwa. [neonatal brain-derived carcinostatic factor (nbcf)–cytocidal action to neuroblastoma cells and molecular characters as a glycoprotein]. *Human cell*, 3(2):137–145, 1990.

Alexander Morgan and Lynette Hirschmann. Overview of biocreative ii gene normalization task. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, 2007.

Sara Mostafavi and Quaid Morris. Using the gene ontology hierarchy when predicting gene function. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 419–427. AUAI Press, 2009.

H. M. Muller, E. E. Kenny, and P. W. Sternberg. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2 (11), 2004.

Christopher J Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and functional genomics*, 5(6-7):509–520, 2004.

David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

Darren A Natale, Cecilia N Arighi, Winona C Barker, Judith A Blake, Carol J Bult, Michael Caudy, Harold J Drabkin, Peter D'Eustachio, Alexei V Evsikov, Hongzhan Huang, et al. The protein ontology: a structured representation of protein forms and complexes. *Nucleic acids research*, 39(suppl 1):D539–D545, 2011.

Natalya F Noy, Nigam H Shah, Patricia L Whetzel, Benjamin Dai, Michael Dorf, Nicholas Griffith, Clement Jonquet, Daniel L Rubin, Margaret-Anne Storey, Christopher G Chute, et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, page gkp440, 2009.

National Library of Medicine. Lvg:lexical variant generatior. http://lexsrv2.nlm.nih.gov/LexSysGroup/Projects/lvg/2012/web/index.html, 2012. Last accessed 2015-04-15.

P Ogren, K Cohen, and L Hunter. Implications of compositionality in the Gene Ontology for its curation and usage. In *Pacific Symposium on Biocomputing*, pages 174–185, 2005.

Philip V. Ogren, K. Bretonnel Cohen, George K. Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. The compositional structure of Gene Ontology terms. *Pacific Symposium on Biocomputing*, pages 214–225, 2004.

Min-Jung Park and Ji-Sook Han. Protective effects of the fermented laminaria japonica extract on oxidative damage in llc-pk1 cells. *Preventive nutrition and food science*, 18(4):227, 2013.

Paul Pavlidis, Jason Weston, Jinsong Cai, and William Stafford Noble. Learning gene functional classifications from multiple data types. *Journal of computational biology*, 9(2):401–411, 2002.

Lourdes Peņa-Castillo, Murat Taşan, Chad L Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, Michele Leone, Andrea Pagnani, Wan Kyu Kim, et al. A critical assessment of mus musculus gene function prediction using integrated genomic evidence. *Genome biology*, 9:S2, 2008.

Judes Poirier, Marie-Claude Delisle, Remi Quirion, Isabelle Aubert, Martin Farlow, Debmoi Lahiri, Siu Hui, Philippe Bertrand, Josephine Nalbantoglu, and Brian M Gilfix. Apolipoprotein e4 allele as a predictor of cholinergic deficits and treatment outcome in alzheimer disease. *Proceedings of the National Academy of Sciences*, 92 (26):12260–12264, 1995.

Martin F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Predrag Radivojac, Wyatt T Clark, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 2013. URL http://dx.doi.org/10.1038/nmeth.2340. Last accessed 2015-04-15.

Panchamoorthy Rajasekar and Carani Venkatraman Anuradha. Fructose-induced hepatic gluconeogenesis: effect of l-carnitine. *Life sciences*, 80(13):1176–1183, 2007.

S. Ray and M. Craven. Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics*, 6 Suppl. 1, 2005. ISSN 1471-2105.

D Rebholz-Schuhmann. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–8, 2008. URL http:/dx.doi.org/10.1093/bioinformatics/btm557. Last accessed 2015-04-15.

Lawrence Reeve and Hyoil Han. Conann: an online biomedical concept annotator. In *Data Integration in the Life Sciences*, pages 264–279. Springer, 2007.

Lawrence H Reeve, Hyoil Han, and Ari D Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, 43(6): 1765–1776, 2007.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

S. B. Rice, G. Nenadic, and B. J. Stapley. Mining protein function from text using term-based support vector machines. *BMC Bioinformatics*, 6 Suppl. 1, 2005a. ISSN 1471-2105.

Simon B Rice, Goran Nenadic, and Benjamin J Stapley. Mining protein function from text using term-based support vector machines. *BMC bioinformatics*, 6(Suppl 1):S22, 2005b.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83 (5):610–615, 2008.

Willie Rodgers, Francois-Michel Lang, and Cliff Gay. Metamap data file builder. http://metamap.nlm.nih.gov/Docs/datafilebuilder.pdf. Last accessed 2015-04-15.

Raul Rodriguez-Esteban. Biomedical text mining and its applications. *PLoS computational biology*, 5(12):e1000597, 2009.

Christophe Roeder, Clement Jonquet, Nigam H Shah, William A Baumgartner, Karin Verspoor, and Lawrence Hunter. A uima wrapper for the ncbo annotator. *Bioinformatics*, 26(14):1800–1801, 2010.

Burkhard Rost, Jinfeng Liu, Rajesh Nair, Kazimierz O Wrzeszczynski, and Yanay Ofran. Automatic prediction of protein function. *Cellular and Molecular Life Sciences CMLS*, 60(12):2637–2650, 2003.

Apache UIMA Sandbox. Conceptmapper annotator documentation. http://uima.apache.org/downloads/sandbox/ConceptMapperAnnotatorUserGuide/ConceptMapperAnnotatorUserGuide.html, 2009. Last accessed 2015-04-15.

Martijn J Schuemie, Rob Jelier, and Jan A Kors. Peregrine: Lightweight gene name normalization by dictionary lookup. *Proceedings of the Biocreative*, 2:23–25, 2007.

A.S. Schwartz and M.A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, volume 8, pages 451–462, 2003.

Burr Settles. ABNER: an open source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

Nigam Shah, Nipun Bhatia, Clement Jonquet, Daniel Rubin, Annie Chiang, and Mark Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(Suppl 9):S14, 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S9-S14. URL http:/dx.doi.org/10.1186/1471-2105-10-S9-S14. Last accessed 2015-04-15.

Hagit Shatkay, Scott Brady, and Andrew Wong. Text as data: Using text-based features for proteins representation and for computational prediction of their characteristics. *Methods*, 2014. ISSN 1046-2023.

Eileen M Shore. Fibrodysplasia ossificans progressiva: a human genetic disorder of extraskeletal bone formation, or—how does one tissue become another? *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1):153–165, 2012.

Matthew S Simpson and Dina Demner-Fushman. Biomedical text mining: A survey of recent progress. In *Mining text data*, pages 465–517. Springer, 2012.

Artem Sokolov and Asa Ben-Hur. Hierarchical classification of gene ontology terms using the gostruct method. *Journal of Bioinformatics and Computational Biology*, 8 (02):357–376, 2010.

Artem Sokolov, Christopher Funk, Kiley Graim, Karin Verspoor, and Asa Ben-Hur. Combining heterogeneous data sources for accurate functional annotation of proteins. *BMC Bioinformatics*, 14(Suppl 3), 2013a.

Artem Sokolov, Christopher Funk, Kiley Graim, Karin Verspoor, and Asa Ben-Hur. Combining heterogeneous data sources for accurate functional annotation of proteins. *BMC Bioinformatics*, 14(Suppl 3):S10, 2013b. ISSN 1471-2105. doi: 10.1186/1471-2105-14-S3-S10. URL http://www.biomedcentral.com/1471-2105/14/S3/S10. Last accessed 2015-04-15.

Irena Spasic, Sophia Ananiadou, John McNaught, and Anand Kumar. Text mining and ontologies in biomedicine: making sense of raw text. *Briefings in bioinformatics*, 6(3):239–251, 2005.

Padmini Srinivasan. Text mining: generating hypotheses from medline. *Journal of the American Society for Information Science and Technology*, 55(5):396–413, 2004.

Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl 1):D535–D539, 2006.

Johannes Stegmann and Guenter Grohmann. Hypothesis generation guided by coword clustering. *Scientometrics*, 56(1):111–135, 2003.

Samuel Alan Stewart, Maia Elizabeth von Maltzahn, and Syed Sibte Raza Abidi. Comparing metamap to mgrep as a tool for mapping free text to formal medical lexicons. In *Proceedings of the 1st International Workshop on Knowledge Extraction and Consolidation from Social Media (KECSM2012)*, 2012.

Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller, Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl 1):D561–D568, 2011.

Lorraine Tanabe and W John Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.

Michael Tanenblatt, Anni Coden, and Igor Sominsky. The conceptmapper approach to named entity recognition. In *International Conference on Language Resources and Evaluation*, 2010.

J Tarabeux, O Kebir, J Gauthier, FF Hamdan, L Xiong, A Piton, D Spiegelman, E Henrion, B Millet, F Fathalli, et al. Rare mutations in n-methyl-d-aspartate glutamate receptors in autism spectrum disorders and schizophrenia. *Translational psychiatry*, 1(11):e55, 2011.

The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000. ISSN 1061-4036.

Rotem Tidhar, Shifra Ben-Dor, Elaine Wang, Samuel Kelly, Alfred H Merrill, and Anthony H Futerman. Acyl chain specificity of ceramide synthases is determined within a region of 150 residues in the tram-lag-cln8 (tlc) domain. *Journal of Biological Chemistry*, 287(5):3197–3206, 2012.

Léon-Charles Tranchevent, Roland Barriot, Shi Yu, Steven Van Vooren, Peter Van Loo, Bert Coessens, Bart De Moor, Stein Aerts, and Yves Moreau. Endeavour update: a web resource for gene prioritization in multiple species. *Nucleic acids research*, 36(suppl 2):W377–W384, 2008.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *AAAI Fall Symposium: Information Retrieval and Knowledge Discovery in Biomedical Text*, 2012.

Karin Van der Borght, Rickard Köhnke, Nathanael Göransson, Tomas Deierborg, Patrik Brundin, Charlotte Erlanson-Albertsson, and Andreas Lindqvist. Reduced neurogenesis in the rat hippocampus following high fructose consumption. *Regulatory peptides*, 167(1):26–30, 2011.

S Van Landeghem, K Hakala, S Rönnqvist, T Salakoski, Y Van de Peer, and F Ginter. Exploring biomolecular literature with evex: Connecting genes through events, homology and indirect associations. *Advances in Bioinformatics*, Special issue Literature-Mining Solutions for Life Science Research:ID 582765, 2012. URL http://dx.doi.org/10.1155/2012/582765. Last accessed 2015-04-15.

Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, 6(1):e1000641, 2010.

C. Verspoor, C. Joslyn, and G. Papcun. The Gene Ontology as a source of lexical semantic knowledge for a biological natural language processing application. In *Proceedings of the SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics*, Toronto, CA, August 2003. URL http://compbio.ucdenver.edu/Hunter_lab/Verspoor/Publications_files/LAUR_03-4480.pdf. Last accessed 2015-04-15.

K. Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L. M. Rocha, and T. Simas. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, 6 Suppl. 1, 2005. ISSN 1471-2105.

Karin Verspoor, Judith Cohn, Susan Mniszewski, and Cliff Joslyn. A categorization approach to automated ontological function annotation. *Protein Science*, 15(6):1544–1549, 2006. ISSN 1469-896X.

Karin Verspoor, Daniel Dvorkin, K. Bretonnel Cohen, and Lawrence Hunter. Ontology quality assurance through analysis of term transformations. *Bioinformatics*, 25 (12):77–84, 2009.

Karin Verspoor, Kevin Bretonnel Cohen, Arrick Lanfranchi, Colin Warner, Helen L. Johnson, Christophe Roeder, Jinho D. Choi, Christopher Funk, Yuriy Malenkiy, Miriam Eckert, Nianwen Xue, William A. Baumgartner Jr., Michael Bada, Martha Palmer, and Lawrence E. Hunter. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, 13(207), 2012.

Karin M. Verspoor. Roles for text mining in protein function prediction. In Vinod D. Kumar and Hannah Jane Tipney, editors, *Biomedical Literature Mining*, volume 1159 of *Methods in Molecular Biology*, pages 95–108. Springer, New York, 2014. ISBN 978-1-4939-0708-3. URL http://dx.doi.org/10.1007/978-1-4939-0709-0_6. Last accessed 2015-04-15.

David Warde-Farley, Sylva L Donaldson, Ovi Comes, Khalid Zuberi, Rashad Badrawi, Pauline Chao, Max Franz, Chris Grouios, Farzana Kazi, Christian Tannus Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl 2):W214–W220, 2010.

Marc Weeber, Rein Vos, Henny Klein, Alan R Aronson, Grietje Molema, et al. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association*, 10(3):252–259, 2003.

D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, W. Helmberg, Y. Kapustin, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 34(Database issue):D5–D12, 2006. 1362-4962.

Anthony J Williams, Rossana Berti, Changping Yao, Rebecca A Prince, Luisa C Velarde, Irwin Koplovitz, Susan M Shulz, Frank C Tortella, and Jitendra R Dave. Inflammatory gene response in rat brain following soman exposure. Technical report, DTIC Document, 2005.

David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006.

A. Wong and H. Shatkay. Protein function prediction using text-based features extracted from biomedical literature: The cafa challenge. *BMC Bioinformatics*, 14 (Suppl 3), 2013.

Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li. Network-based global inference of human disease genes. *Molecular Systems Biology*, 4(1), 2008.

Lixia Yao, Anna Divoli, Ilya Mayzus, James A Evans, and Andrey Rzhetsky. Benchmarking ontologies: bigger or better? *PLoS Computational Biology*, 7(1):e1001055, 2011.

Hong Yu, Vasileios Hatzivassiloglou, Carol Friedman, Andrey Rzhetsky, and W John Wilbur. Automatic extraction of gene and protein synonyms from medline and journal articles. In *Proceedings of the AMIA Symposium*, page 919. American Medical Informatics Association, 2002.

Hong Yu, Won Kim, Vasileios Hatzivassiloglou, and W John J. Wilbur. Using medline as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *J Biomed Inform*, June 2006. ISSN 1532-0480.

Y Ze-Min, C Wei-Wen, and W Ying-Fang. [research on differentially expressed genes related to substance and energy metabolism between healthy volunteers and splenasthenic syndrome patients with chronic superficial gastritis]. *Chinese journal of integrated traditional and Western medicine*, 33(2):159–163, 2013.

Dongqing Zhu, Dingcheng Li, Ben Carterette, and Hongfang Liu. Integrating information retrieval with distant supervision for gene ontology annotation. *Database*, 2014:bau087, 2014.

Qinghua Zou, Wesley W Chu, Craig Morioka, Gregory H Leazer, and Hooshang Kangarloo. Indexfinder: a method of extracting key concepts from clinical texts for indexing. In *AMIA Annual Symposium Proceedings*, volume 2003, page 763. American Medical Informatics Association, 2003.

Pierre Zweigenbaum. Question answering in biomedicine. In *Proceedings Workshop on Natural Language Processing for Question Answering, EACL*, volume 2005, pages 1–4. Citeseer, 2003.

Pierre Zweigenbaum, Dina Demner-Fushman, Hong Yu, and Kevin B Cohen. Frontiers of biomedical text mining: current progress. *Briefings in bioinformatics*, 8(5): 358–375, 2007.