ProPheno 1.0: An Online Dataset for Accelerating the Complete Characterization of the Human Protein-phenotype Landscape in Biomedical Literature

Morteza Pourreza Shahri[1] **and Indika Kahanda**[1]

[1]**Gianforte School of Computing, Montana State University, USA**

Corresponding author:
Indika Kahanda[1]

Email address: indika.kahanda@montana.edu

ABSTRACT

Identifying protein-phenotype relations is of paramount importance for applications such as uncovering rare and complex diseases. One of the best resources that captures the protein-phenotype relationships is the biomedical literature. In this work, we introduce ProPheno, a comprehensive online dataset composed of human protein/phenotype mentions extracted from the complete corpora of Medline and PubMed Central Open Access. Moreover, it includes co-occurrences of protein-phenotype pairs within different spans of text such as sentences and paragraphs. We use ProPheno for completely characterizing the human protein-phenotype landscape in biomedical literature. ProPheno, the reported findings and the gained insight has implications for (1) biocurators for expediting their curation efforts, (2) researches for quickly finding relevant articles, and (3) text mining tool developers for training their predictive models. The RESTful API of ProPheno is freely available at http://propheno.cs.montana.edu.

INTRODUCTION

Proteins are the workhorses of life, and they perform a wide range of operations in cells. Thousands of proteins work together to provide the functionality of cells. However, when a gene is mutated, a malfunction of protein may occur which can lead to a genetic disorder (NIH, 2018).

The observation of phenotypes is important in studying genetic disorders. In the medical context, a phenotype can be characterized as a deviation from normal morphology or behavior (Robinson, 2012). The study of phenotypes in medicine consists of detailed understanding of the phenotypic abnormalities associated with each disease (Robinson, 2012). Variations of genes and proteins cause functional changes and identifying the effects of their mutations is necessary for understanding the resulting phenotype, i.e. the observed disease state (Baker and Rebholz-Schuhmann, 2009). Furthermore, many patients suffer from rare diseases caused by genomic variants, i.e. diseases caused by disruptions in regular gene expression. However, many variants are quite rare, making genotype-phenotype correlations dubious and clinical interpretations difficult (Firth et al., 2009). One way to increase certainty would be identifying patients who share the same or overlapping gene variants and phenotypic characteristics (Firth et al., 2009). Therefore, finding relations between proteins and phenotypes can be considered vital for applications such as finding treatments and cures for rare diseases.

Since biologists, researchers, and scientists report their findings and observations from wet-lab experiments and clinical studies in biomedical literature, it can be considered one of the most valuable resources for extracting protein-phenotype relations. Therefore, exploring the feasibility of extracting relations between proteins and phenotypes mentioned in biomedical literature through text mining and machine learning has recently gained significant attention (Singhal et al., 2016; Korbel et al., 2005; Goh et al., 2006; Khordad and Mercer, 2017). However, the first main hurdle in developing text mining models is the absence of a gold-standard dataset of bio-entity mentions, i.e. proteins and phenotypes, which

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27479v2 | CC BY 4.0 Open Access | rec: 2 May 2019, publ: 2 May 2019

encompasses the complete corpora of published biomedical articles. In this work, we attempt to resolve this issue by developing such a comprehensive dataset.

Human Phenotype Ontology (HPO) is a standardized vocabulary which covers a wide range of phenotype abnormalities observed in human diseases (Köhler et al., 2013). HPO is composed of five sub-ontologies among which *Phenotypic abnormality* is the main sub-ontology which describes clinical abnormalities. Each sub-ontology is organized in a hierarchical structure where more general terms are close to the top while more specific terms are closer to the bottom. Each pair of terms in the hierarchy are linked with a *is-a* relationship. In this paper, we use *phenotypes* and *HPO terms*, interchangeably.

HPO website[1] provides gold-standard annotations for a large collection of human proteins. However, currently, only a small portion of known human proteins have HPO annotations (Köhler et al., 2013). But, it is believed that there are many other human proteins which are associated with diseases and hence should be annotated with HPO terms (Peter Robinson, personal communication, 2015). Biocuration, which is the process of extracting knowledge from unstructured text and storing the data in knowledge bases, is the primary technique for expanding the biological knowledge bases such as the HPO database. Biocuration is usually performed manually with the help of computational tools (International Society for Biocuration, 2018). This process is considered tedious and the computational tools are also evolving in terms of efficiency and accuracy. One of the the first steps in bicuration of human protein-phenotype information is identifying and extracting protein and phenotype names from biomedical literature. Currently, there is no publicly available dataset of proteins and phenotype names that covers the majority of the entire biomedical literature corpora.

In this paper, we introduce ProPheno, which is an online and publicly accessible dataset composed of proteins, phenotypes (HPO terms), and their co-occurrences (co-mentions) in text which are extracted from Medline abstracts and PubMed Central (PMC) Open Access full-text articles using a sophisticated in-house developed text mining pipeline. This dataset covers all terms in the *Phenotypic abnormality* sub-ontology. Using the ProPheno data, we also conduct a comprehensive characterization of the protein-phenotype landscape in biomedical literature. The findings from this characterization has implications for biocurators and researchers working in related fields as well as practitioners in the area of developing automated text mining pipelines for biocuration. One direct application of the ProPheno data is the generation of features for computational models that can predict phenotypes for human proteins; this task is one of the subtasks in the CAFA (The Critical Assessment of protein Function Annotation algorithms) competition (Jiang et al., 2016). In a recent study (Pourreza Shahri and Kahanda, 2018), we demonstrated that co-mentions of proteins and phenotypes (extracted from ProPheno dataset) play an important role in improving the prediction of HPO terms associated with a given protein. ProPheno is accessible through a RESTful API, which can be used in many programming language as well as a web interface (online demo version) for online access.

METHOD

We developed the text mining pipeline[2] shown in Figure 1, which is an extension of the pipeline described elsewhere (Pourreza Shahri and Kahanda, 2018). This pipeline extracts proteins/phenotypes and their co-mentions by consuming 27,590,898 Medline abstracts (downloaded on 07/01/2017) and 1,873,381 PMC full-text articles (downloaded on 3/15/2018) as the input. PMC full-text articles were downloaded from the PubMed website in XML format. A small portion of these full-text articles (140,370) did not contain any text, so we removed them from consideration. We employed *PubMed XML Parser* (Achakulvisut and Acuna, 2015) to extract the paragraphs from the remaining 1,733,012 full-text articles, and stored the corresponding paragraphs in separate files. It is worth mentioning that to avoid duplicating abstracts, we do not take the abstract section of full-text articles into account.

We employed NCBO Virtual Appliance (NCBO Annotator) from BioPortal (Jonquet et al., 2009; Noy et al., 2009) for extracting HPO terms from the literature. Protein mentions were retrieved from the literature using LingPipe (Carpenter, 2007). We used UniProt (Consortium, 2018) synonyms[3] of proteins to improve the coverage when extracting proteins from literature.

We also considered other alternatives to extract these entities such as OBO annotator (Taboada et al., 2014) and Bio-Lark CR (Groza et al., 2015a) for extracting phenotype names, and GNormPlus (Wei

[1] https://hpo.jax.org/app

[2] Image created using https://www.draw.io

[3] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27479v2 | CC BY 4.0 Open Access | rec: 2 May 2019, publ: 2 May 2019

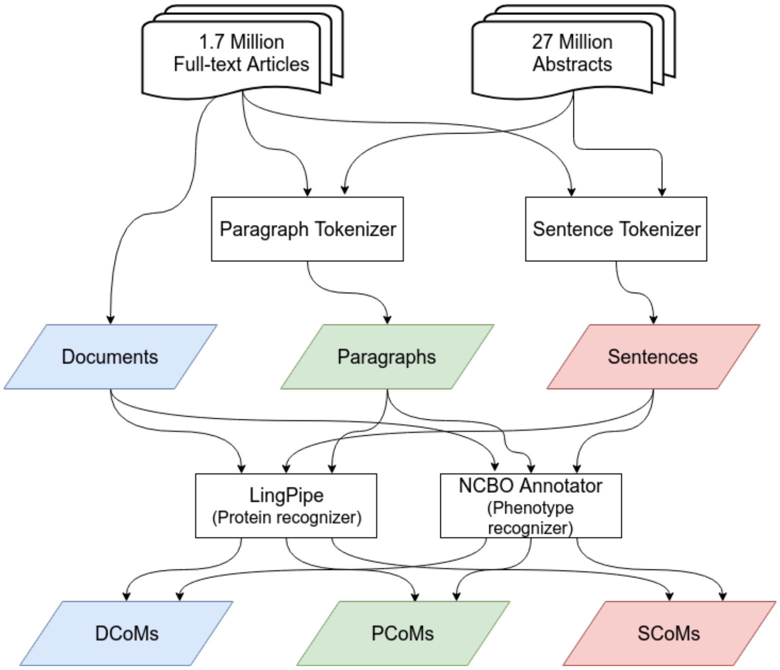


Figure 1. Overview of the text mining pipeline for extracting protein and phenotype names as well as their co-occurrences within different spans of text.

et al., 2015) and ABNER (Settles, 2005) for extracting protein names. However, most of these systems either did not provide desirable results or were difficult to access or use. We selected the list of phenotype extraction tools from a study by Groza et al. (Groza et al., 2015b).

The introduced dataset comprises UniProt identifiers and HPO identifiers for proteins and phenotypes extracted from biomedical literature, respectively, which are categorized by the abstracts and full-text articles. This dataset also provides co-mentions of proteins and phenotypes in different spans of text. We used three spans: (1) sentence-level co-mentions (SCoM) which occur in a single sentence (2) paragraph-level co-mentions (PCoM) which occur in a single paragraph (i.e. across multiple sentences), and (3) document-level co-mentions (DCoM) which occur in a single document (i.e. across multiple paragraphs).

We have previously worked on extracting co-mentions in a limited setting (Pourreza Shahri and Kahanda, 2018) in which we showed that SCoMs and PCoMs can improve the performance of protein-phenotype prediction tools. In an effort to primarily make these co-mentions readily available for other researchers in the filed, ProPheno was born. Figure 2 shows an example of a co-mention of a protein and a phenotype in a sentence (i.e. a sentence-level co-mention). Figure 3 is a screenshot of ProPheno demo version, which shows the list of first 10 HPO terms extracted from abstracts. In this figure, each row depicts an occurrence of a phenotype mentioned in text, and "Start Location" and "End Location" indicated the actual location of occurrence.

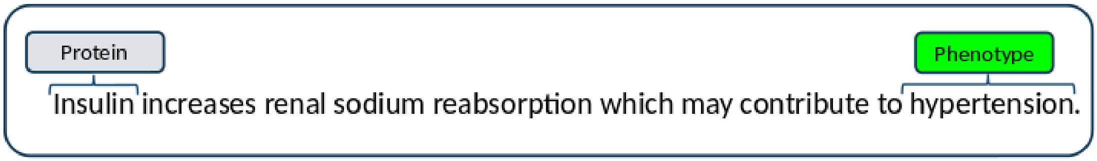
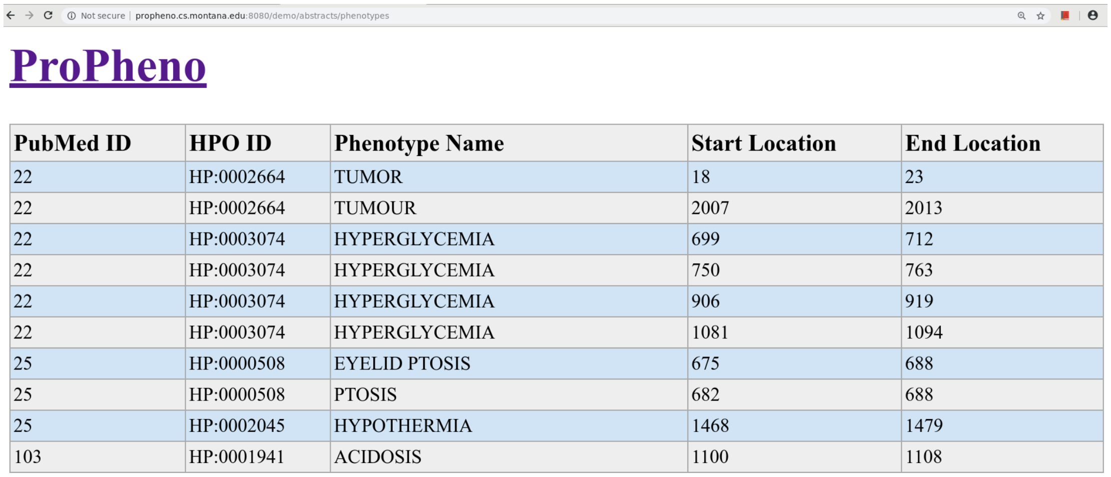

Protein
Phenotype
Insulin increases renal sodium reabsorption which may contribute to hypertension.

Figure 2. An example of a sentence-level protein-phenotype co-mention which is extracted from the article PMID: 10855734. The corresponding UniProt ID and HPO ID for the protein and the phenotype are P01308 and HP:0000822, respectively.



ProPheno

PubMed ID	HPO ID	Phenotype Name	Start Location	End Location
22	HP:0002664	TUMOR	18	23
22	HP:0002664	TUMOUR	2007	2013
22	HP:0003074	HYPERGLYCEMIA	699	712
22	HP:0003074	HYPERGLYCEMIA	750	763
22	HP:0003074	HYPERGLYCEMIA	906	919
22	HP:0003074	HYPERGLYCEMIA	1081	1094
25	HP:0000508	EYELID PTOSIS	675	688
25	HP:0000508	PTOSIS	682	688
25	HP:0002045	HYPOTHERMIA	1468	1479
103	HP:0001941	ACIDOSIS	1100	1108

Figure 3. The list of first 10 HPO terms mentioned in the first abstracts ordered by PubMedID shown in the online demo version. Each row shows an occurrence of an HPO term in an abstract. ”Start Location” and ”End Location” show the position of the matched phenotype in text, e.g. Start Location = 98 and End Location = 103 show that the phenotype has been mentioned in text starting from index 98 and ends at index 103.

RESULTS

In this section we present detailed statistics and analysis on the mentions of proteins and phenotypes in different spans of text.

Proteins and Phenotypes

Table 1 shows the total number of the unique and all proteins and phenotypes in abstracts and full-text articles. Since full-text articles include more details about proteins and phenotypes, intuitively, we expect more terms in full-text articles. The average number of proteins and phenotypes in abstracts and full-text articles supports this claim. By looking at the unique number of proteins and phenotypes in abstracts versus full-text articles, we observe that there are a few proteins and phenotypes which were extracted only from either abstracts or full-text articles. In this study, we consider various combination of words which are detected by NCBO Annotator that are matched with a term in the HPO database. For instance, NCBO Annotator returns ”prostate cancer”, i.e. HP:0012125, and ”cancer”, i.e. HP:0002664, from ”... prostate cancer ...”, and both words can be matched with corresponding HPO terms in the HPO database. This can be attributed as the main reason for having a large number of extracted phenotypes. However, the number of unique phenotype names is relatively low in comparison with the total number of phenotypes in the HPO dataset. The main reason for this discrepancy is that we chose to work with the phenotypes for which there are 10 or more annotations in the HPO database.

Table 1. Stats of Protein/Phenotype Mentions in Biomedical Literature

	Abstracts	Full-texts	Total
Unique proteins	2,178	2,389	2,512
All proteins	1,807,246	2,173,695	3,980,941
Avg. number of proteins	2.11	4.63	-
Unique phenotypes	2,224	2,227	2,277
All phenotypes	30,954,930	32,639,095	63,594,025
Avg. number of phenotypes	3.58	24.8	-

Figures 4 and 5 show the distribution of unique proteins and phenotypes in abstracts and full-text articles, respectively[4]. As mentioned before, there are a few proteins and phenotypes which were extracted from either abstracts or full-text articles. We also observe that our pipeline can detect 74% of the proteins and 92% of the phenotypes curated in the HPO database.

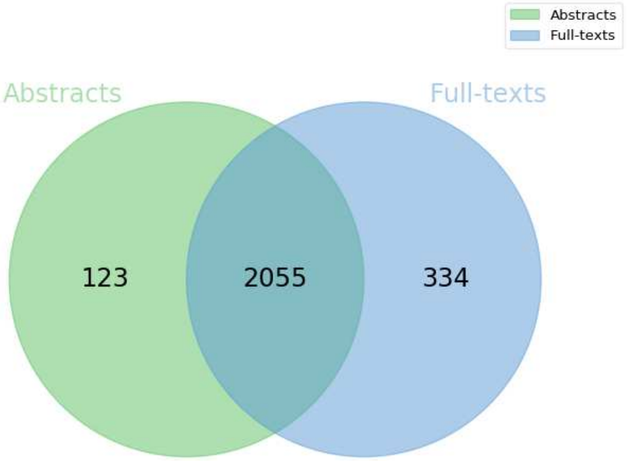


Figure 4. The Distribution of Unique Proteins in Abstracts and Full-texts

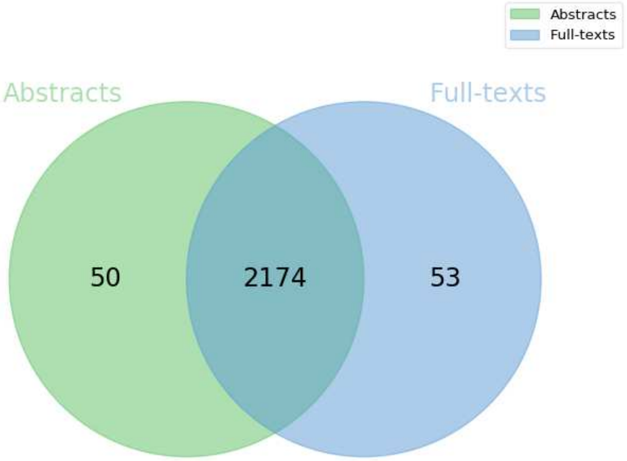


Figure 5. The Distribution of Unique Phenotypes in Abstracts and Full-texts

[4] Diagrams generated by https://github.com/tctianchi/pyvenn

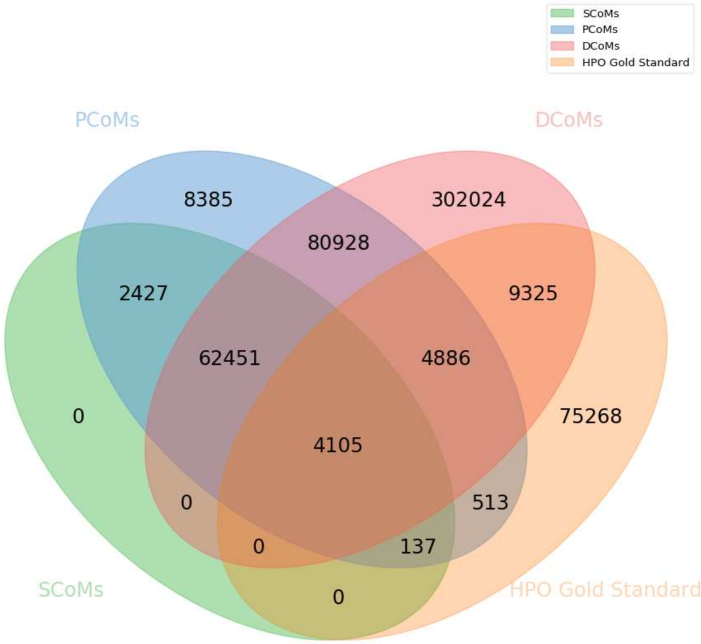


Figure 6. The Distribution of Unique Pairs of Co-mentions in SCoMs, PCoMs, DCoMs, and protein-HPO annotations curated in the HPO database. Each of the ovals shows the number of common unique co-mentions between the four data sources.

Co-mention of Proteins and Phenotypes in Text

As mentioned in the previous section, co-mentions are defined as the co-occurrences of proteins and phenotypes within a certain span of text extracted from our text mining pipeline. Statistics on these co-mentions in various spans of text, i.e. sentences, paragraphs, and full-text documents, are given in Table 2. The distribution of unique pairs of proteins and phenotypes in SCoMs, PCoMs, DCoMs, and the HPO gold standard are shown in Figure 6. We observe that 75,268 unique pairs in the HPO gold standard were not extracted by our pipeline. There are two possible reasons that can prevent the pipeline to extract these pairs. First, either LingPipe or NCBO Annotator may have failed to extract the corresponding proteins or phenotypes, respectively. Second, they may have been either not co-occurred in at least a document or one entity is mentioned in the abstract while the other entity is in the body of the article. Besides, there are 10,812 unique pairs in PCoMs (2,427 in common with SCoMs) which are not recognized by the DCoMs. This is because we treat the abstracts as paragraphs, and since we only have one copy of the abstracts, DCoMs do not contain those unique pairs.

Analysis of Protein and Phenotype Named Entities

Figure 7 shows the distribution of protein names in SCoMs, PCoMs, and DCoMs. The number of less frequent proteins in SCoMs is more than proteins with high frequency, and the number of proteins in for each frequency in SCoMs is higher than corresponding number in PCoMs and DCoMs. This observation suggests that the larger spans of text are able to identify more proteins.

Table 2. Statistics of Co-mentions Extracted from both Medline and PMC. We consider abstracts as paragraphs, so we do not have Document-level information for the abstracts.

Full-text articles				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	1,845	1,498	49,686	693,005
Paragraph-level	2,185	1,896	122,522	4,323,395
Document-level	2,362	2,126	463,719	99,818,140

Abstracts				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	1,684	1,461	42,774	496,969
Paragraph-level	1,975	1,827	102,881	4,116,999
Document-level	-	-	-	-

All				
Span	Unique proteins	Unique HPO terms	Unique co-mentions	Total co-mentions
Sentence-level	1,998	1,623	69,120	1,189,974
Paragraph-level	2,313	1,976	225,403	8,440,394
Document-level	2,362	2,126	436,719	99,818,140

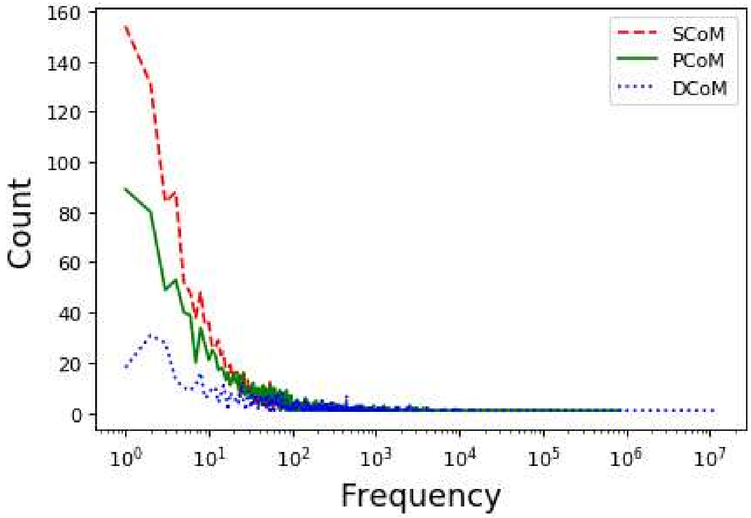


Figure 7. The Distribution of Protein Mentions

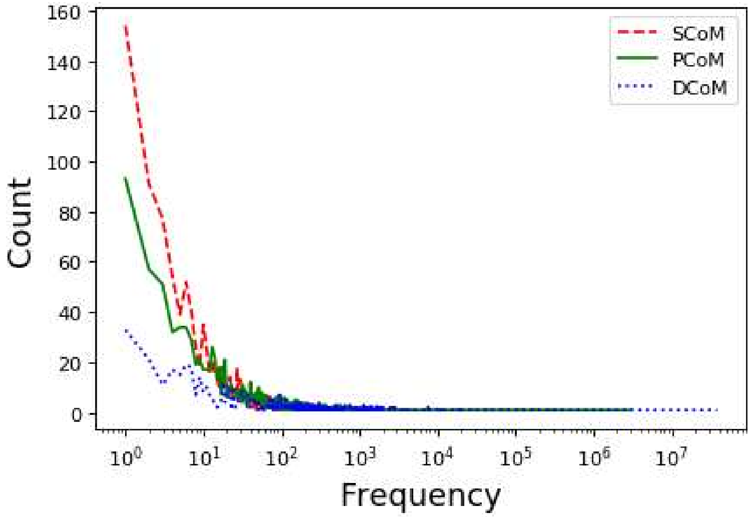


Figure 8. The Distribution of HPO-term Mentions

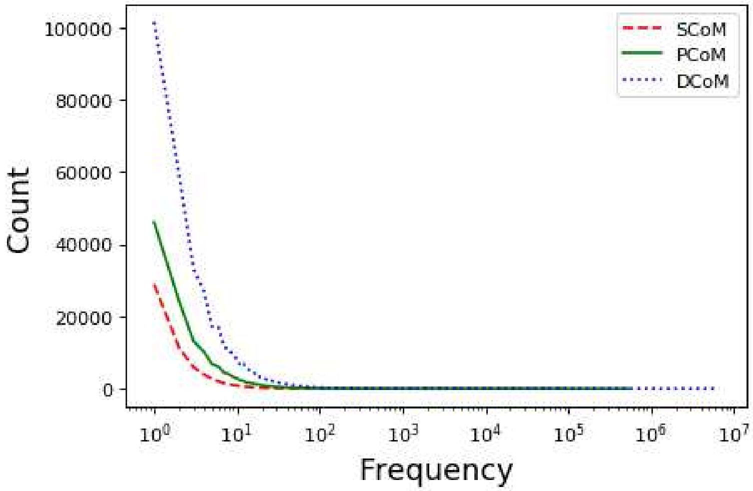


Figure 9. The Distribution of the Unique pair mentions

The same can be observed in Figure 8 which shows distribution of HPO terms in SCoMs, PCoMs, and DCoMs. However, in Figure 9 which shows the distribution of unique protein and HPO-term pairs, occurrences of unique pairs for each frequency in DCoMs is higher than corresponding values in SCoMs and PCoMs. One of the reasons of this observation can be that many of the unique pairs occur in sentences and do not occur more in larger spans of text. Therefore, larger spans of text have more occurrences of unique pairs for every frequency.

Figure 10 shows the distribution of the depths (with respect to the hierarchy) of all HPO terms recognized in the literature. This plot conveys that more specific HPO terms are less frequent in the literature. One of the reasons of this observation can be that there are less number of mentions of more specific terms in the literature. Figure 11 demonstrates the distribution of unique HPO terms detected in the literature and HPO terms from the HPO gold standard. We observe that our pipeline is able to detect the majority of HPO terms from the HPO gold standard.

In addition, the evolution of the protein-phenotype landscape in biomedical literature is shown in Figure 12. We observe a slight increase in the number of unique proteins, and a relatively higher increase in the number of unique HPO terms between 2009 and 2018.

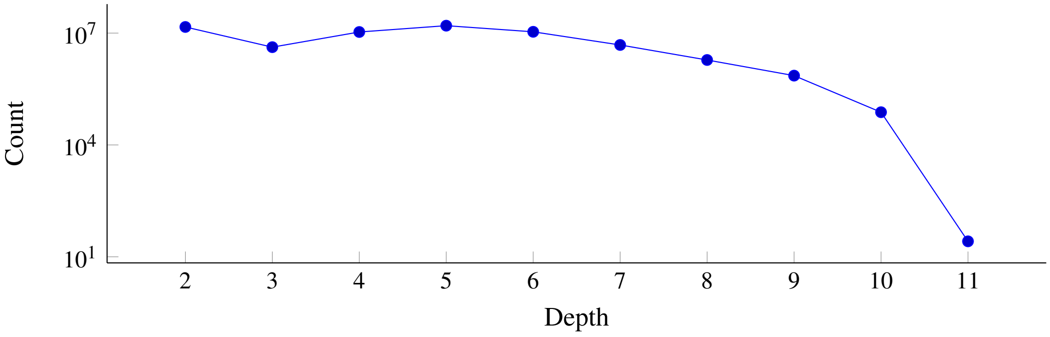


Figure 10. Distribution of the depths of terms detected in literature

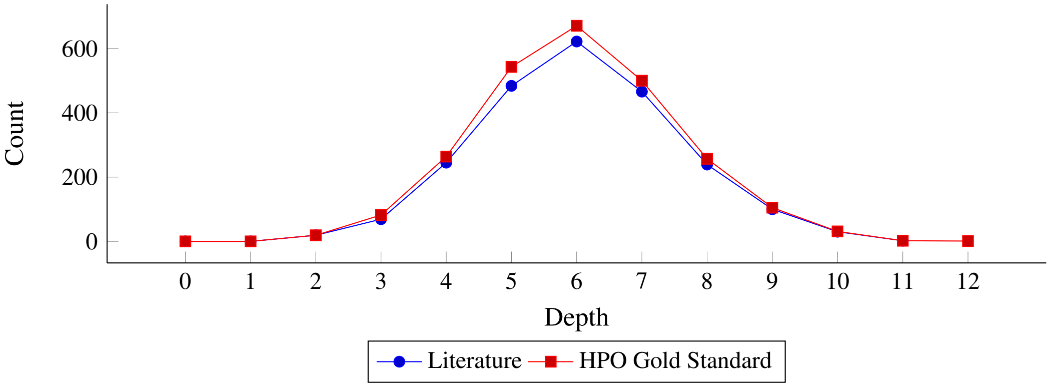


Figure 11. Distribution of the Depths of Unique Terms Detected in Literature and HPO Gold Standard

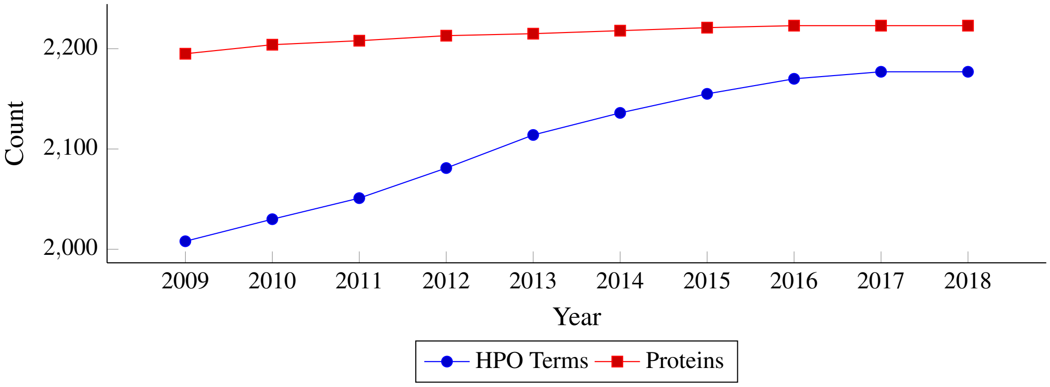


Figure 12. The evolution of the protein-phenotype landscape in Biomedical Literature

CONCLUSIONS AND FUTURE WORK

In this paper we presented a dataset of proteins and phenotypes (HPO labels) in the entire biomedical corpora derived from Medline abstracts and PubMed Open Access full-text articles. This dataset was generated using an expanded text mining pipeline from one of our previous work (Pourreza Shahri and Kahanda, 2018).

We also reported detailed analysis and statistics on the mentions of proteins and phenotypes in the entire corpora along with co-mentions of these entities in various spans of biomedical text. Additionally, we presented the evolution of biomedical literature over the period of 2009-2018.

In this study, we used bio-entity recognizers which had shown good performance in identifying bio-entities in biomedical text. However, bio-entity recognition is still a challenging problem. Consequently, advances in the performance of bio-entity recognizers would enhance the ability of our pipeline for correctly detecting more proteins and phenotypes in text.

Despite detecting millions of proteins and phenotypes in text, it is understood that a portion of the co-mentions are false positives (the mere occurrence of two entities within a certain span does not constitute a valid relationship). Therefore, the next step would be developing a context-sensitive co-mention classifier or a filter to remove these false positives. The various sections in a published article include different types of information (e.g. information contained in the introduction versus results sections); this high-level location information about co-mentions could be beneficial for biocuration. Therefore, a future step is to include the section labels for each entity pair. This will provide the capability of listing the co-mentions in a ranked order according to the confidence scores predicted by the classifier. In this study, we limit the scope to the Phenotypic Abnormality sub-ontology. In the future, we plan to include the other sub-ontologies of HPO. Moreover, in this study we considered all the phenotypes extracted using NCBO annotator (which have 10 or more annotations in the HPO database). This includes both abstract (i.e. general) terms and specific terms. In the future, we will filter out the more abstract terms and keep the more specific terms.

Since the biomedical literature is expanding exponentially, we plan to develop an automated online tool to extract all types of co-mentions on demand, and remove the need to update the dataset regularly.

ACKNOWLEDGMENTS

We thank Scott Dowdle, the Systems Manager of the Gianfforte School of Computing at Montana State University, for setting up the ProPheno server.

REFERENCES

Achakulvisut, T. and Acuna, D. E. (2015). Pubmed Parser. `http://doi.org/10.5281/zenodo.159504`.

Baker, C. J. and Rebholz-Schuhmann, D. (2009). Between proteins and phenotypes: annotation and interpretation of mutations.

Carpenter, B. (2007). LingPipe for 99.99% recall of gene mentions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume 23, pages 307–309.

Consortium, U. (2018). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515.

Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R. M., and Carter, N. P. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533.

Goh, C.-S., Gianoulis, T. A., Liu, Y., Li, J., Paccanaro, A., Lussier, Y. A., and Gerstein, M. (2006). Integration of curated databases to identify genotype-phenotype associations. *BMC genomics*, 7(1):257.

Groza, T. et al. (2015a). Automatic concept recognition using the Human Phenotype Ontology reference and test suite corpora. *Database*, 2015:bav005.

Groza, T., Köhler, S., Doelken, S., Collier, N., Oellrich, A., Smedley, D., Couto, F. M., Baynam, G., Zankl, A., and Robinson, P. N. (2015b). Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015.

International Society for Biocuration (2018). Biocuration: Distilling data into knowledge. *PLoS biology*, 16(4):e2002846.

Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., et al. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome biology*, 17(1):184.

Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The open biomedical annotator. *Summit on translational bioinformatics*, 2009:56.

Khordad, M. and Mercer, R. E. (2017). Identifying genotype-phenotype relationships in biomedical text. *Journal of biomedical semantics*, 8(1):57.

Köhler, S., Doelken, S. C., et al. (2013). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, 42(D1):D966–D974.

Korbel, J. O., Doerks, T., Jensen, L. J., Perez-Iratxeta, C., Kaczanowski, S., Hooper, S. D., Andrade, M. A., and Bork, P. (2005). Systematic association of genes to phenotypes by genome and literature mining. *PLoS biology*, 3(5):e134.

NIH (2018). How can gene mutations affect health and development? `https://ghr.nlm.nih.gov/primer/mutationsanddisorders/mutationscausedisease`.

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

Pourreza Shahri, M. and Kahanda, I. (2018). Extracting co-mention features from biomedical literature for automated protein phenotype prediction using PHENOstruct. In *10th International Conference on Bioinformatics and Computational Biology, BICOB 2018*, pages 123–128.

Robinson, P. N. (2012). Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–780.

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

Singhal, A., Simmons, M., and Lu, Z. (2016). Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine. *PLoS computational biology*, 12(11):e1005017.

Taboada, M., Rodríguez, H., Martínez, D., Pardo, M., and Sobrido, M. J. (2014). Automated semantic annotation of rare disease cases: a case study. *Database*, 2014.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27479v2 | CC BY 4.0 Open Access | rec: 2 May 2019, publ: 2 May 2019