Extracting Co-mention Features from Biomedical Literature for Automated Protein Phenotype Prediction using PHENOstruct

Morteza Pourreza Shahri and Indika Kahanda
Gianforte School of Computing, Montana State University
Bozeman, MT 59717, USA
(mpourrezashahri, indika.kahanda)@montana.edu

Abstract

Human Phenotype Ontology (HPO) is a recently introduced standard vocabulary for describing disease-related phenotypic abnormalities in human. Since experimental determination of HPO categories for human proteins is a highly resource-consuming task, developing automated tools that can accurately predict HPO categories has gained interest recently. In our previous work, we developed PHENOstruct, an automated phenotype prediction tool that uses input features generated from heterogeneous data sources including standard bag-of-words features extracted from biomedical literature. In this work, we introduce novel co-mention features which are based on co-occurrences of protein names and HPO terms within a specified span of text. Our experimental results indicate that utilizing co-mentions significantly improves the overall performance and that the most effective span is the paragraph-level. This is the first study that uses a knowledge-based approach for generating literature features for the task of automated protein phenotype prediction. These findings have implications for practitioners interested in developing automated biocuration pipelines for phenotypes.

1 Introduction

Phenotypes can be described as any observable characteristics of an organism which have fascinated researchers' interests since the relationship between a gene and its phenotypic manifestation was discovered [15]. The Human Phenotype Ontology (HPO) provides a bioinformatics resource which offers a framework for the analysis of phenotypic abnormalities associated with human disease [11]. HPO was originally populated based on databases, such as OMIM (Online Mendelian Inheritance in Man) [6], which contain information about rare diseases. Each single protein is linked to a set of HPO terms based on the diseases caused by mutation to the corresponding genes. Currently, only a small portion of human proteins (about 3,500) have HPO annotations, and researchers believe there are more genes related to human diseases (P. Robinson, personal communication, July 12, 2014). Manually annotating proteins with HPO categories through wet-lab experiments and/or clinical studies is a highly-resource-consuming task, and over the last few years there has been a growing interest in developing automated tools to predict protein-HPO term annotations [10, 12, 13, 19]. In fact, automated protein-HPO term prediction was one of the tasks in the recent CAFA challenge [7].

The HPO is composed of independent sub-ontologies that describe various aspects of phenotypes [11]. The main sub-ontology is *Phenotypic abnormality* and it describes clinical abnormalities. The *Mode of inheritance* sub-ontology describes phenotypes according to inheritance patterns and contains terms such as *Autosomal dominant*. The *Mortality/Aging* sub-ontology similarly describes the age of death and contains terms such as *Neonatal death* or *Sudden death*. Finally, the *Clinical modifier* sub-ontology is composed of terms such as *Incomplete penetrance*, and describes typical modifiers of clinical symptoms [11]. Throughout this paper, we use the terms Organ, Inheritance, and Onset, for referring to the Phenotypic abnormality, Mode of inheritance and Clinical modifier, respectively. Within each sub-ontology, categories are arranged in a Directed Acyclic Graph (DAG) structure.

In our previous work, we developed PHENOstruct [10], which is the first computational method for automated prediction of protein-HPO terms. It uses a Structured Support Vector Machine (SSVM) model for predicting hierarchically consistent HPO labels. PHENOstruct employs several heterogeneous data sources as input: protein-protein interactions, disease variants, experimentally validated functional annotations, and biomedical literature, and uses protein-HPO term annotations extracted from the HPO website as the class labels. PHENOstruct used simple Bag-of-Words (BoW) features obtained from the biomedical literature in which all words occurring in the sentences that contain protein names are used as

978-1-943436-11-8 / copyright ISCA, BICOB 2018
March 19-21, 2018, Las Vegas, Nevada, USA

features. However, this is a knowledge-free approach in which information on actual phenotypes mentioned in the literature is not utilized.

Over the last few years, there have been several other automated protein-HPO term prediction tools [10, 12, 13, 19]. Notaro et al. [13] proposed a two-step method that consists of a flat learning in the first step and a hierarchical combination of the predictions in the second step. Valentini et. al. presented a novel Hierarchical Top-Down algorithm that assigns a single classifier to each HPO term and based on the hierarchical structure of DAG, it can correct the predictions [19]. Moreover, Notaro et al. proposed an algorithm that exploits the information from the ontology terms which specifies the phenotype information related to each human gene [12]. However, none of these methods use literature as their input while PHENOstruct extracts literature features using a knowledge-free approach (i.e. BoW) as opposed to a knowledge-based approach in which phenotype information is also considered.

Undoubtedly, the most comprehensive resource on biological findings, including disease-related phenotypes, is the biomedical literature. Therefore, extracting bio-entities from literature and linking them to bio-ontologies such as HPO has attracted interest within the text mining community recently [5]. This approach has high potential for exploiting the data from a variety of patient reports, case studies, and controlled trials [5]. In a related study, GOstruct 2.0 [9] utilized a natural language processing (NLP) pipeline to successfully exploit information on protein function (i.e. Gene Ontology or GO terms [1]) from the literature. Similarly, Funk et. al. [4] conducted a comprehensive study on evaluating the usage of the literature feature for the task of protein-GO term prediction. They, in addition to simple BoW features, extracted protein names and GO terms co-mentions (co-occurrences of protein names and GO terms within a short span of text) from biomedical literature, and demonstrated the utility of a knowledge-based approach for that task. Furthermore, in our previous work that uses the BoW model with PHENOstruct, we found that the majority of the most important tokens extracted from literature (i.e. the tokens that were assigned the highest weights in the trained SSVM) consist of names of proteins, genes, and diseases [10]. This suggested that applying a knowledge-based approach in extracting features would be more effective for phenotype prediction. Moreover, co-mentions have the added value that they are easy to verify by a human curator [10].

In this work, we conduct the most comprehensive evaluation of extracting literature features for the task of protein phenotype prediction. We use a knowledge-based approach and extract protein-phenotype co-mentions (co-occurrences of protein names and HPO terms within a specified span of text) from an extremely large collection of biomedical literature. Using the various co-mention features as input to PHENOstruct, we demonstrate the utility of this approach for the task of automated protein-HPO term prediction. Outcomes of this study have implications for the bio-curation community as well as text mining practitioners interested in utilizing literature for protein phenotype prediction.

The rest of the paper is organized as follows: Section 2 describes the co-mention features, the text mining pipeline used for obtaining the said features as well as the experimental setup, section 3 discusses the key observations from the experiments, and Section 4 presents conclusions and future directions.

2 Methodology

2.1 Data

In this work, we use CAFA3 Targets (http://biofunctionprediction.org) as the reference set of input proteins. We use features generated from protein-protein interactions (downloaded on 09-26-17), disease variants (downloaded on 07-05-17), experimentally validated GO annotations (downloaded on 07-21-17), as well as simple BoW features generated from biomedical literature as input for PHENOstruct (as described elsewhere [10]). Combination of variants, protein-protein iterations and GO features is referred to as VNG. In addition to the simple BoW features, we introduce novel protein-HPO term co-mention features as described below. We use UniProt synonyms of proteins to improve the coverage when extracting protein names from literature. In terms of labels, we use protein-HPO term annotations extracted from the HPO website on 07-18-17. Table 1 depicts the statistics on the HPO labels used for this study. We ignored *Mortality/Aging* sub-ontology in our experiments.

Table 1: Number of proteins, HPO categories, and annotations.

Sub-ontology	Proteins	HPO categories	Annotations
Organ	3,407	2,872	291k
Inheritance	3,049	15	10.3k
Onset	1,053	20	4.9k

2.1.1 Literature Features

We employed 27 million Medline abstracts and 1.6 million full text articles for obtaining the literature features. We generated two different sets of literature

Table 2: Statistics of co-mentions extracted from both Medline and PubMed

Organ				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	2,306	2,475	102,726	1,962,332
Paragraph-level	2,348	2,475	157152	7,292,398
Non-sentence-level	2,181	2,475	137,486	5,423,845
Inheritance				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	1,710	12	5,029	100,086
Paragraph-level	1,763	12	5,930	370,656
Non-sentence-level	1,496	12	4,929	283,740
Onset				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	399	16	1,126	14,965
Paragraph-level	511	16	1,948	74,602
Non-sentence-level	493	16	1,811	59,886

features: (1) Simple bag-of-words (BoW) features [10], (2) co-mention features. Details of the feature generation is described below.

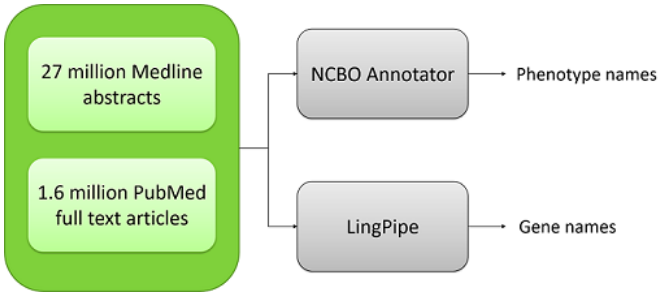

Figure 1: Overview of the NLP pipeline for extracting features from the literature.

Bag-of-words Features

Bag-of-words (BoW) is a knowledge-free feature representation which is broadly used in text mining applications. We retrieved all the sentences which had an occurrence of a protein name as candidates and extracted all the words in those sentences. For each sentence, we first lowercased all the words in the sentence and then removed the highly frequent words (stop words). All the remaining words and their counts were used as feature-value pairs. A protein is represented as a vector of variables, each of which is the count of that specific word.

Co-mentions Features

In this work, we introduce the novel protein-HPO co-mention (CoM) features which are computed from co-occurrences of the protein names and HPO terms within a specified text span. Three text spans were considered for obtaining co-mentions: sentence-level, non-sentence-level and paragraph-level. Sentence-level co-mentions (SCoM) occur in a single sentence and paragraph-level co-mentions (PCoM) are proteins and HPO terms which occur in a single paragraph (i.e. across multiple sentences). Non-sentence-level co-mentions (NSCoM) are obtained by subtracting SCoMs from PCoMs. Note that SCoMs and NSCoMs are proper subsets of PCoMs. Each protein is represented by a vector in which all the HPO terms co-occurred with the protein and their counts specify the feature-value pairs. Statistics on these co-mention features are given in Table 2.

Text Mining Pipeline

We developed the NLP pipeline shown in Figure 1 in order to obtain the BoW and CoM features. We used NCBO Virtual Appliance from BioPortal [8, 14] to extract all the phenotype names from the literature. Protein name mentions were retrieved from the literature files using LingPipe [3] trained on GeneTag [18]. In our preliminary studies, we also considered other alternatives to extract these entities such as OBO annotator [17] and Bio-Lark CR [5] for extracting phenotype names and GNormPlus [20] and ABNER [16] for extracting protein names. However, most of these systems were either difficult to access or did not provide desirable results.

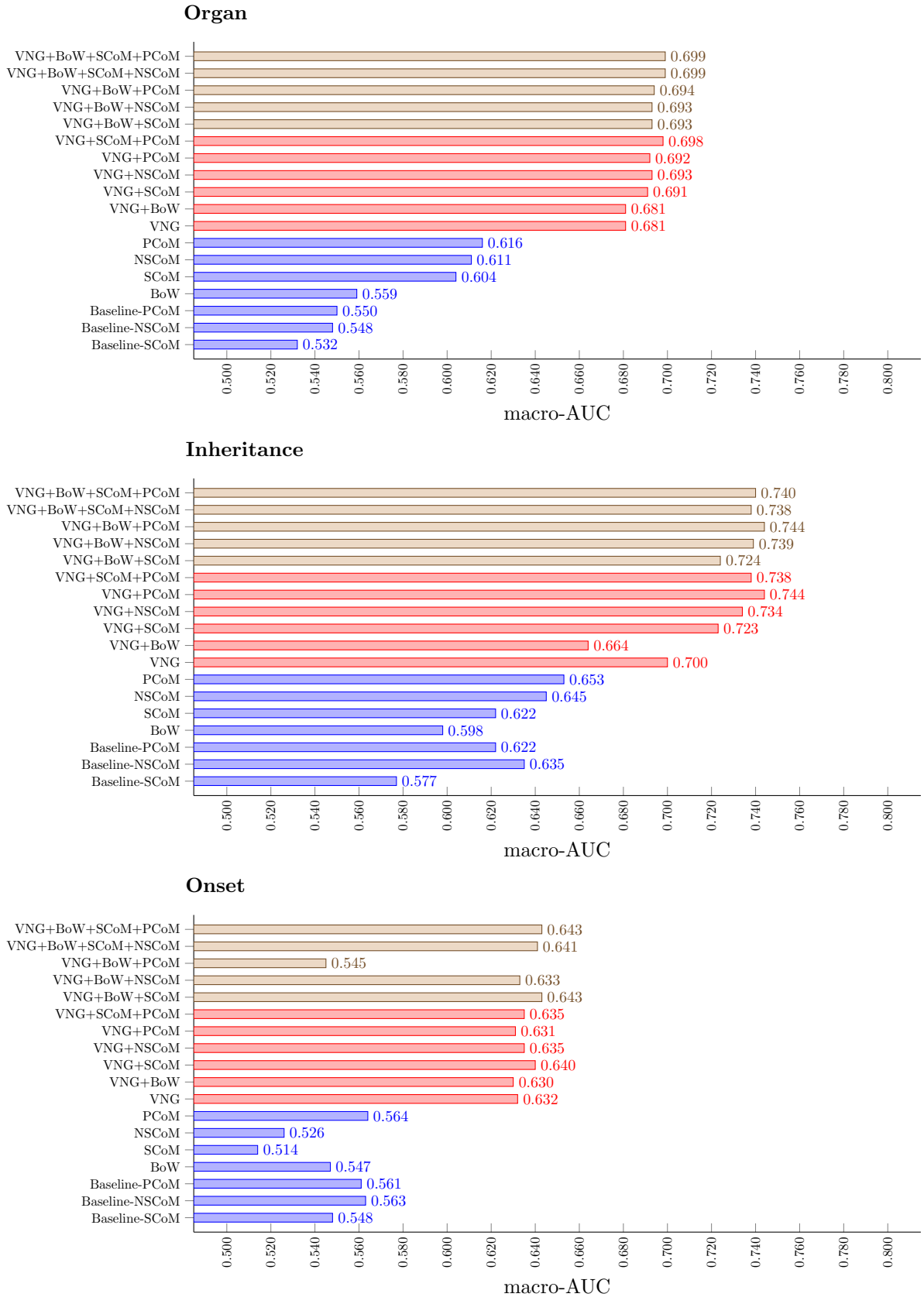

Figure 2: PHENOstruct's performance with different combinations of data sources (VNG: Variants+Network+GO).

2.2 PHENOstruct

As previously mentioned, PHENOstruct is the first computational method for automated protein phenotype prediction problem [10]. It can capture information from the inter-relationships between the HPO labels and has the advantage of not having to train multiple classifiers. Moreover, predicted labels are hierarchically consistent. PHENOstruct employs a Structured SVM model for HPO term prediction. For each test protein provided to the trained model, it outputs a set of HPO labels and corresponding confidence scores.

2.3 Experimental Setup

In order to establish baselines, we utilized the SCoMs, NSCoMs, and PCoMs as the final predictions themselves (i.e. without any machine learning) along with their associated frequencies as the confidence scores. As mentioned before, PHENOstruct provides confidence scores for each prediction. Considering the structure of HPO, all HPO annotations and predictions are expanded via the *true path rule* to the root node of HPO. This rule states that any annotation to a certain term implicitly indicates annotations to all its ancestors. Macro-AUROC (Area Under the Receiver Operative Curve) was used as the evaluation measure for the predictions. We use a five-fold cross-validation setting for all our experiments. Separate experiments were carried out for each sub-ontology. In order to compare the experiments, we compute p-values using paired t-tests by considering only the leaves. All the experiments were performed on a system running Linux Fedora 26 with a 24-cores processor and 128 GB of memory. The average running time for each experiment on the Organ, Inheritance, and Onset sub-ontologies are 36 hours, 20 minutes, and 5 minutes, respectively. Note that PHENOstruct is not compared to other protein-phenotype prediction tools [10, 12, 13, 19] because of the availability of only research-quality code.

3 Results and Discussion

In order to evaluate the effectiveness of the newly introduced co-mention features, a set of ablation studies were carried out by feeding different combinations of features into PHENOstruct as input. Our experimental results demonstrate that, when using individual co-mention features as the input, the paragraph-level co-mentions (PCoMs) provide the best performance in all three sub-ontologies (see Figure 2). PCoMs consistently beat SCoMs and NSCoMs for all three sub-ontologies. These observations suggest that the paragraph level is the span that is overall better if you are interested in using a single set of co-mention features. Bada et al. [2] describes that co-mentions do not necessarily occur in the same sentence; this may be the justification for the relatively superior performance of PCoMs.

Moreover, PCoMs by themselves consistently outperform BoWs in all three sub-ontologies suggesting that knowledge-based approaches are better than knowledge-free methods (P-values for Organ, Inheritance, and Onset are 7.8E-31, 7.6E-01, and 6.2E-01, respectively). This observation is also true for both SCoMs and NSCoMs in many of the cases. Moreover, co-mention features combined with other data sources give better performance compared to using BoW features combined with other data sources.

Another key observation is that, as expected, SCoMs, NSCoMs, and PCoMs outperform the baseline-SCoM, baseline-NSCoM, and baseline-PCoM, respectively, in both Organ and Inheritance sub-ontologies. However, this is not the case in Onset sub-ontology; further investigation is required for finding the underlying reason.

Literature features by themselves provide lower performance compared to when using them in conjunction with other types of data sources. However, the best performance in all three sub-ontologies is obtained by using literature features along with other data sources (P-values for Organ, Inheritance, and Onset are 1.7E-61, 1.4E-01, and 2.4E-02, respectively). This suggests literature features are highly complementary to other data sources.

Regardless of the data sources/features used, aligning with our previous experimental observations [10], PHENOstruct provides best performance in the Inheritance sub-ontology closely followed by the Organ sub-ontology.

4 Conclusion and Future Work

In this work we conducted a comprehensive study on evaluating a variety of literature features for the task of protein phenotype prediction using PHENOstruct. We demonstrate that knowledge-based features (i.e. co-mention features) helps improve the overall performance and are more effective than their knowledge-free counterparts. Moreover, we find that paragraph span best serves this purpose; PCoMs are the most valuable source of information in comparison with the other literature feature sets. However, we conclude that using PCoMs as an individual data source does not provide the best performance, and it needs to be used as a complementary set of features to obtain the most optimum performance.

This study opens up many other avenues for future investigation. By carefully analyzing the the performances of baselines, we notice that our NLP pipeline is generating a large number of false positives (data not shown). In other words, not every co-occurrence of a protein and a phenotype represent a valid relationship. Since our current work does not consider the context surrounding these entity words, the next step would be to develop a context-sensitive co-mention filter/classifier for removing these false positives and improving the overall quality of generated co-mentions. Moreover, this classifier by itself can serve as an important component in a fully automated biocuration pipeline for phenotypes.

References

[1] Michael Ashburner et al. Gene Ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[2] 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. *BioLink–an ISMB Special Interest Group. Berlin, Germany: Proceedings of BioLINK SIG*, 2013:13–19, 2013.

[3] Bob Carpenter. LingPipe for 99.99% recall of gene mentions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume 23, pages 307–309, 2007.

[4] 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.

[5] Tudor Groza et al. Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database*, 2015:bav005, 2015.

[6] 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.

[7] Yuxiang Jiang et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):184, 2016.

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

[9] Indika Kahanda and Asa Ben-Hur. GOstruct 2.0: Automated Protein Function Prediction for Annotated Proteins. In *BCB*, 2017.

[10] Indika Kahanda, Christopher Funk, Karin Verspoor, and Asa Ben-Hur. PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources. *F1000Research*, 2015.

[11] Sebastian Köhler et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876, 2017.

[12] Marco Notaro et al. Ensembling Descendant Term Classifiers to Improve Gene–Abnormal Phenotype Predictions. *Proceedings of CIBB*, page 1, 2017.

[13] Marco Notaro, Max Schubach, Peter N Robinson, and Giorgio Valentini. Prediction of Human Phenotype Ontology terms by means of hierarchical ensemble methods. *BMC bioinformatics*, 18(1):449, 2017.

[14] Natalya F Noy et al. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic acids research*, 37(suppl_2):W170–W173, 2009.

[15] Anika Oellrich et al. The digital revolution in phenotyping. *Briefings in bioinformatics*, 17(5):819–830, 2015.

[16] Burr Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[17] Maria Taboada, Hadriana Rodríguez, Diego Martínez, María Pardo, and María Jesús Sobrido. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014, 2014.

[18] Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics*, 6(1):S3, 2005.

[19] Giorgio Valentini et al. Prediction of Human Gene-Phenotype Associations by Exploiting the Hierarchical Structure of the Human Phenotype Ontology. In *IWBBIO (1)*, pages 66–77, 2015.

[20] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015, 2015.