UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

To be or NOT to be: The Impact of Negative Annotation in Biomedical Semantic Similarity

Lina Andreia Gama Aveiro

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Professora Doutora Cátia Luísa Santana Calisto Pesquita

2021

Resumo Alargado

A investigação biomédica gera um enorme fluxo de dados anualmente, com uma tendência crescente. A única forma de analisar esta quantidade de dados de forma rápida e eficiente é utilizando métodos digitais, contudo dados biomédicos são gerados em linguagem natural, que na sua forma não tratada, não é compatível com computação. Uma das características da linguagem natural que é de difícil tradução é o significado das palavras. Para guardar a informação de forma correta, sinónimos, por exemplo, têm de ser considerados como semelhantes apesar das palavras poderem ser totalmente diferentes.

As ontologias biomédicas reúnem as condições necessárias para o armazenamento de dados biomédicos de forma acessível a homem e máquina, por isso são muito usadas para resolver este problema. Técnicas de semelhança semântica utilizam as características inerentes da ontologia, como hierarquia, para retornar um valor numérico que reflete a proximidade de significado entre conceitos. As ontologias são utilizadas também para anotar dados, ou seja, associar entidades com as classes presentes na ontologia, incluindo informação sobre esta, as evidências consideradas, quem a criou e quando. As anotações têm diversas aplicações e motivaram o desenvolvimento de novas ferramentas para estruturar conhecimento e analisá-lo.

Apesar das anotações serem cruciais no desenvolvimento da bioinformática, as ontologias biomédicas continuam incompletas num aspeto chave, na presença de anotações negativas. Usualmente as anotações baseiam-se em associar um facto com uma entidade – anotação positiva - mas as anotações que descrevem que um facto não está associado com uma entidade, ou anotações negativas, são raras nas maiores ontologias biomédicas. O número de anotações negativas é teoricamente muito maior do que o número de anotações positivas, contudo anotações negativas são raras em bio-ontologias. O problema é que as ontologias não estão a considerar a Open World Assumption, onde a falta de uma anotação não pode ser interpretada como sendo uma anotação negativa. Os métodos existentes para o cálculo de semelhança semântica normalmente ignoram as anotações negativas e utilizam apenas as anotações positivas para evitar esta limitação. Contudo, devido à dificuldade em distinguir informação negativa de informação desconhecida, a geração de exemplos negativos para metodologias que precisam deles, como aprendizagem automática, fica comprometida.

Para compreender o impacto desta limitação, neste trabalho as medidas de semelhança semântica pairwise Best-Match Average(BMA) e Resnik foram adaptadas para desenvolver PolarBMA e PolarResnik. Estas utilizam tanto anotações positivas como negativas para o cálculo de semelhança semântica. Uma metodologia foi desenvolvida, em que ambas as medidas são usadas em combinação. A PolarResnik é utilizada para obter a semelhança de classes e em seguida a PolarBMA utiliza essa semelhança para calcular a semelhança semântica final. A dificuldade em incorporar anotações negativas no cálculo de semelhança semântica é causada pela herança das anotações negativas ser oposta à das anotações positivas. Enquanto que, para as anotações positivas, anotação a uma dada classe implica anotação a todas as superclasses, para as anotações negativas anotação a uma classe implica anotação a todas as suas subclasses. A medida PolarResnik considera esta informação no cálculo de semelhança de classes, e depois

a PolarBMA utiliza estes valores e calcula a semelhança semântica penalizando os pares de entidades com polaridades opostas.

Estas medidas foram testadas em duas aplicações muito importantes para a área de bioinformática, a previsão de interação proteína-proteína e previsão de doença. Para o primeiro caso, foram escolhidos pares de proteínas de um dataset curado com uma score de interação, de forma a que se soubesse se os pares são compostos por proteínas que interagem ou que não interagem. As proteínas foram depois enriquecidas com anotações retiradas da Gene Ontology para que cada uma possa ser representada como um conjunto das suas anotações. Devido à raridade de anotações negativas presentes na ontologia, foram geradas anotações negativas, com diferentes níveis de frequência e para cada ramo da Gene Ontology: Componente Celular, Função Molecular e Processo Biológico.

Para a previsão de doença, foram selecionadas trinta e três doenças mendelianas, e 50 pacientes sintéticos foram criados para cada uma. Os fenótipos associados às doenças têm diferentes penetrâncias, que dependem da frequência com que esses fenótipos se manifestam. A penetrância varia com o sexo, por isso foi gerado, com igual probabilidade, um género para cada paciente. Estes foram depois enriquecidos com anotações dos fenótipos consoante os da doença descritos na Human Phenotype Ontology, e a penetrância de cada um. Foram também geradas anotações negativas para cada paciente a partir das anotações negativas existentes nas doenças. Foram ainda adicionadas, em diferentes experiências, ruído e imprecisão. O ruído consiste na adição de anotações que são aleatórias e não relacionadas com a doença, para representar sintomas que pacientes podem reportar, mas que são apenas coincidências. A imprecisão surge quando são anotados sintomas mais gerais ou imprecisos do que o sintoma original. Foram ainda adicionadas mil doenças aleatórias da Human Phenotype Ontology para dificultar a tarefa de classificação e mimicar condições reais.

Para avaliar o desempenho das medidas que consideram a polaridade das anotações, foram utilizadas as medidas originais BMA e Resnik para comparação, medidas que têm um bom desempenho numa generalidade de aplicações. A metodologia foi a mesma que a utilizada para PolarResnik e PolarBMA mas sem a utilização de anotações negativas para o cálculo da semelhança.

Foi então calculada a semelhança semântica para todas as experiências, utilizando as medidas polares e as medidas originais. Para o caso de previsão de interação proteína-proteína, foi computada a semelhança semântica entre as proteínas de cada par, enquanto que, para a previsão de doenças, foi determinada a semelhança entre os pacientes e as doenças candidatas.

Foi então verificado se as semelhanças obtidas pelas diferentes medidas estavam a permitir prever corretamente se as proteínas interagiam, ou se os pacientes sintéticos sofriam de uma doença específica previamente determinada. Fez-se um ranking de frequência cumulativa para os 10 primeiros lugares, correspondentes às 10 doenças mais prováveis ou as 10 proteínas com maior probabilidade de interagir com uma proteína alvo, para cada medida. Esta medida foi escolhida devido a aplicação da previsão de doenças para ajuda de diagnóstico, onde é preferível ter uma lista das doenças mais prováveis para depois um profissional de saúde poder fazer a decisão final. Também foi calculada a curva de Precision-Recall e ROC AUC para as experiências de interação proteína-proteína, assim como a Average Precision para a previsão de doenças. Estas são medidas de avaliação mais utilizadas tendo em conta as propriedades dos dados.

No caso das proteínas, as medidas polares tiveram um desempenho semelhante às medidas originais, exceto no ramo de Função Molecular e numa experiência de Componente Celular, o que indica que as anotações negativas podem ter uma maior importância na função da proteína e onde na célula a proteína tem a sua função, do que nos processos a que pertence. No caso da previsão de doença, as medidas polares tiveram uma melhoria de 10 por cento no seu desempenho, comparativamente às medidas não

polares. Até nos casos em que não foram adicionadas anotações negativas aos pacientes, as medidas polares tiveram uma melhor performance porque utilizavam as anotações negativas das doenças, o que dá enfase à importância das ontologias terem anotações negativas. É de salientar que o número de anotações negativas, até com a adição de anotações geradas, continuou a ser um número muito reduzido quando comparado com o número de anotações positivas. Assim sendo, se até a introdução de um número reduzido de anotações negativas teve impacto na previsão de doenças, é possível concluir que as anotações negativas têm impacto na semelhança semântica e é um tópico que carece de mais estudo.

Palavras chave: Semelhança Semântica, Ontologia biomédica, Anotação negativa, Previsão Interação Proteína-Proteína, Previsão de doença

Abstract

Classical Semantic Similarity Measures did not consider negative annotations in similarity computation, and the impact that these annotations can have in this data mining technique is not well studied. As such, this work aims to understand how the addition of negative annotations impacts semantic similarity. To do so, two pairwise similarity measures, Best-Match Average and Resnik, were adapted to create the polar measures PolarBMA and PolarResnik. These were evaluated in two currently relevant scopes: protein-protein interaction prediction and disease prediction against the original measures. Pairs of proteins where the proteins were known to interact or not were taken from STRING and enriched with positive and negative annotations from the Gene Ontology. Synthetic patients were created as sets of annotations taken from the Mendelian diseases they were designed to have, as well as possible noise or imprecise annotations. Then semantic similarity was computed with both polar and non-polar measures between proteins in pairs and between patients and candidate diseases including the Mendelian diseases, as well as random diseases taken from the Human Phenotype Ontology.

To evaluate if the polar measures performed well in comparison to the baseline, a ranking according to semantic similarity was made for each measure and scope for evaluation and the rank cumulative frequencies were plotted. ROC AUC and Precision-Recall curves were also determined for the Protein-Protein interaction(PPI) prediction, as well as average precision for the disease prediction dataset. In PPI prediction, polar measures had an increased performance in the Molecular Function branch for both experiments where negative annotations were added and also in one of the experiments with the Cellular Component branch. In the disease prediction scope, polar measures had an improved performance of approximately ten percent. This improvement was verified in all disease prediction experiments, even with the addition of noise and imprecision. Considering the results obtained, this work concludes that negative annotations have an impact on semantic similarity, but the amplitude of this impact requires further study.

Keywords: Semantic Similarity, Biomedical Ontologies, Negative Annotation, Protein-Protein Interaction Prediction, Disease Prediction

Acknowledgments

First and foremost, I must thank Professor Catia Pesquita, for her assistance and dedication in this project. It was not an easy year considering the current state of the world, but with her motivation and availability I managed to do my best despite the circumstances. Her broad knowledge on the topic and excellent problem-solving skills helped me overcome the obstacles that appeared through the project, and her positive but rigorous attitude helped me grow and I will always consider her a role model.

I am also grateful for the LASIGE research unit, most specifically the LiSeDa Lab for providing a community of networking and knowledge sharing. I will remember our weekly meetings and events kindly and I will always appreciate the constructive criticism and support that helped me grow as a scientist.

Finally I thank my family and friends for their support, and for being there for me through the best and worse times. The current pandemic required restraint in regard to social contacts and required work from home, but my friends and family were always available to chat remotely or to give me quiet and work friendly conditions at home when I needed.

Index

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Unlike other scientific fields, in life sciences knowledge is usually expressed in natural language, which is very ambiguous and heterogeneous so it can not be easily reduced to mathematical form, thus becoming difficult to compute (Bodenreider et al., 2006). One way to translate facts from natural language into data that can be computed is by using ontologies. These represent knowledge in conceptual schema, with classes and relationships, that can be compared, enriched with metadata and with specific properties that allow for computation, such as axioms (Hoehndorf et al., 2015). These have gained popularity in the biomedical field in recent years and have become an area of international effort and interest (Hoehndorf et al., 2015).

Ontologies can be used to describe data through ontology-based annotation which associates an entity with a class, combining this assertion with information, including the evidence considered and who created it. Data mining has a wide range of limitations in the biomedical field because classical machine learning algorithms can not interpret the meaning of the information they process. Semantic similarity is a data mining technique that tackles this problem, because semantic similarity measures return a numerical value that reflects the closeness in meaning between entities (Pesquita, Faria, Falcão, et al., 2009). Semantic similarity measures, when used in ontologies for data mining, take advantage of the axioms and definitions to find the similarity between annotated entities. This is extremely useful and has been widely applied to the diagnosis of disease based on phenotype similarity (Hoehndorf et al., 2015).

Most annotations used in biomedical ontologies are positive annotations, these are the standard annotations that state the fact that a biomedical entity is described by an ontology class. For the Human Phenotype Ontology (HPO) (Köhler, Gargano, et al., 2020), for example, diseases have positive annotations that associate the phenotypes with the disease that causes them. In the case of the Gene Ontology (GO) (Ashburner et al., 2000; Consortium et al., 2020), proteins are positively annotated with the molecular functions they are known to carry out, biological processes and cellular components they belong to. However the annotation sets available for these ontologies are incomplete when it comes to negative annotations, i.e., annotations that indicate that a biological entity is not described by an ontology class. This is important because oftentimes the absence of a positive annotation causes confusion because it is not explicit if there is no knowledge about the annotation or if the annotation is indeed negative. Since biomedical knowledge is very incomplete, biomedical ontology annotations reside under the open world assumption (OWA), where the lack of an annotation between an entity and a class cannot be interpreted as evidence for that annotation to be false. This makes it impossible to distinguish between cases where there is true lack of a property versus cases where there is no knowledge of about it. Explicit negative

annotations for when there is such evidence would mitigate this issue.

Usually the computation of semantic similarity uses only positive annotations to determine similarity, e.g., two patients suffer from *tongue atrophy*, so they are similar. However, this method ignores relevant information, since classes can be similar if they share negative annotations, such as two patients that are similar if they don't suffer from *tongue atrophy*, a feature that is treated as missing data by existing semantic similarity algorithms. Negative annotations can help complete the semantic representation of biological entities and thus potentially support more insightful and relevant semantic similarity measures.

Theoretically the number of negative annotations present in an ontology should be significantly higher that the number of positive annotations, so the study of their influence in semantic similarity applications is key to estimate how important negative annotation are for data mining techniques.

1.2 Objectives

This work aimed to investigate if considering negative semantic annotations in biomedical semantic similarity computation influences the performance of semantic similarity applications for protein-protein interaction (PPI) prediction and disease prediction. These are areas of the biomedical field that can benefit greatly from data mining techniques to save resources and help physicians. Previous related work identified the need to use automated techniques for PPI prediction, as well as the limitations in GO caused by the small number of negative semantic annotations present in the ontology (Youngs et al., 2014). As for disease prediction, the utility of data mining tools as well as ontologies has been identified(Masino et al., 2014), but negative annotations' impact on these techniques has not been studied in depth.

1.2.1 Research questions

This project was structured to answer three research questions:

1. How can negative semantic annotations be considered in semantic similarity computation? Taxonomical semantic similarity measures do not consider negative semantic annotations. This work will develop a novel approach to tackle this limitation.

2. How does the addition of negative semantic annotations influence Protein-Protein interaction prediction based on semantic similarity? This work will evaluate if the results obtained from addition of negative semantic similarity improves protein-protein prediction. Evaluation will be made by ranking results according to ground truth and analysing how they relate.

3. How does the addition of negative semantic annotations influence disease prediction based on semantic similarity? This work will evaluate if the results obtained from addition of negative semantic similarity improves the patient's disease prediction. Evaluation will be made by ranking results according to ground truth and analysing how they relate.

1.3 Document Structure

This document is structured as follows:

- Chapter 2 describes the state of the art, that introduces the concepts necessary to understand the methodology, as well as related work.

- Chapter 3 is the about Methods and Data, and focuses on the development of PolarResnik and PolarBMA, and on the creation of the PPI prediction and disease prediction datasets.
- Chapter 4 is about the experimental design, and precises how the negative annotations were generated and further characterizes how the datasets were designed. It also includes the dataset characteristics and evaluation metrics description.
- Chapter 5 presents and discusses the results obtained.
- Finally, Chapter 6 revolves around the conclusions, as well as contributions and future work.

Chapter 2

State of the art

This chapter describes the concepts and techniques that are required to understand the context of this work. It is structured as follows:

- Section 2.1 defines ontologies and specifies their 4 key features, structure and applications. Here the two ontologies used for this dissertation, the Gene Ontology and the Human Phenotype Ontology, are described in more depth.
- Section 2.2 focuses on semantic annotations and their key components on the Gene Ontology, as well as in the introduction of negative semantic annotations.
- In 2.3 semantic similarity is introduced in an ontology context. The concept of Information Content is defined and Node based and edge based approaches are discussed. Pairwise similarity measures are distinguished from group wise semantic similarity measures, and Resnik and BMA are reviewed.
- Finally, 2.4 is dedicated to related work that developed methodologies related to the main problematic addressed on this dissertation, inclusing the scopes of disease and protein-protein interaction prediction, and semantic similarity use with ontologies.

2.1 Ontology

An ontology is a set of logic axioms that form a model of a portion of reality (Bodenreider et al., 2006), and it can be used to represent and share knowledge about a domain by modelling it and the relationships within (Hoehndorf et al., 2015). Functionally, ontologies possess four key features:

1. Classes and Relations: The main components of an ontology, representing entities of the domain and the relationships between them. Can be uniquely identified by IRI's which allow for data integration across platforms.
2. Domain Vocabulary: Labels are associated to classes and relations to represent them beyond their identifiers. Provides terms, such as synonyms, that cover the domain so that natural language processing can be achieved.
3. Metadata and Descriptions: Specifies the meaning of classes and relates these classes to information on other data sources. Guarantees that classes can be understood and described by experts consistently.

4. Axioms and Formal Definitions: allow computation and automation because they can be accessed through software. Ontologies can be represented in OWL or other similar languages and then processed to extract domain knowledge.

Ontologies have important applications, including data integration to Natural Language Processing (Rubin et al., 2007). Biomedical ontologies are historically represented in the Open Biomedical Ontologies (OBO) language, designed specifically for the characteristics of biomedical ontologies and focuses on being readable by humans and machines, as well as easily used (Dessimoz et al., 2017). However, the Web Ontology Language (OWL), a more powerful language in terms of logical statements and reasoning support which is considered the standard in Computer Science, started growing within the Biomedical Ontologies community.

```
<!-- http://purl.obolibrary.org/obo/GO_0002942 -->

    <owl:Class rdf:about="http://purl.obolibrary.org/obo/GO_0002942">
        <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/GO_0030488"/>
        <obo1:IAO_0000115>The process whereby a guanine residue in a transfer RNA is methylated twice at the N2 position.</obo1:IAO_0000115>
        <oboInOwl:created_by>hjd</oboInOwl:created_by>
        <oboInOwl:creation_date>2012-11-16T16:07:20Z</oboInOwl:creation_date>
        <oboInOwl:hasOBONamespace>biological_process</oboInOwl:hasOBONamespace>
        <oboInOwl:id>GO:0002942</oboInOwl:id>
        <rdfs:label>tRNA m2,2-guanine biosynthesis</rdfs:label>
    </owl:Class>
    <owl:Axiom>
        <owl:annotatedSource rdf:resource="http://purl.obolibrary.org/obo/GO_0002942"/>
        <owl:annotatedProperty rdf:resource="http://purl.obolibrary.org/obo/IAO_0000115"/>
        <owl:annotatedTarget>The process whereby a guanine residue in a transfer RNA is methylated twice at the N2 position.</owl:annotatedTarget>
        <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/XMLSchema#string">GOC:hjd</oboInOwl:hasDbXref>
        <oboInOwl:hasDbXref>ISBN:155581073X</oboInOwl:hasDbXref>
    </owl:Axiom>
```

Figure 2.1: Example of a class and axiom from the Gene Ontology written in OWL

Most biomedical ontologies are organized as directed acyclic graphs, with 'child' classes (sometimes referred to as terms), that are specific instances or components of 'parent' or 'ancestor' classes which are more general. It is not a rigorous hierarchy because child classes can have one or more parent classes. This structure resembles a hierarchy in a sense that a child class possesses all the properties of parent classes (Dessimoz et al., 2017).

2.1.1 Gene Ontology

The Gene Ontology is the most adopted ontology by the biomedical field, it contains information about genes and their products. It was created to standardize terms that can then be used to annotate different Model Organism Databases (MOD) in an unambiguous manner (Ashburner et al., 2000; Consortium et al., 2020). Each gene or gene product is annotated with a set of GO-terms, in a process called ontology-based annotation. A GO-annotation describes how and where a gene functions, as well as what processes it contributes to(Rubin et al., 2007).

For this purpose GO is divided into three branches:

1. Biological Process: the process that a gene product's activity contributes to, like 'DNA repair'.
2. Cellular Component: where in a cell is a gene product's activity, an example is 'ribosome'.
3. Molecular Function: molecular activities of a gene product, such as 'catalysis'. These activities can be performed by a gene product or a molecular complex of gene products.

2.1.2 Human Phenotype Ontology

The HPO describes phenotypes and human disease, with the purpose of integrating this knowledge across scientific fields and platforms. HPO includes several subontologies, ranging from 'Phenotypic

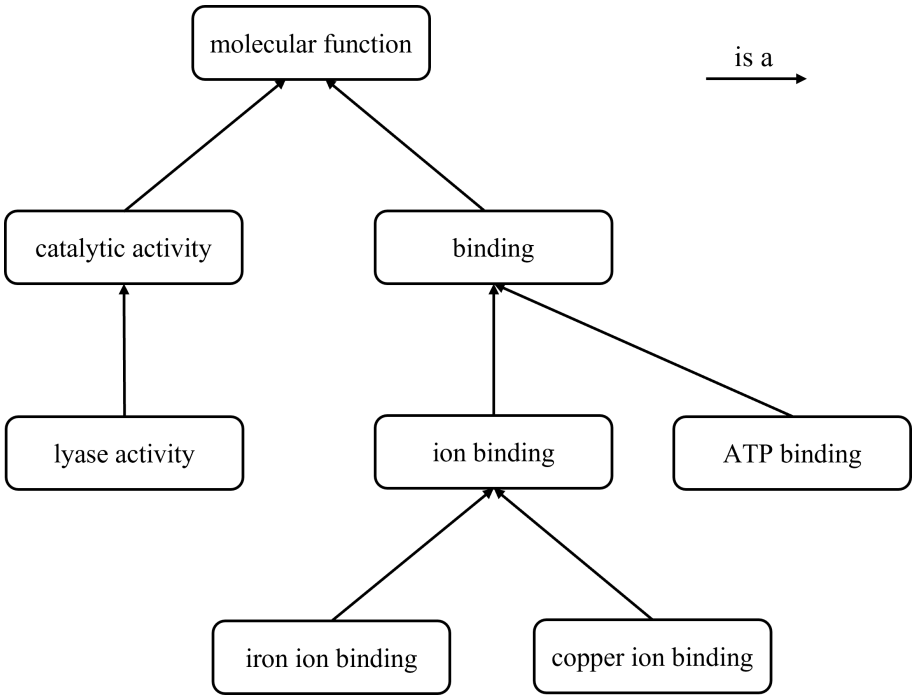


Figure 2.2: Schematic view of an excerpt of the gene ontology

abnormality' to 'Mortality', with the 'Mode of inheritance' being of interest to this work because it describes if a disease has Mendelian inheritance (Köhler, Gargano, et al., 2020).

In this ontology, annotations work as HPO phenotypic profiles, so patients can be described very specifically as a set annotations, which will allow several applications such as personalized medicine and diagnosis. Another point of interest is that for the Mendelian diseases, there is a degree of penetrance for each phenotype depending on the sex of the patient. This is very useful for generating synthetic patients that allow studies of rare diseases that would not have enough patients otherwise. (Pesquita, Faria, Falcao, et al., 2009)

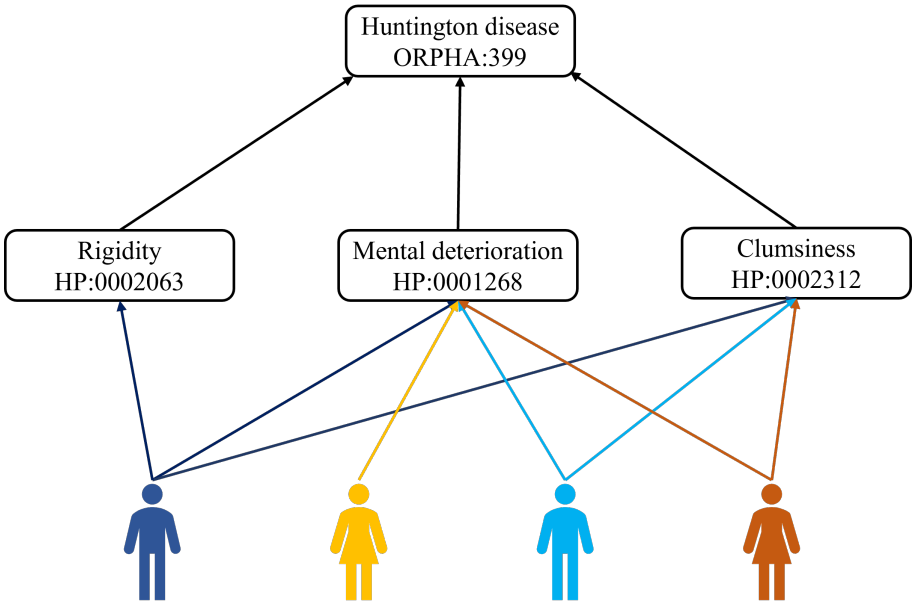


Figure 2.3: Schematic view of a Mendelian disorder (Huntington's disease) and its HPO annotations. Each annotation of the disease has a specific frequency and patients are represented as sets of annotations.

Figure 2.3 shows a representation of Huntington disease, a Mendelian disorder with Autosomal dominant inheritance, with several annotation terms in HPO. These terms, such as Rigidity (with the term identifier HP:0002063), describe the symptoms that patients can present, so patients, shown in different colours in the image, can be seen as sets of annotations. Annotations have different frequencies according to penetrance, and this is evident in Figure 2.3, as Rigidity that is an occasional annotation, is shown in one patient, while Clumsiness is a Frequent annotation, and Mental deterioration, a very frequent symptom of Huntington disease, is present in every patient in the sample.

2.2 Semantic Annotation

One of the main uses of bio-ontologies is for the annotation of data, which captures information about classes in a way that can be accessed by computers and understood by humans (Hill et al., 2008). Ontology-based annotations associate an entity with a class, combining this assertion with information including the evidence considered and who created it (Hoehndorf et al., 2015). Annotation can be performed manually by curators or by automatic processes, the first option generates high quality annotations but cannot keep up with the amount of data high-throughout technologies produce, while the second option is faster but still struggles to achieve the same quality (Couto et al., 2006). Semantic annotations focus on providing meaning so that it can be incorporated into computation.

GO annotations are structured with four key components: Gene Product, GO-term, Reference and Evidence. Additional optional information can be added, such as qualifiers or date. One example of a GO annotation in the format Gene Product | GO-term | Reference | Evidence can be:

- LIG1 | lagging strand elongation | PMID:21873635 | IBA

LIG1 is the Gene Product code for DNA ligase, that in this example is annotated with the GO term lagging strand elongation. IBA is the code for Inferred from Biological aspect of Ancestor evidence.

Annotations can be inferred through:

- Experimental evidence means that experiments directly support the annotation.
- phylogenetic evidence infer relationships among genes by reconstructing evolutionary events or by branch positions in phylogenetic trees.
- Computational analysis stems from sequence or other data analysis.
- Automatically generated data

(Hill et al., 2008)

In the case of positive annotations, the true-path rule applies, whereby an annotation to a given class implies annotation to all of its superclasses. In Figure 2.4, DNA ligase is annotated with the molecular function lagging strand elongation, this annotation is positive so it is known that DNA ligase has this function. Since this annotation is positive it implies annotation to DNA strand elongation and DNA metabolic process, but not to lagging strand initiation, because it is a child class, not a superclass.

Negative semantic annotations are distinguished from positive annotations by the presence of the 'NOT' qualifier, and share the same main structure as positive annotations. An example of a negative annotation in the format Gene Product | Qualifier | GO-term | Reference | Evidence can be:

- CAPN6 | NOT | proteolysis | PMID:21873635 | IBA

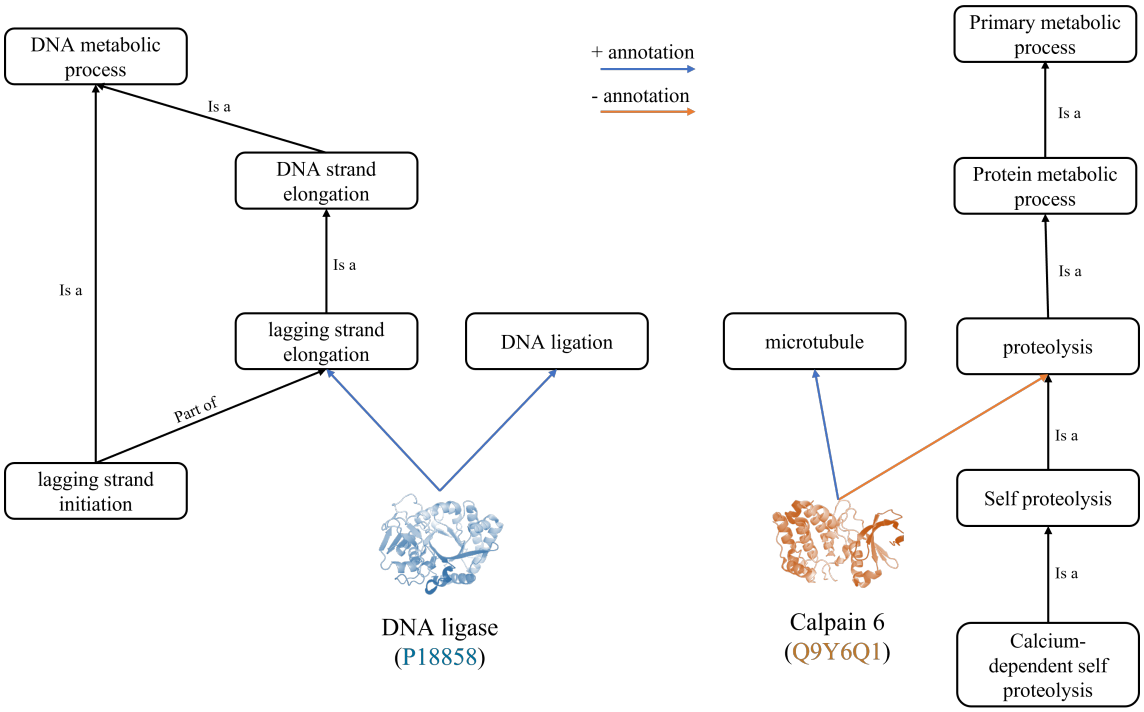


Figure 2.4: Schematic view of positive and negative GO annotations for the DNA ligase and Calpain 6 proteins.

CAPN6 represents Calpain 6, that is annotated with NOT participating in proteolysis. For negative annotations, inheritance is actually reversed, a negative annotation to a given class implies annotations to all of its subclasses. The Calpain 6 negative annotation to proteolysis is represented in Figure 2.4, and since it is a negative annotation it will imply annotation only to Self proteolysis and Calcium dependent self proteolysis. The superclasses are not considered so the process is the opposite of the path shown for positive annotations.

2.3 Semantic Similarity

To mine information from the semantic annotations in an ontology, it is important to measure how similar entities are to each other. To compare a set of annotations it is necessary to compute their similarity value in a manner that conveys meaning. Semantic similarity measures use semantic representations afforded by ontologies to compare the meaning of concepts or entities. This similarity is then represented in numerical form. Figure 2.5 shows two proteins that are similar in the fact that both perform the same function of iron ion binding, but differ due to one having lyase activity while the other performs ATP binding.

The semantic representations allows us to for instance distinguish between concepts with similar names and different meanings. If the meaning of a concept is not taken into account, 'Apoptosis' and 'Apoplexy' are more similar than 'Apoptosis' and 'Programmed cell death', however if meaning is considered and semantic similarity is used, then 'Apoptosis and 'Programmed cell death' are more similar.

Most semantic similarity measures consider only taxonomic relationships when evaluating similarity and explore the concept of the 'true-path ´ rule, whereby an annotation to a given class, implies annotation to all of its superclasses. However, considering negative annotations, requires different procedures when this type of reasoning is applied, since the 'true-path ´ needs to be reversed (Vesztrocy et al., 2020).

Semantic similarity measures fall into two categories: measures for class similarity and measures for

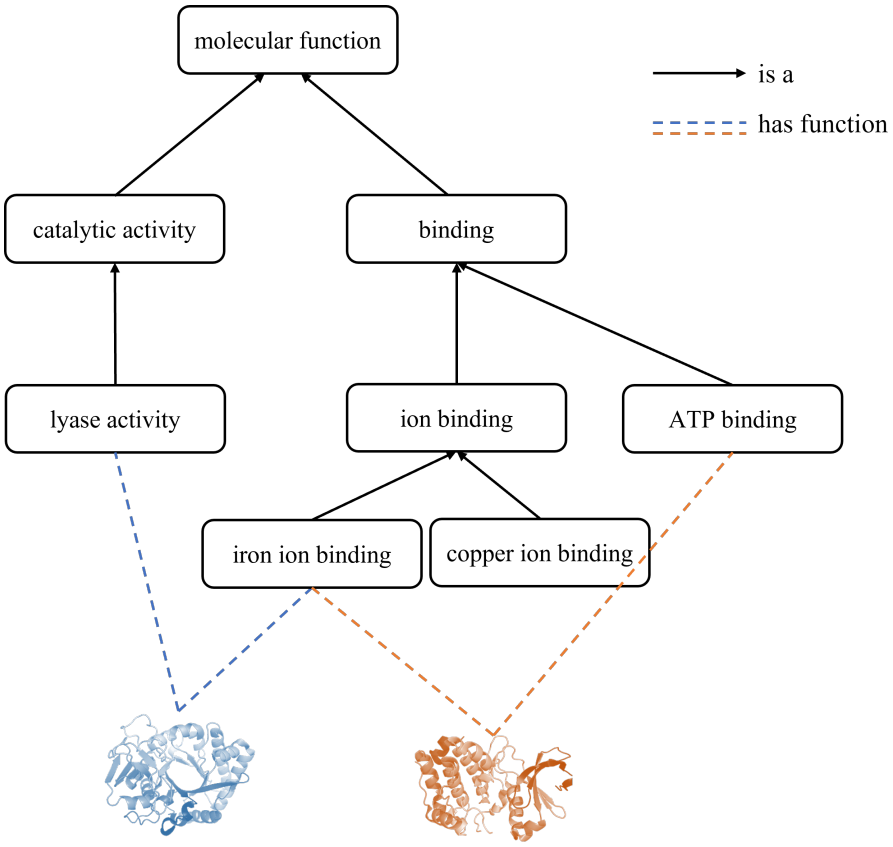


Figure 2.5: Schematic view of an excerpt of the gene ontology with protein annotations.

annotated entity similarity.

2.3.1 Class similarity

Semantic similarity can be employed to measure the similarity between two classes defined in the same ontology. Since ontologies are structured as graphs, ontology classes are defined as nodes and relations as edges. When calculating semantic similarity one may use node-based or edge-based approaches.

Edge-based methods are relation oriented and count the edges in the path that connects the two classes that are being considered (see Figure 2.6, red path) and define it as the distance between entities. Classes that are closer are determined to be more similar than distant classes. Node-based approaches use the properties of the classes, as well as their ancestors (see Figure 2.6) and descendants to compute similarity. (Wu et al., 2013)

Node based methods consider node properties. Several node-based similarity measures exist, with one of the most popular being Resnik's (Resnik, 1995). Given two concepts u and v, Resnik's similarity is calculated by obtaining the common ancestor that maximizes the IC between u and v - Most Informative Common Ancestor (MICA) (Harispe et al., 2015).

$$sim_{Resnik}(u,v) = IC(MICA(u,v))$$

(Harispe et al., 2015)

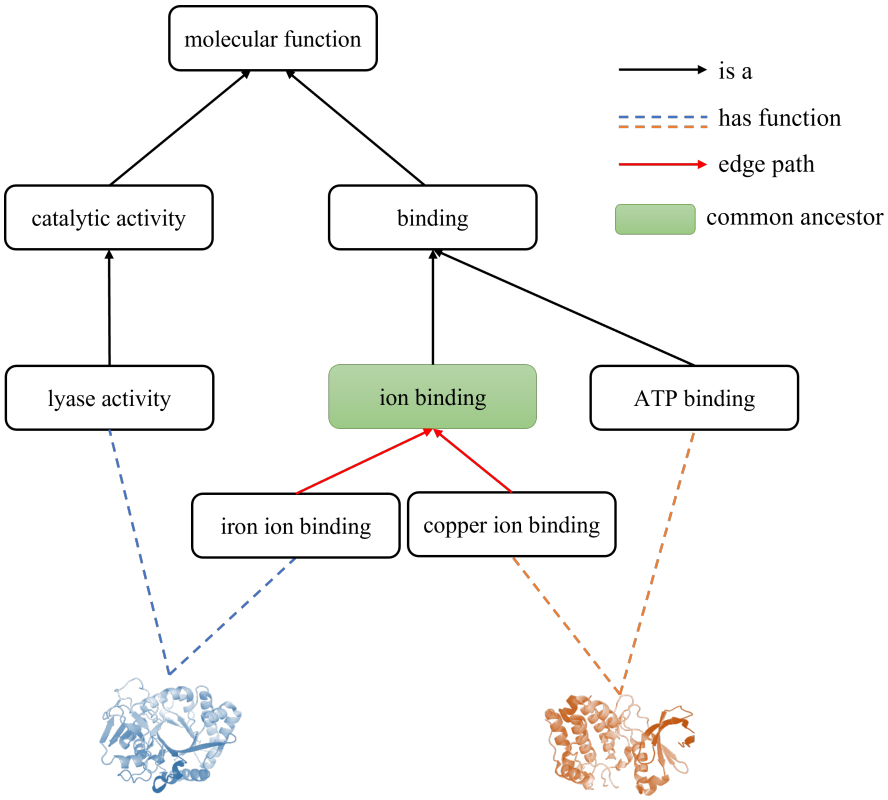


Figure 2.6: Schematic view of an excerpt of the gene ontology describing node-based (green box) and edge-based (red path) approaches.

2.3.1.1 Information Content

To measure similarity between two ontology classes one must first determine how informative the classes and their ancestors or descendants are. According to Information Theory , a concept with a greater specificity conveys more information(Resnik, 1995). Knowing that a protein has a binding function gives less information than specifying a copper ion binding function, for example. Similarity must reflect this principle, so similarity measures can use the Information Content (IC) of classes to determine how much information they share. IC was originally defined in Resnik 1995 (Resnik, 1995) , and is inversely proportional to the probability to encounter an instance c in a set of instances.

The IC of a concept c is then quantified as:

$$IC_{Resnik} = -logP(c)$$

This formula indicates that as the probability of c increases, the more general it is so the IC is smaller. If c is rare, then its probability decreases and IC increases.

Another method to quantify the IC is Seco2004, which assumes that concepts with more hyponyms are less informative since they require further differentiation. Otherwise, concepts that are leaf nodes are very specific and hence more informative (Seco et al., 2004). Formally:

$$IC_{Seco} = 1 - \frac{log(hypo(c)+1)}{log(max_{wn})}$$

(Seco et al., 2004) where hypo returns the number of hyponyms of a concept c, and

$$max_{wn}$$

is the constant associated to the maximum number of concepts in the taxonomy. (Seco et al., 2004)

2.3.2 Entity Similarity

Entity similarity measures fall into two groups: pairwise and groupwise approaches.

2.3.2.1 Pairwise Similarity

Pairwise approaches are based on pairwise comparisons of the two sets of classes that annotate each entity to compare. Popular approaches include taking the average or maximum of all pairwise similarities. The Best-Match Average (BMA) has shown good results. (Pesquita, Faria, Bastos, et al., 2008)

To compute semantic similarity using BMA, given 2 classes c1 and c2 and their respective set of annotations annot(c1) and annot(c2), the average similarity of each term t1 in annot(c1) and its most similar term t2 in annot(c2) are calculated and then averaged with their reciprocal as is shown in 2.1

$$\text{Sim}_{\text{BMA}}(c_1, c_2) = \frac{\sum_{t_1 \in \text{Annots}(c_1)} \text{sim}(t_1, t_2)}{2|\text{Annots}(c_1)|} + \frac{\sum_{t_2 \in \text{Annots}(c_2)} \text{sim}(t_1, t_2)}{2|\text{Annots}(c_2)|} \tag{2.1}$$

2.3.3 Groupwise Similarity

Groupwise similarity measures compare more than 2 concepts simultaneously. Three approaches can be used: set, graph and vector. In the set-based approach, direct annotations are organized into term sets to represent an entity, then set similarity techniques are used (Pesquita, Faria, Falcão, et al., 2009; Teng et al., 2013). Approaches can be further described as direct or indirect, depending on whether the sets are compared to their information characterised in taxonomy or if pairwise measures are used as an intermediate step (Harispe et al., 2015).

SimGIC is a straightforward and well performing method of this type and is defined in equation 2.2:

$$\text{SimGIC}(e_1, e_2) = \frac{\sum_{c \in \{\text{Annots}(e_1) \cap \text{S}(e_2)\}} \text{IC}(c)}{\sum_{c \in \{\text{Annots}(e_1) \cup \text{S}(e_2)\}} \text{IC}(c)} \tag{2.2}$$

2.4 Related Work

2.4.1 Protein-Protein Interaction Prediction

Proteins have a ubiquitous influence in biological processes of a living cell and in the cellular systems of all organisms.To understand how proteins work, it is essential to study how they interact with each other, since proteins have very complex pathways. The prediction of protein-protein interactions is shedding light into how these networks work to regulate and allow the functioning of an organism.

The process of validating which proteins interact and how is very resource-intensive when done in wet-lab experiments (Mohamed et al., 2010), as well as sensitive and time consuming since few complexes can be studied at a time (Gonzalez et al., 2012). It has been estimated that 300.000 PPI exist in humans (Hart et al., 2006) and the number of experimentally proven interactions is much smaller (McDowall et al., 2008).

For these reasons, and with the rapid growth of computation in the life sciences, it became clear that PPI prediction with automated methods would be a cost-effective and fast alternative to lab experimentation. Computational methods are usually based on structure, sequence and expression data (Gonzalez. Kann 2012). Databases, such as STRING, were created to store and predict PPI based on annotations and experimental information (McDowall et al., 2008). Machine Learning (ML) methods have gained popularity since ML algorithms have been applied for PPI prediction using the existing labelled data from lab experiments to train, as well as indirect information from GO and sequence data (Mohamed et al., 2010).

The Critical Assessment of protein Function Annotation algorithms is a global effort to improve computational annotation of protein function. This initiative has improved protein function prediction , however the CC GO-branch annotations remain illusive (Zhou et al., 2019).

2.4.1.1 Protein-Protein Interaction Prediction in Biomedical Ontologies

The use of the Gene Ontology has become a powerful tool to predict PPI, especially when coupled with semantic similarity. Jain and Bader, 2010, developed the Topological Clustering Semantic Similarity algorithm, that computes similarity between GO-terms to accurately distinguish true from false protein interactions (Jain et al., 2010). Wu et al., 2013 attempted to tackle that same limitation, with an Edge-based and IC-based hybrid methodology. Both methods improved PPI prediction and solidified the role of semantic similarity and GO in PPI prediction.

2.4.2 Negative Annotations in Biomedical Ontologies

Negative examples, which in the PPI scope indicate that a protein does not have a certain function for example, are essential for ML algorithms, however ontologies and databases are incomplete and focus mostly on positive examples. Youngs et al., 2014 designed two algorithms to predict negative examples, which then populated the general use NoGO database. Fu et al., 2016 then developed a novel approach to select negative examples of a protein in a protein-protein interaction context, taking advantage of the hierarchical semantic similarity between GO terms and GO hierarchy. Both of these methodologies tackled the problem of the lack of negative annotations in GO, since even though ML algorithms need negative examples of proteins to predict PPI, GO is very incomplete in negative annotation. The number of functions that a protein has are logically smaller than the number of functions it does not have, so theoretically the number of negative annotations in GO should be several degrees of magnitude larger than the number of positive annotations. Still, negative annotations are somewhat rare and difficult to confirm in lab conditions, so techniques like NegGOA can tackle that limitation.

Vesztrocy et al., 2020 targeted the limitation that OWA must be taken into consideration to avoid underestimation of precision due to lack of negative annotaions, especially in Critical Assessment of protein Function Annotation (CAFA) challenges. To tackle this issue, a benchmark was created to include a balanced set of curated positive and negative annotations.

2.4.3 Disease Prediction

Human disease is a very complex topic, not only in identifying but also choosing the best treatment for a specific ailment. To help physicians, ontologies like the HPO have been created and machine learning algorithms trained to predict disease presence or characteristics (Dahiwade et al., 2019), since oftentimes the symptoms can be to cryptic for a doctor to find the cause early.

Heart disease is one of the leading causes of death today, and to diagnose it early and efficiently is challenging due to the complex structure of risk factors and symptoms (Mohan et al., 2019). Various machine learning methodologies have been designed for a better prediction of heart disease and its extent (Mohan et al., 2019; Palaniappan et al., 2008; Kohli et al., 2018).

Another ailment that is significant in mortality statistics and needs early detection and urgent treatment is cancer. Often it is difficult to correctly determine the malignity of clump cells early, and ML algorithms have been trained to do this task according to parameters about the cell characteristics (Kohli et al., 2018).

A different application for machine learning algorithms in disease prediction is to determine the progression rate. Amyotrophic Lateral Sclerosis, for example, has a progression rate that varies greatly among patients, so studies have been conducted to try to find a pattern, through machine learning algorithms, that can predict how fast the disease progresses for a more personalised medical treatment (Teixeira, 2019).

As medicine evolves digitally and more knowledge is found, ontologies and data mining techniques are becoming more valuable tools, so it is of interest to know if the addition of negative annotations can improve disease prediction

2.4.4 Semantic Similarity Measures in the Human Phenotype Ontology

Köhler, Schulz, et al., 2009 developed the Phenomizer, a tool to help experts in the differential diagnostic process. This was evaluated using synthetic patients, with HPO terms, and noise and imprecision conditions to simulate clinic conditions and give a significance score to candidate diseases. Masino et al., 2014 had a similar methodology, and used HPO to create synthetic patients as sets of HPO annotations of a given Mendelian disease and created an algorithm that ranked genes from HPO according to semantic similarity. This was then evaluated by determining the rank of the causal gene of the Mendelian disease associated with a given synthetic patient. An interesting characteristic of these methodologies is that noise and imprecision were added to the patient annotations to mimic real world conditions. This is important because real patients can be ambiguous in their description of phenotype, or even add unrelated symptoms to their description, and clinical reports can be incomplete or imprecise. The latter study found that semantic similarity is effective in ranking the causative gene of a disease as long as imprecision is mitigated.

Chapter 3

Methods and Data

This chapter focuses on the development of PolarResnik and PolarBMA, semantic similarity measures that take into consideration negative annotations, and the data used for their evaluation, as well as how the semantic similarity was calculated. It is organized in the following manner:

- 3.1 is an overview of this work's methodology.
- 3.2 describes the challenge of adapting Resnik1995 to include annotation polarity, as well as how the adapted measure PolarResnik works.
- The section 3.3 is dedicated to PolarBMA, and has a similar structure to section 3.2, specific to BMA's adaptation to consider negative semantic annotation. This section also includes the pseudo-code of PolarBMA.
- 3.4 describes protein-protein interaction data, its manipulation and inclusion of generated negative annotations. This section also covers how the semantic similarity between proteins in a pair was calculated.
- 3.5 introduces the disease prediction dataset, how synthetic patients were created, and negative annotation generation, as well as semantic similarity calculation between patients and diseases.

3.1 Overview

The goal of this work was to determine if the incorporation of negative semantic annotations can improve semantic similarity applications. To do so, a semantic similarity measure that considers negative annotations into the computation of semantic similarity had to be designed, as well as datasets that included a suitable number of negative and positive semantic annotations. The classical measures BMA and Resnik perform very well in varied applications so they were adapted to include negative annotations as a novel semantic similarity measure, composed of PolarBMA which introduces negative semantic annotations into BMA, coupled with PolarResnik, that takes into consideration the polarity of annotations. To test whether these measures were effective, two scenarios were designed, using BMA with Resnik1995 as comparison. The first scenario was to predict protein-protein interaction, and the second to predict the disease of synthetic patients.

3. METHODS AND DATA

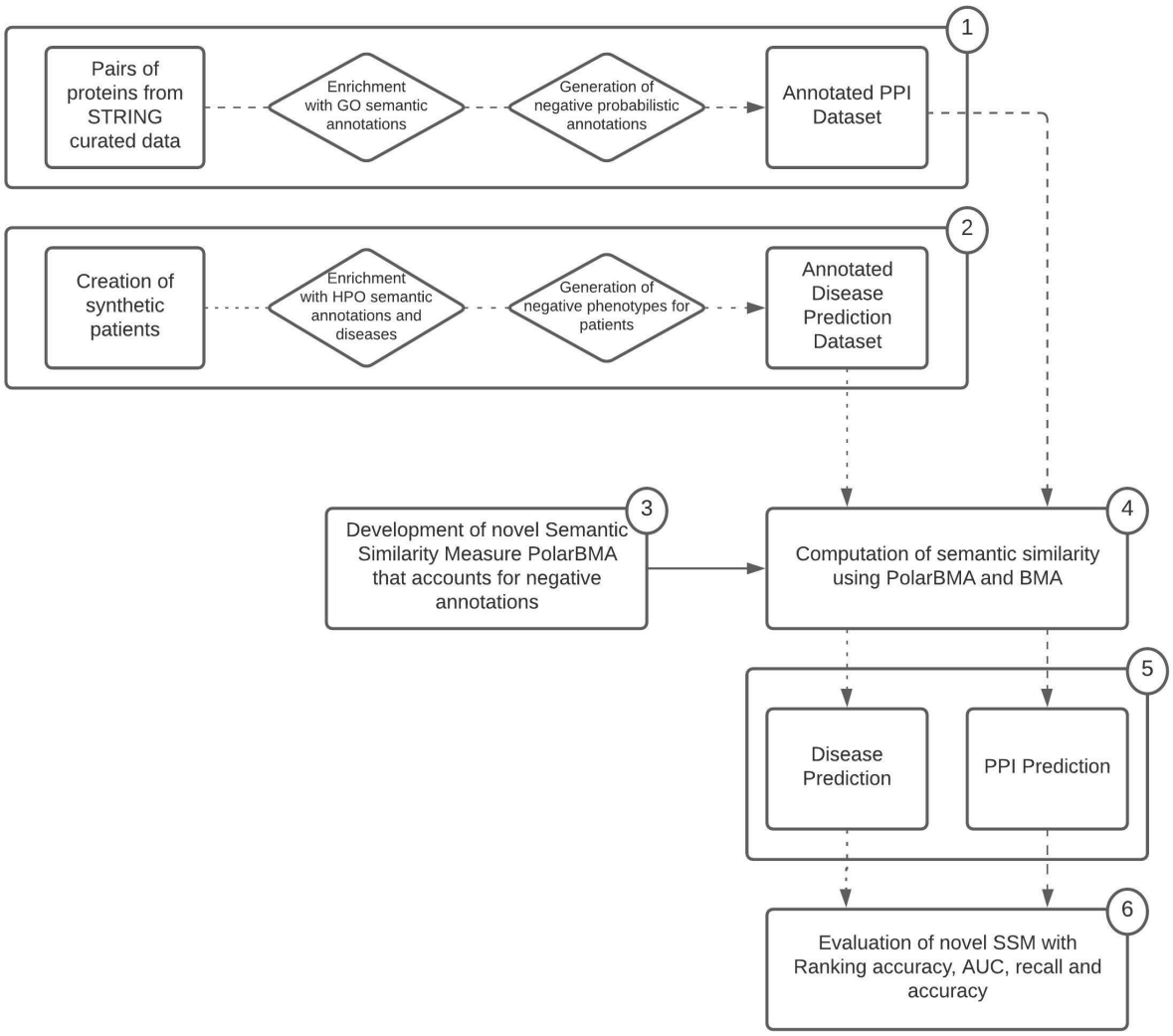


Figure 3.1: Overview of the 6-step methodology: 1 - Creation of a PPI annotated dataset; 2 - Creation of an annotated Disease Prediction Dataset; 3 - Development of a novel semantic similarity measure; 4- Semantic Similarity computation; 5- Prediction; 6 - Evaluation

3.2 PolarResnik

PolarResnik is a novel class similarity measure that takes into consideration the polarity of the annotations made with each class being compared. The biggest challenge to create this measure was to determine how to calculate shared IC between annotations of different polarity. For negative annotations it does not make sense to use the most informative common ancestor, since negative annotations have an inverse inheritance in comparison to positive annotations. For instance, the annotation 'upper limb pain' inherits relations from the ancestors, while 'NOT upper limb pain' inherits them from the descendants. Upper limb pain is a limb pain, but if a patient doesn't have upper limb pain he might still have limb pain, but she won't have hand pain, or finger pain.

For this reason, to calculate similarity using PolarResnik, descendants are considered for negative annotations, while ancestors are used to positive annotations. Considering classes c1 and c2, there are four possible scenarios:

1. Both classes are positive annotations. Then the similarity is calculated as in usual Resnik1995:

$$sim_{Resnik}(c1,c2) = IC(MICA(c1,c2))$$

2. Both classes are negative annotations. Then the similarity will be calculated using the common descendant with the highest IC - Most Informative Common Descendant (MICD):

$$sim_{PolarResnik}(c1,c2) = IC(MICD(c1,c2))$$

3. c1 is a positive annotation and c2 is a negative annotation. Then the ancestors of c1 and descendants of c2 are gathered and similarity is calculated using the common class with the highest IC - Most Informative Common Class (MICC):

$$sim_{PolarResnik}(c1,c2) = IC(MICC(c1,c2))$$

4. c1 is a negative annotation and c2 is a positive annotation. Then the descendants of c1 and ancestors of c2 are gathered and similarity is calculated using the common class with the highest IC -Most Informative Common Class:

$$sim_{PolarResnik}(c1,c2) = IC(MICC(c1,c2))$$

3.3 PolarBMA

PolarBMA is an entity similarity measure of the pairwise type that takes into consideration the polarity of the annotations made to the two entites it compares. To create PolarBMA, the biggest challenge was to determine how to integrate negative semantic similarity without introducing bias, while obtaining satisfactory results. To do this, one must consider how polarity influences similarity. A given pair of terms will be more similar if they share polarity. For example, if two patients are annotated with 'upper limb pain', similarity will be 1, while if one patient has 'upper limb pain' and another patient has 'NOT upper limb pain', they will be opposites.

To overcome this challenge, for each pair of entities (A,B), positive annotations from A were paired up with positive annotations from B, to create a list of all possible pairs of positive annotations (PtoP).

Negative annotations from A were paired with negative annotations from B to create a list of all pairs of negative annotations (NtoN). Finally, positive annotations from A and negative annotations from B were paired, as well as negative annotations from A and positive annotations from B, to create two sets of all possible mixed polarity pairs, depending on the order of the negative/positive annotation in the pair (PtoN and NtoP).

Pairwise similarity was calculated using PolarResnik, depending on the polarity of the pair. For each entity pair, four matrices were computed, one for each polarity - PtoP, NtoN, PtoN and NtoP - containing the annotations' pairwise similarities. Then the best match for each annotation was found and if the polarity of the pair with the best match was mixed, then the score was inverted. This was done to reflect the opposition of polarity in the similarity score. At last, the best matches average was computed and averaged with its reciprocal.

The pseudo-code for PolarBMA is as follows:

Algorithm 1 PolarBMA pseudo-code

```
Pairs ← All combinations between annotations of entity A and entity B, including their polarity
simMatrix ← To store all similarities between annotations in A x B
bestMatches
for pair(c_A, c_B) in Pairs do
    sim(c_A, c_B) ← PolarResnik(c_A, c_B)
    simMatrix.add(sim(c_a, c_B), polarity(c_A, c_B))
for c_A in simMatrix do
    bestMatch ← pair(c_A, c_B)withmax(sim)
    maxScore ← sim(bestMatch)
    if pol == "equal" then
        bestMatches.add(maxScore)
    else
        bestMatches.add(−maxScore)
for c_B in simMatrix do
    bestMatch ← pair(c_B, c_A)withmax(sim)
    maxScore ← sim(bestMatch)
    if pol == "equal" then
        bestMatches.add(maxScore)
    else
        bestMatches.add(−maxScore)
interimScore = (sum(bestMatches)/len(bestMatches))
scores.add(interimScore)
score = sum(scores)/2
```

3.4 Protein-Protein Interaction

3.4.1 Data

To predict protein-protein interaction using semantic similarity, the first step was to make a repository of proteins. Pairs of proteins were taken from a curated STRING dataset, and proteins from pairs with a score higher than 950 out of 1000 were selected. The proteins were then enriched with their annotations out-of the Gene Ontology so that proteins can then be represented as sets of semantic annotations.

Semantic annotations in GO have the qualifier 'NOT' if the annotation is negative, meaning that a

gene product has been experimentally proven not to carry out a specific activity, or lost that function (i.e., it would be expected for the gene to carry out that function given its phylogeny, but function was lost). Hence, the annotations that possess this qualifier were classified as negative while the annotations without it were considered as positive. Proteins that did not have at least one positive annotation were discarded from the data. Two annotation files for the same set of proteins were created, one with both positive and negative annotations and another only with positive annotations. The first file was used to compute similarity with polar semantic similarity measures(SSM), so it contained the proteins and their semantic annotations with an added '+' or '-', according to the presence or absence of the NOT qualifier. The second annotation file was required to calculate similarity with traditional SSM, so it contained the same proteins, however negative semantic annotations or any kind of polarity were not considered.

3.4.2 GO-branches

GO annotations can be classified into 3 branches: Molecular Function (MF), Cellular Component (CC), and Biological Process (BP). It can be of interest to study the branches separately so negative annotations for each protein were separated according to their GO branch, then proteins were placed in groups according to the branch of their negative annotations. Since proteins can have negative annotations of different branches, it was possible for a protein to belong to more than one branch group.

The previously selected pairs from STRING were considered as pairs of proteins that interact, because of the high similarity score, so the dataset was scanned to find pairs where both proteins were of the same group of annotation branch. These pairs were defined as the interacting pairs, then 10000 pairs of proteins of the same branch group were randomly generated and the pairs that did not have a score higher than 950 or were not present in the STRING dataset were defined as pairs that do not interact.

Pairs of proteins that interact and pairs that do not were joined, according to the branches to which their negative annotations belonged, and three datasets were created MF, CC and BP. So a pair was assigned to the dataset MF if both proteins had at least one negative annotation that belonged to the MF GO branch. Then a forth dataset was generated as the union of pairs from the first three datasets: ALL. This is important to evaluate the impact of negative annotations in specific branches, but also in the whole picture.

3.4.3 Semantic Similarity

To calculate the semantic similarity of pairs from the four datasets without considering negative annotations, the annotation files without negative annotations or polarity indication were used. First, a dictionary with the protein pairs and their IC values was computed using Seco 2004 or Resnik1995, then the pairwise similarity between pairs of annotations was calculated with Resnik1995 and BMA.

The same pairs were then used to calculate polar semantic similarity, using annotation files with negative annotations and indication of polarity. To do so, the IC values were computed as previously, however, PolarResnik was used for class similarity calculation and PolarBMA was chosen to compute entity similarity. The obtained score was then used for PPI prediction. Higher similarity scores indicated a higher probability that a pair interacts, so for each protein, pairs were ranked by semantic similarity. For simplicity, and because in both baseline and polar measures, both BMA and Resnik are used in combination with each other, the Resnik and BMA experiments will be called BMA, and the PolarResnik and PolarBMA combination will be called PolBMA.

3.5 Disease Prediction

3.5.1 Data

To predict which disease a synthetic patient was labelled with, the methodology from Masino et al., 2014 was adapted. 33 Mendelian diseases were taken from related work (Masino et al., 2014), and a file was made with phenotypes from Human Phenotype Ontology for each disease. It is possible for an illness to have an absent phenotype, which is a negative annotation from HPO that includes 'NOT', that indicates that a disease does not cause that phenotype. On the positive annotations of the disease, which are the phenotypes without the 'NOT', the penetrance of the phenotype is also present. Penetrance indicates the likelihood that a patient that suffers from a given disease presents a specific symptom. 50 synthetic patients were then generated for each disease. To create the patients, the presence of a phenotype on a patient depended on the disease's penetrance and the gender of the patient, defined randomly with an equal probability to both genders. Negative phenotypes do not have penetrance so to decide if a patient would be assigned a negative annotation, a probability N was defined (N=[0,0.25,0.5]). Three datasets were created with patients as sets of positive and negative annotations, according to N. For example, 'Aarskog-Scott syndrome' is annotated with 'Ptosis', which has a penetrance of 0.5061, so approximately half of the synthetic patients, generated for this disease will present this symptom. It is also annotated with 'NOT Decreased fertility', so that disease is know to not decrease fertility. Depending on the dataset, a patient generated with that illness will have a probability N to have that negative annotation.

1000 diseases were taken randomly from HPO to mix with the original diseases and add complexity to the disease prediction. These were enriched with their positive and negative annotations and annotation files were made, according to whether it would be used for polar semantic similarity or not-polar traditional semantic similarity computation. Each file contained the 33 Mendelian diseases and their annotations, 1000 random diseases and their annotations, and finally 1650 generated patients and their generated annotations. For the calculation of polar semantic similarity, the annotations had '+' or '-' to indicate their polarity, while for the non-polar similarity, the negative annotations were removed and no indication of polarity was included. Disease annotations include all available annotations for the disease, both positive and negative.

3.5.2 Semantic Similarity

Semantic similarity was calculated between each of the 1650 patient and the 1033 possible diseases. The calculation followed the same steps as the protein-protein interaction semantic similarity.

Chapter 4

Experimental Design

To evaluate the polar similarity measures, the chosen baseline was computed using Resnik1995 or Seco2004 for IC calculation, then Resnik1995 and BMA for pairwise semantic similarity. The measures were evaluated in two different scenarios that are significant to the biomedical field: Protein-protein interaction prediction and disease prediction. This chapter is divided into the following sections:

- 4.1 describes the different datasets and how they were created for the Protein-Protein Interaction data. It specifies how negative annotations were generated and the final datasets characteristics.
- The next section, 4.2, evaluates the developed measures in the disease prediction scope. It focuses on the parameters that add real world conditions to the dataset: the negative annotation generation and the noise and imprecision addition.
- 4.3 is dedicated to which evaluation metrics were chosen and why.

4.1 Protein-Protein Interaction

4.1.1 Negative Semantic Annotations Generation

Since negative annotations are rare in GO, to investigate the impact of a higher proportion of negative annotations, more negative annotations were automatically generated using the probabilistic method proposed by NegGOA (Fu et al., 2016).

To find probable negative annotations, all proteins and annotations from GO were gathered, and a dictionary was made containing annotations as keys, with the value of other annotations that were present in the same proteins. For clarity, consider three proteins with 2 annotations each:

- P1: A1, A2
- P2: A2, A3
- P3: A1, A2

A1 appears in the same protein as A2 in P1 and P3, A2 shares P1 and P3 with A1 and P2 with A3, and finally A3 shares a protein with A2. This information is saved in a dictionary:

- A1:[A2,A2]
- A2:[A1,A3,A1].

- A3:A2

Then, for every possible pair protein-annotation, the keys of the dictionary contained in the set of annotations of the protein, excluding the annotation in the pair, were retrieved and the frequency of the annotation was calculated for each entry. Afterwards, the total number of proteins in which the annotation appears was counted. Finally, the probability that the annotation was not negative is equal to the average of the frequencies divided by the total. For example, for the pair P1-A1, the frequency of A1 in A2's dictionary entry is 2/3 and in A3's entry is 0. A1 appears in 2 proteins, so the probability that A1 is not a negative annotation of P1 is the average of 0.33 and 0: 0.17.

An annotation A was considered to be a negative annotation for a protein P if the probability calculated in the previous step for the pair P-A was lower than a given threshold.

Negative annotations can be generated with different thresholds, representing the probability that an annotation will not be negative, and different thresholds will produce datasets with higher quality annotations or larger less specific datasets. The chosen thresholds were 0.05 and 0.1, so two annotation files were created per threshold, one with polarity and one without. The process of protein pair selection was repeated for each threshold and 4 datasets were created according to threshold: ALL0.1 and ALL0.05. Then 6 more files were created to evaluate both threshold and GO branch: MF 0.1, CC 0.1, BP 0.1, MF 0.05, CC 0.05 and BP 0.05. So a pair would be added to CC 0.1, if both proteins had a negative annotation for the CC GO branch after the generation of negative annotations with a 0.1 threshold.

4.1.2 Final Dataset Characteristics

In Table 4.1, the mean and median number of annotations per protein is shown for each dataset. For the datasets that do not contain generated negative annotations the mean and median number of annotations per protein is similar between different GO-branches. However, when generated annotations are taken into consideration, in dataset 0.05, the GO-branches CC and MF has significantly less annotations than the other branches while the BP GO-branch has more annotations. This difference stems from the number of positive annotations in each protein that appears to get lower in CC and MF when negative annotations are generated. It seems counter intuitive that as negative annotations are generated, the mean number of positive annotations decreases. This happens because, since only proteins with at least one negative annotation are included, when these are generated the number of eligible proteins increases. These new proteins had a lower mean number of annotations so the overall mean decreases.

Table 4.1: Number of annotations in the PPI Dataset

	Annotations	Positive Annotations	Negative Annotations
All mean	18.54	17.32	1.22
All mean	18.54	17.32	1.22
CC mean	20.79	19.35	1.44
CC median	15	14	1
MF mean	19.6	18.19	1.43
MF median	14	13	1
BP mean	20.40	19.01	1.38
BP median	15	13	1

Table 4.2: Number of annotations in the PPI Dataset 0.05

	Annotations	Positive Annotations	Negative Annotations
All mean	17.98	16.68	1.30
All median	13	12	1
CC mean	13.57	12.40	1.17
CC median	9.5	8.5	1
MF mean	15.46	14.15	1.31
MF median	10	9	1
BP mean	20.38	19.08	1.30
BP median	15	14	1

Table 4.3: Number of annotations in the PPI Dataset 0.1

	Annotations	Positive Annotations	Negative Annotations
All mean	18.08	16.78	1.32
All median	13	12	1
CC mean	16.74	18.43	1.31
CC median	10	8.5	1
MF mean	17.52	16.15	1.37
MF median	12	10	1
BP mean	20.80	19.48	1.32
BP median	15	14	1

4.2 Disease Prediction

To evaluate PolarResnik and PolarBMA on whether they can be generalized with satisfactory results, the synthetic patients data was used with different parameters.

4.2.1 Negative Annotation Generation

First, the probability of adding a negative annotation to the patient was defined. This was tested in three thresholds, 0, 0.25 and 0.5 because, even though negative phenotypes do not have penetrance so they should be present with a probability of 1, an issue with negative annotations is their confusion with missing annotations. When a physician fills a patient file, for example, the patient is usually annotated with his symptoms so negative annotations are not added, however it cannot be assumed that a missing annotation is a negative annotation. When questionnaires are used, it is more likely that negative annotations will be generated since the options include 'NO' when a patient does not present a symptom. With all that in consideration 3 thresholds were chosen to try to mimic real life conditions.

4.2.2 Noise and imprecision

To add more realism to the synthetic patients test, noise and imprecision were added. Noise attempted to mimic symptoms that are not related to the patient's pathology but coincidentally appear at the same time. To add it to the annotations, the methodology from Masino et al., 2014 was followed and a number

(half of the total annotations of a given patient) of random annotations were added to each patient. Six more annotation files were created following the same process described in the methodology, but containing noise annotations.

Imprecision tackled the lack of detail provided in symptom description, like cases where a phenotype would be described as a less specific symptom, for example a patient that would have 'foot pain' instead of 'ankle pain' in their file. The methodology to add imprecision to the annotations was changed from the original paper, as half of each patient's annotations were replaced by an annotation from a parent node, in case of a positive annotation, or from a child node, if the annotation was negative. The methodology was changed because originally all annotations were replaced by any superclass, which is too severe of a change to be realistic. Then six more annotation files were created in the same manner as the six original files, but including imprecision annotations.

Finally six files were created combining imprecision with noise, to simulate real life conditions. The noise annotations were added to the imprecision files, as per Masino et al., 2014 methodology. In total 24 datasets were created to test polar measures, considering the value of N and if the conditions were optimum (without noise nor imprecision), with noise, imprecision, or both.

4.2.3 Final dataset characteristics

In Tables 4.4 and 4.5, the mean and median number of annotations for each disease is shown. Mendelian diseases have a greater number of positive annotations than the randomly chosen ailments from HPO.

As for Tables 4.6 through 4.9, they contain information about the number of annotations of patients for each dataset depending on the value of the negative threshold and the presence of noise or/and imprecision. Within each table the number of annotations is consistent with the number of negative annotations increasing with the Negative threshold as expected. In 4.7 and 4.9, the number of negative annotations with the null negative threshold is not zero, which is unexpected, but easily explained by the fact that the noise introduced was chosen from a pool that contained both positive and negative noise annotations.

Table 4.4: Number of annotations in the Mendelian diseases

	Annotations	Positive Annotations	Negative Annotations
Mean	49.03	48.9	0.13
Median	47.5	47	0

Table 4.5: Number of annotations in the other diseases

	Annotations	Positive Annotations	Negative Annotations
Mean	18.68	18.57	0.1
Median	14	14	0

Table 4.6: Number of annotations in the synthetic patients without noise or imprecision.
NAP:negative annotations probability

NAP	Measure	Annotations	Positive Annotations	Negative Annotations
0	Mean	7.83	7.83	0
	Median	7	7	0
0.25	Mean	7.89	7.84	0.03
	Median	7	7	0
0.5	Mean	7.94	7.88	0.06
	Median	7	7	0

Table 4.7: Number of annotations in the synthetic patients with noise.
NAP:negative annotations probability

NAP	Measure	Annotations	Positive Annotations	Negative Annotations
0	Mean	13.04	13.01	0.03
	Median	11	11	0
0.25	Mean	13.11	13.04	0.07
	Median	11	11	0
0.5	Mean	13.18	13.08	0.09
	Median	11	11	0

Table 4.8: Number of annotations in the synthetic patients with imprecision.
NAP:negative annotations probability

NAP	Measure	Annotations	Positive Annotations	Negative Annotations
0	Mean	7.60	7.60	0
	Median	7	7	0
0.25	Mean	7.66	7.62	0.03
	Median	7	7	0
0.5	Mean	7.72	7.66	0.06
	Median	7	7	0

Table 4.9: Number of annotations in the synthetic patients with noise and imprecision.
NAP:negative annotations probability

NAP	Measure	Annotations	Positive Annotations	Negative Annotations
0	Mean	16.44	16.41	0.03
	Median	14	14	0
0.25	Mean	14.66	14.58	0.08
	Median	13	12	0
0.5	Mean	14.92	14.82	0.10
	Median	13	13	0

4.3 Evaluation metrics

To evaluate how the developed measures performed against the baseline in PPI, different evaluation metrics were used. The most commonly used metrics are precision and recall. Recall is the proportion of true positives cases that are correctly classified (Hossin et al., 2015), and is defined by the following equation:

$$Recall = \frac{tp}{tp+tn}$$

Where tp stands for true positives and tn for true negatives, a true positive happens when a true positive class is correctly classified as positive, and true negatives are true negative classes that are correctly predicted as negative. Precision measures the positive cases that are predicted correctly from the total predicted cases of the positive class(Hossin et al., 2015), and is defined by:

$$Precision = \frac{tp}{tp+fp}$$

Fp is the false positive cases, where a true positive class is wrongly assumed to be negative. F-Measure is the harmonic mean between precision and recall(Hossin et al., 2015) and is represented by:

$$F-Measure = \frac{2*Precision*Recall}{Precision+Recall}$$

Since the results from the Semantic Similarity computation are continuous and a threshold to separate positive cases from negative cases is necessary, a precision-recall curve was chosen. The PR curve is a good choice for binary classification with continuous values because it displays performance for a range of thresholds(Boyd et al., 2013).

Receiver Operator Characteristic (ROC) curves show how the True positive rate varies with the false positive rate. Since there are only two classes and the positive class is not imbalanced with the negative class(Davis et al., 2006) in this work, then this evaluation metric is adequate. The True Positive Rate and False Positive Rate are calculated as follows:

$$TPR = \frac{tp}{totalpositives}$$

$$FPR = \frac{fp}{totalnegatives}$$

(Davis et al., 2006)

After drawing the curve, it is useful to calculate the Area Under the Curve (AUC), which is a value closely related to the quality of the classification(Cortes et al., 2003). It is determined by the following formula:

$$AUC = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{x_i>y_j}}{mn}$$

,

where x1 through xm is the output of a classifier on the positive examples and y1 through yn its output on the negative examples (Cortes et al., 2003).

For disease prediction, since it is a multiclass classification, a common precision and recall measure to use is Average Precision. It is easy to use and returns a numerical value that can be quickly interpreted

without needing a range of thresholds or plots. It can be thought of as the area under the Precision vs Recall curve and is given by the following formula:

$$AP = \int Precision(t)d[Recall(t)]$$

(Su et al., 2015)

It is important to consider other existing measures depending on the context used. For the case of disease prediction, for example, it is of interest to determine a top of probable diseases to help the decision making of health professionals. With that in consideration, an interesting metric to use is the **rank cumulative frequency**(Masino et al., 2014). The evaluation using rank was done to include the 10 most likely diseases of a patient or the 10 proteins that a given protein is more likely to interact, depending on the scenario of the evaluation.

To do this, for every patient or protein, a ranking was made according to the value of semantic similarity, where the highest value of similarity was ranked first in the ranking and the lowest was last. This ranking was made once for PolBMA and once for BMA. Then, the number of times a given rank appeared in the list of all patients ranks was counted, so that each ranking could have a frequency. For example, if BMA computed the highest similarity for the correct disease for 50 patients out of 100, Rank 1 would have a frequency of 0.5. Once all the Rank frequencies were known, ranks 1 through 10 were organized in an ascending order to create a rank cumulative frequency display.

Chapter 5

Results and Discussion

This chapter shows the results of the evaluating the methodology according to the experimental design. It is structured to consider each prediction scope individually, with section 5.1 focusing on PPI prediction results, section 5.2 on disease prediction results and section 5.3 discussing the overall results.

5.1 Protein-Protein Interaction Prediction

A first step in the evaluation was the computation of ROC and Precision-Recall curves. Figure 5.1 displays the ROC curves for the different PPI datasets. Plots in the first column correspond to results obtained without generating negative annotations and BMA consistently performs better than PolBMA. When negative annotations are introduced,PolBMA performs better in the MF branch for the negative annotations threshold at 0.05 and 0.1, and on the CC branch for 0.05. AUC values are shown in Table 5.1. Figure 5.2 shows the Precision-recall curves for the same datasets and it corroborates the results of the ROC curves.

Table 5.1: Protein-Protein interaction AUC ROC for IC Seco2004

		Negative Annotations Threshold		
GO	SSM	0	0.05	0.1
ALL	BMA	**0.920**	0.904	0.903
	PolBMA	0.780	0.852	0.862
MF	BMA	0.921	0.938	0.931
	PolBMA	0.854	0.957	**0.969**
BP	BMA	**0.920**	0.881	0.872
	PolBMA	0.831	0.806	0.816
CC	BMA	0.857	0.861	0.894
	PolBMA	0.793	**0.929**	0.835

However, ROC and Precision and Recall are global metrics, which are well suited to evaluate performance over a dataset as a whole, but that fail to capture more fine-grained aspects. To support a more specific evaluation, one that is more suitable to cases where prioritization is a goal, rank cumulative distribution curves were calculated for the top 10 predictions. Figures 5.3 and 5.4 show the rank cumulative distribution with IC Seco2004 and IC Resnik1995 respectively. In the first column of either

5. RESULTS AND DISCUSSION

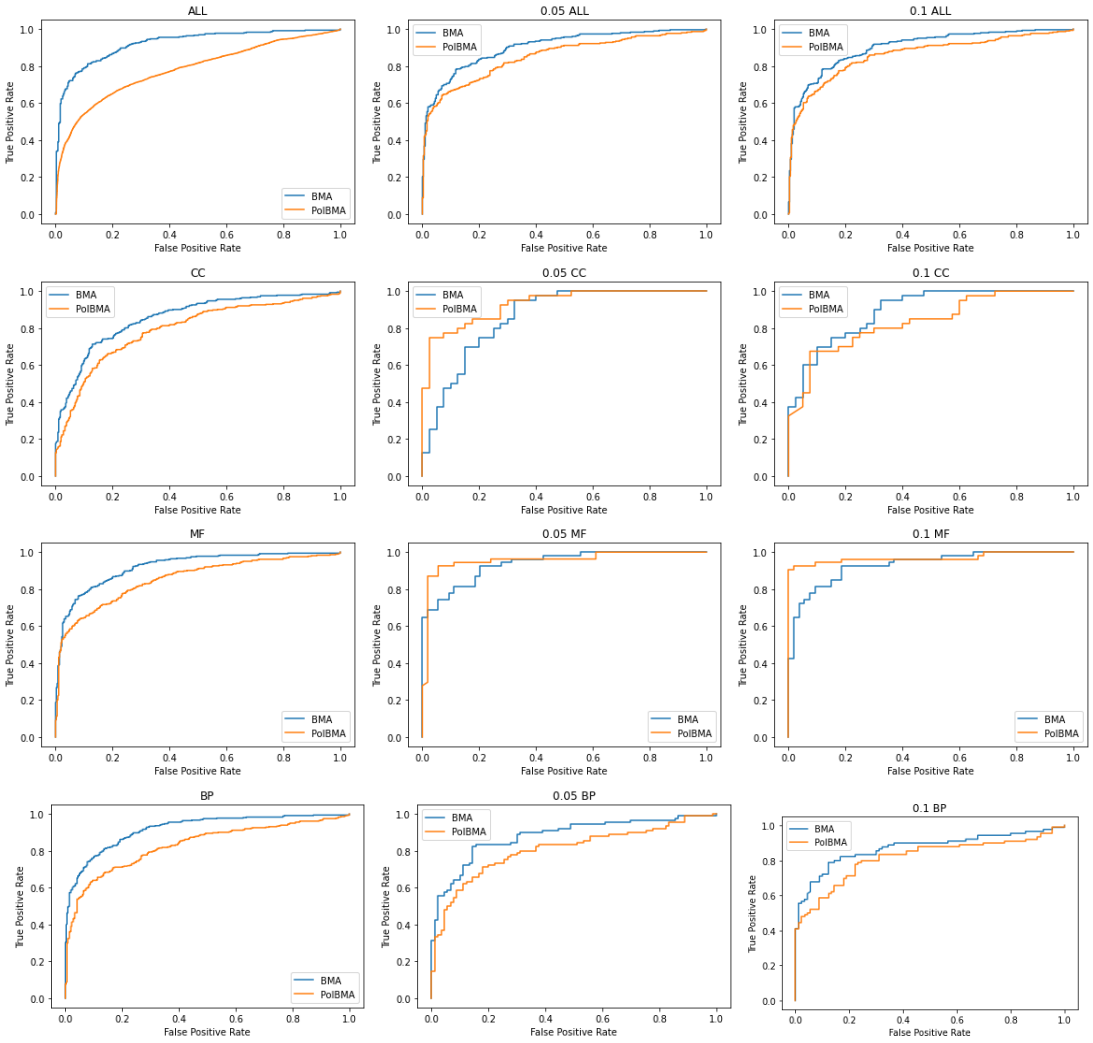
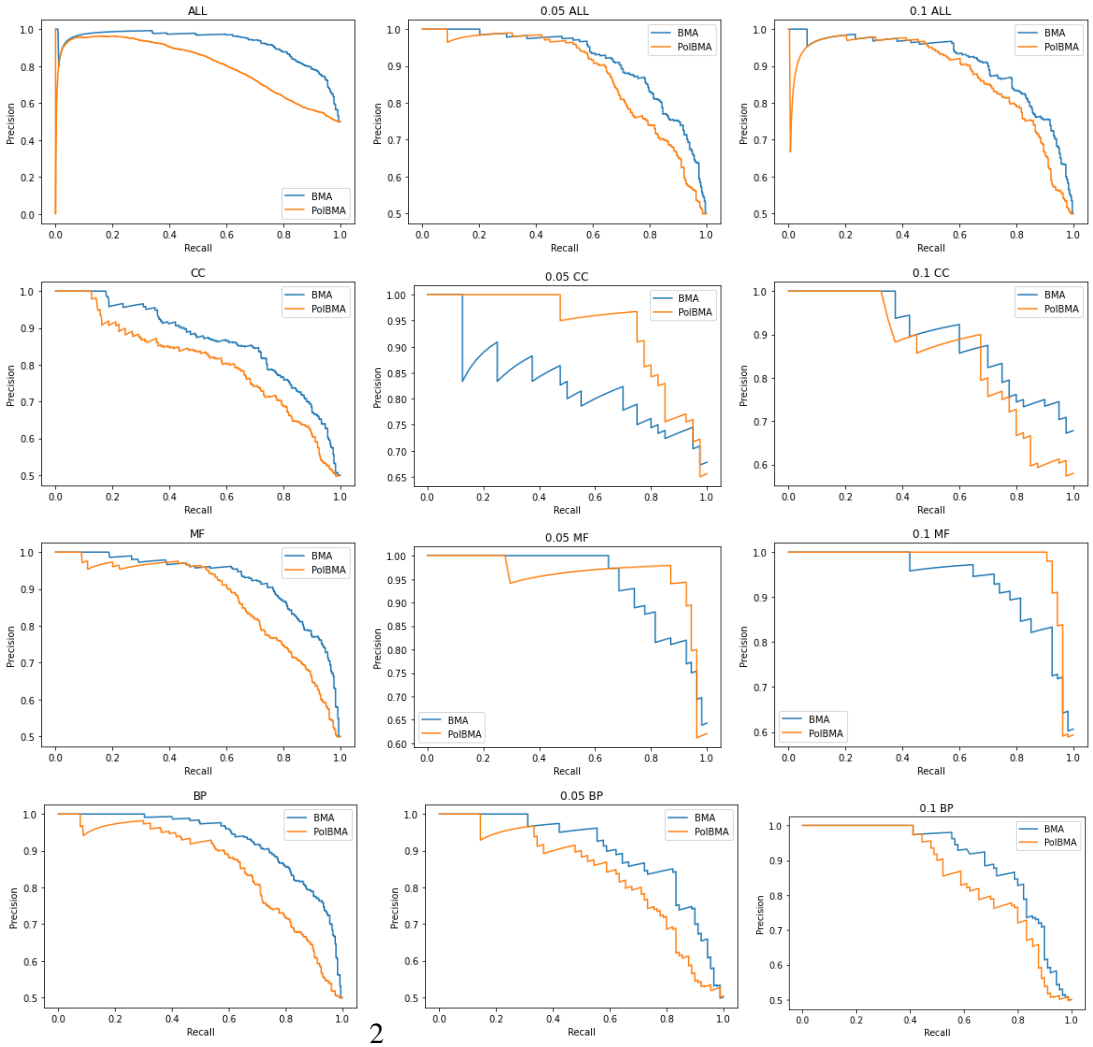


Figure 5.1: Protein-Protein interaction ROC curves for IC Seco2004



2

Figure 5.2: Protein-Protein interaction precision-recall curves for IC Seco2004

figure, no generated negative annotations were considered, and there was no significant improvement in PPI prediction in any of the GO branches when using PolBMA, the results were similar to the baseline. However in the experiment PPI 2 ALL with IC Seco2004, where branches were not discriminated showed an improvement in rank PPI prediction.

In the second and third columns, negative annotations were generated with a threshold of 0.05 and 0.1 respectively, and an improvement was seen in the CC branch of both thresholds in Figure 5.3, while the other experiments showed similar results with the baseline.

When IC Resnik1995 was used, the results were generally similar, except for the MF and CC GO-branches. For CC, the results got a less marked improvement, as it only showed a small difference against the baseline in PPI CC 2 0.05 and a significant improvement only after rank 3 in PPI CC 2 0.1. As for MF, the results improved with Resnik1995, as in PPI MF 2 0.05 and PPI MF 2 0.1 the polar SS measure was significantly better between ranks 4 and 7. PPI BP 2 0.05 had a small but consistent improvement over the baseline in Figure 5.4.

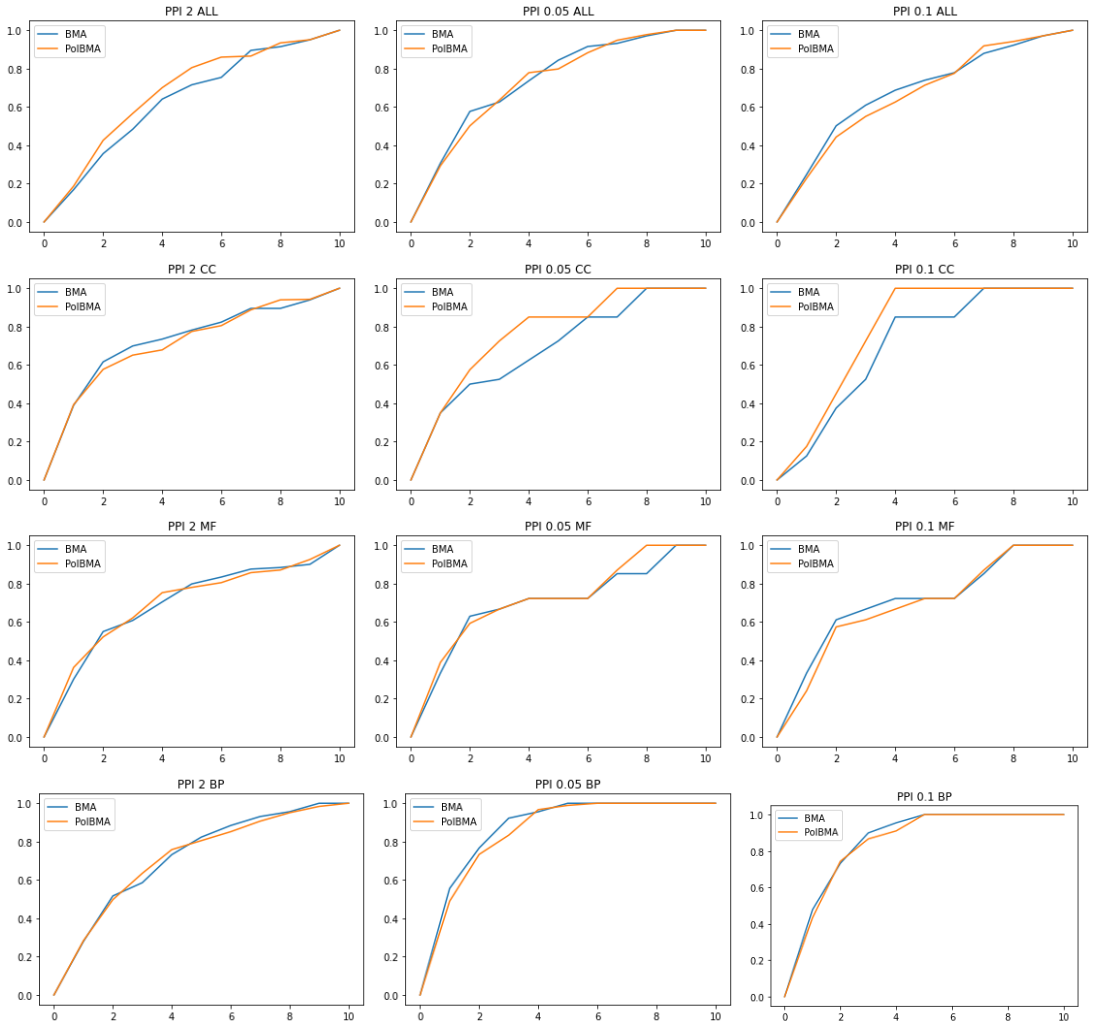


Figure 5.3: Protein-Protein interaction Rank Cumulative Frequency for protein pairs that interact, IC Seco2004

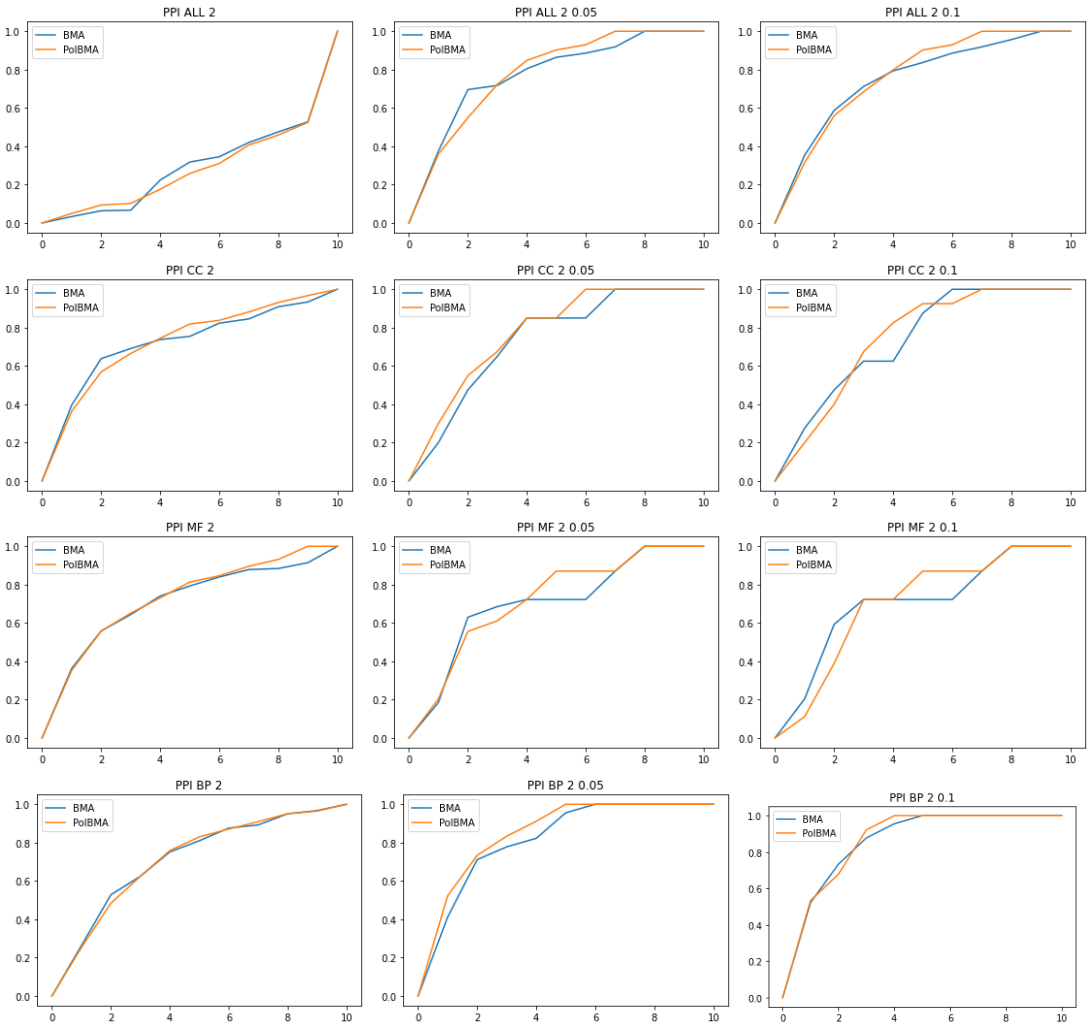


Figure 5.4: Protein-Protein interaction Rank Cumulative Frequency for protein pairs that interact IC Resnik1995

5.2 Disease Prediction

Figures 5.5 and 5.6 show the results of the Disease Prediction evaluation using IC Seco and IC Resnik respectively. In each of the figures, the first column represents the cumulative frequency of 4 different experiments for the PolBMA and the baseline for 3 negative annotation datasets, and the second column shows the same information but in a zoomed-in perspective. The results were consistent for every IC measure and threshold, the polar SSM was significantly better than the baseline, by approximately 10 percent. The best result, with about 0.92 frequency of correct disease prediction was PolBMA, then PolBMA imprecision, BMA, PolBMA Noise, BMA imprecision, PolBMA imprecision and noise, BMA noise and BMA imprecision and noise. Tables 5.2, 5.3 and 5.4 show the same cumulative frequencies, but in a numerical form to show the results in precise values.

It is relevant to note that even when patients do not have negative annotations, PolBMA still outperforms BMA because PolBMA takes into account the negative annotations of diseases when computing the semantic similarity.

Finally, Table 5.5 shows the results for the Average Precision for the different disease prediction datasets. PolBMA has an improved performance of 0.05 to 0.14 in comparison to the baseline, and the greatest difference was in the experiment with a negative threshold of 0.5 and no noise or imprecision.

Table 5.2: Disease Rank Cumulative Frequency for N=0.0

Rank	BMA	PolBMA	BMA imp	PolBMA imp	BMA noise	PolBMA noise
1	0.83	0.92	0.76	0.84	0.69	0.76
2	0.87	0.96	0.83	0.93	0.75	0.83
3	0.88	0.97	0.85	0.94	0.80	0.88
4	0.88	0.97	0.86	0.95	0.81	0.89
5	0.89	0.98	0.87	0.96	0.82	0.90
6	0.89	0.98	0.87	0.96	0.82	0.90
7	0.89	0.98	0.87	0.96	0.83	0.91
8	0.89	0.98	0.88	0.96	0.84	0.93
9	0.89	0.98	0.88	0.97	0.84	0.93
10	0.89	0.98	0.88	0.97	0.85	0.93

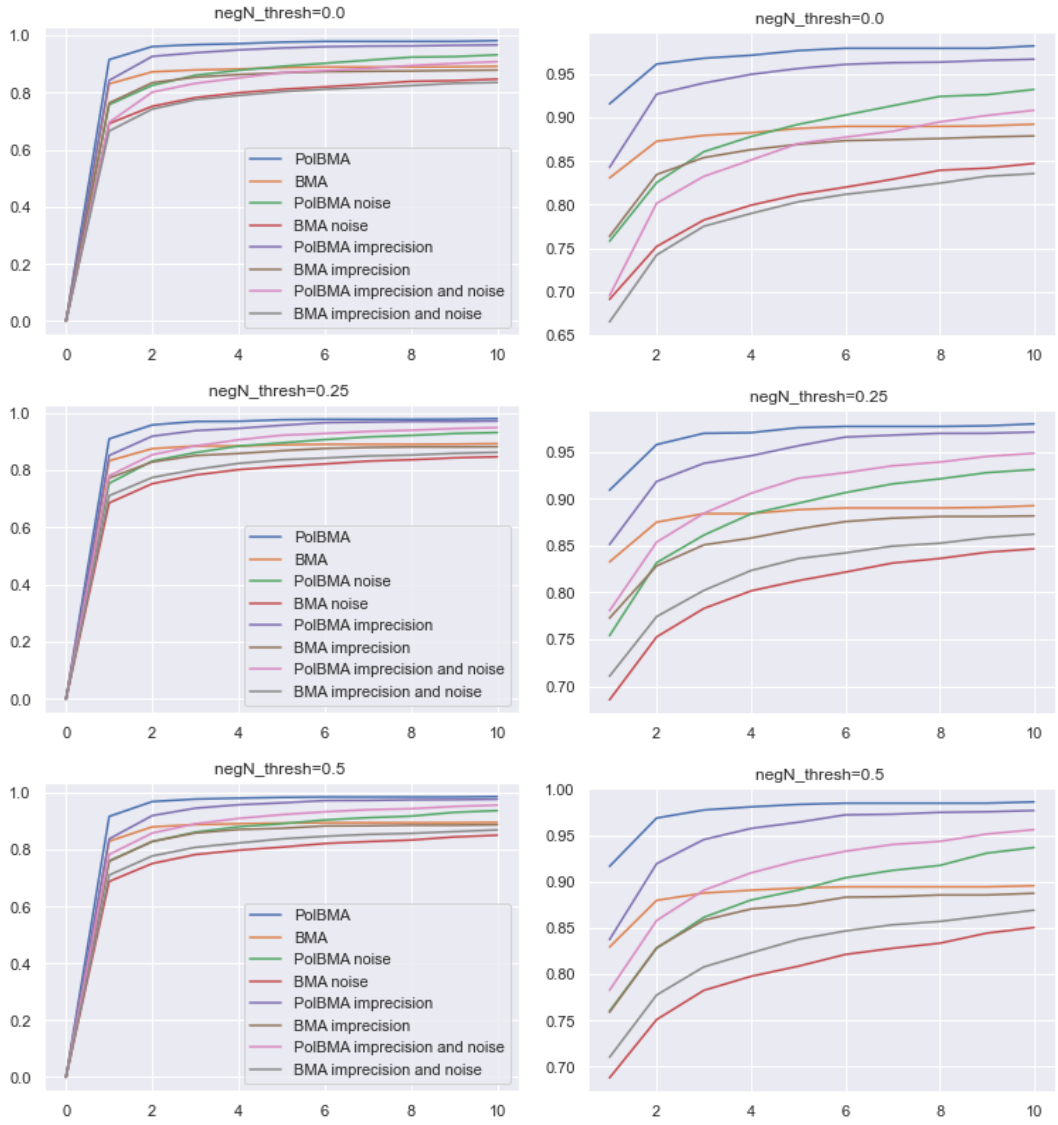


Figure 5.5: Disease Rank Cumulative Frequency using Seco004 as IC measure: Each line of graphs represents cumulative frequencies of a given negative annotations threshold, first on a y scale of 0 to 1, followed by a zoomed version with a y scale from 0.7 to 1

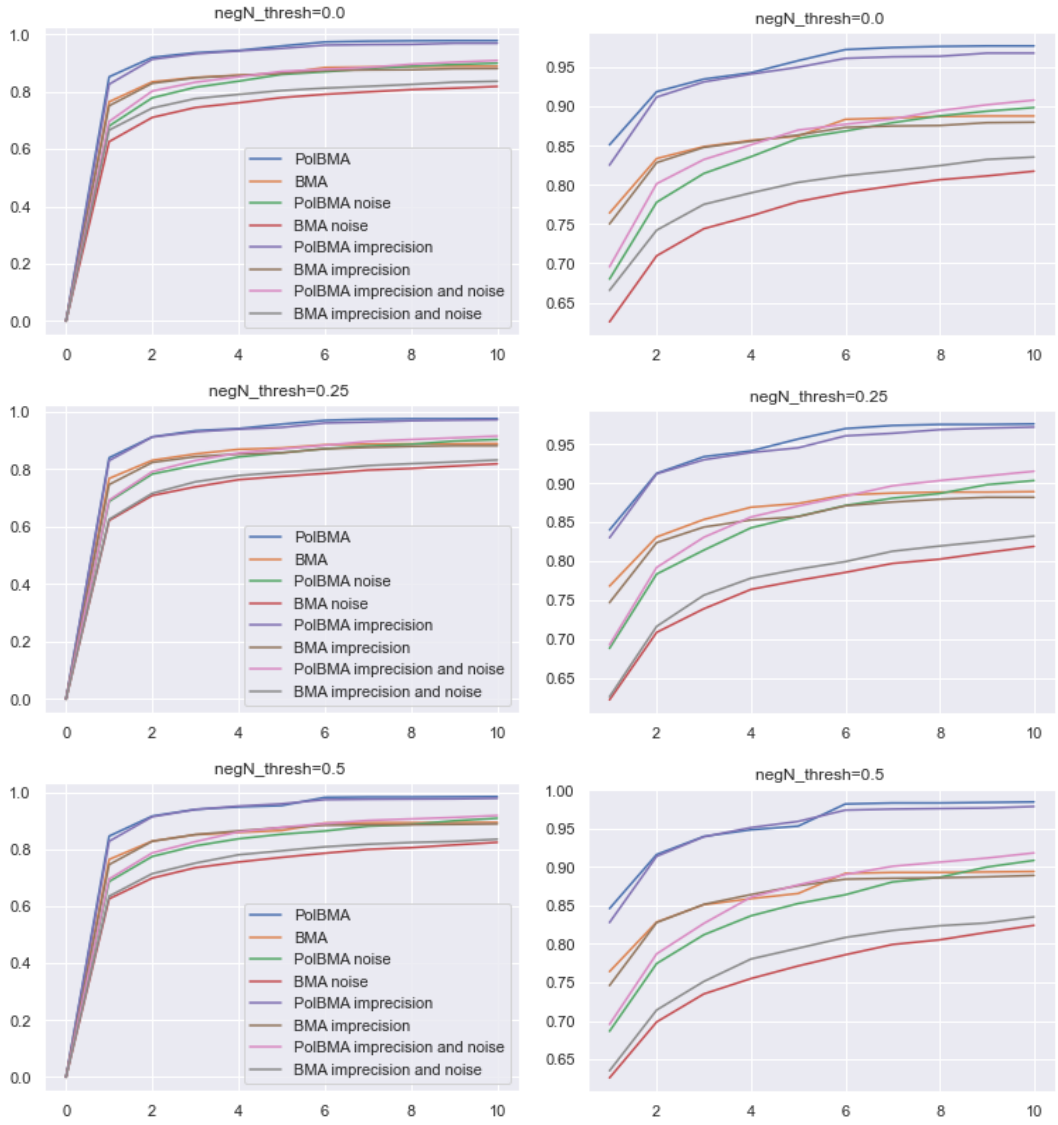


Figure 5.6: Disease Rank Cumulative Frequency using Resnik 1995 as IC measure: Each line of graphs represents cumulative frequencies of a given negative annotations threshold, first on a y scale of 0 to 1, followed by a zoomed version with a y scale from 0.7 to 1

Table 5.3: Disease Rank Cumulative Frequency for N=0.25

Rank	BMA	PolBMA	BMA imp	PolBMA imp	BMA noise	PolBMA noise
1	0.83	0.90	0.77	0.85	0.69	0.75
2	0.88	0.96	0.83	0.92	0.75	0.83
3	0.88	0.97	0.85	0.94	0.78	0.86
4	0.88	0.97	0.86	0.95	0.80	0.88
5	0.89	0.98	0.87	0.96	0.81	0.89
6	0.89	0.98	0.88	0.97	0.82	0.90
7	0.89	0.98	0.88	0.97	0.83	0.92
8	0.89	0.98	0.88	0.97	0.84	0.92
9	0.89	0.98	0.88	0.97	0.84	0.93
10	0.89	0.98	0.88	0.97	0.85	0.93

Table 5.4: Disease Rank Cumulative Frequency for N=0.5

Rank	BMA	PolBMA	BMA imp	PolBMA imp	BMA noise	PolBMA noise
1	0.83	0.92	0.76	0.84	0.69	0.76
2	0.88	0.97	0.83	0.92	0.75	0.83
3	0.89	0.98	0.86	0.95	0.78	0.86
4	0.89	0.98	0.87	0.96	0.80	0.88
5	0.89	0.98	0.87	0.96	0.80	0.89
6	0.89	0.98	0.88	0.97	0.82	0.90
7	0.89	0.98	0.88	0.97	0.83	0.91
8	0.89	0.98	0.89	0.97	0.83	0.92
9	0.89	0.98	0.89	0.98	0.84	0.93
10	0.90	0.99	0.89	0.98	0.85	0.94

Table 5.5: Average Precision

Measure	NegThreshold=0	NegThreshold=0.25	NegThreshold=0.5
PolBMA	0.78	0.7	0.79
BMA	0.65	0.65	0.65
Measure	NegThreshold=0 Noise	NegThreshold=0.25 Noise	NegThreshold=0.5 Noise
PolBMA	0.5	0.51	0.52
BMA	0.43	0.42	0.42
Measure	NegThreshold=0 Imprecision	NegThreshold=0.25 Imprecision	NegThreshold=0.5 Imprecision
PolBMA	0.61	0.62	0.64
BMA	0.51	0.50	0.52
Measure	NegThreshold=0 Both	NegThreshold=0.25 Both	NegThreshold=0.5 Both
PolBMA	0.53	0.54	0.56
BMA	0.45	0.45	0.45

5.3 Discussion

Results from the PPI evaluation in the GO show that a polar measure is only significantly and consistently better than a non polar measure in the MF branch and the CC branch for the 0.05 threshold. This could indicate that negative annotations have a greater impact on where a protein is active, and what function it performs rather than on what processes it participates in. However, when looking at the dataset properties, the CC and MF branches have a smaller ratio of positive to negative annotations. In average, proteins of the CC and MF datasets with generated negative annotations have 12 to 14 positive annotations and 1 negative annotation in the 0.05 threshold and MF has 16 positive to 1 negative annotation in the 0.1 threshold, while the other branches have 16 to 19 positive annotations to 1 negative annotation in the first case, and 18 to 20 positive annotations to 1 negative annotation in the second threshold. This could mean that negative annotations improve results when they are more common.

As for the disease prediction evaluation, PolBMA consistently improves prediction. Both Average precision and rank cumulative frequency show an improvement in performance of about 10%. For about 10% of the patients, BMA would classify the target illness with a very low rank, while PolBMA did not. This could be due to similar illnesses where a key negative annotation provided the needed evidence to tell them apart. The presence of negative annotations in the diseases has a bigger impact on the increased performance of PolBMA than the different thresholds of negative annotations on the patients, which gives emphasis to the limitations associated with incomplete ontologies. From related work, noise was expected to have a significant impact on the disease prediction, and it did decrease the ranking cumulative frequency performance in about 15%. This impact was similar in PolBMA and BMA. Imprecision also decreased the performance in about 10%. Imprecision did not impact the results as much as noise because the semantic similarity measures take advantage of the taxonomic relationships between terms to determine that imprecise classes are still similar to the original symptoms. In the experiment that considers both noise and imprecision, results were slightly better than noise and worse than imprecision. This result was unexpected because noise and imprecision both negatively impact performance, but it appears imprecision is able to mitigate some of the noise induced limitations.

In the disease prediction dataset characteristics, Mendelian diseases have more annotations then the random ailments that were added from HPO to cause noise. The greater number of annotations could introduce bias toward the thirty-three Mendelian diseases, however this was introduced in both polar and non polar measures so it shouldn't impact their comparison and patterns, only the cumulative frequency values could be overestimated.

Disease prediction showed a more consistent improvement when negative annotations were used in Disease prediction than PPI prediction. Human disease has a more pressing need for negative annotations, probably because similar diseases can be differentiated by a key negative phenotype that is not considered in conventional semantic similarity measures.

Chapter 6

Conclusions

This work tested whether the addition of negative semantic annotations to semantic similarity measures would impact semantic similarity. To do so, BMA and Resnik pairwise semantic similarity measures were adapted to incorporate negative annotation in their computation, and PolBMA and Polar Resnik were developed. To evaluate these, and to determine if negative annotations had an impact on semantic similarity, PPI prediction and disease prediction datasets were created.

Results show that because of the scarcity of negative annotations in mainstream ontologies, common evaluation measures might overlook their impact, however when negative annotations are generated to PPI and disease prediction datasets, semantic similarity measures that consider them can have better performance than conventional baselines. Even though the PPI dataset did not show consistent improvement in PPI prediction when polar semantic similarity measures were used, the disease prediction dataset had a 10 percent improvement in performance when negative annotations were added to semantic similarity computation. The results lead to the conclusion that negative annotations have an impact on semantic similarity, that depends on the scope of the research and on the amount of negative annotations considered. The magnitude of this impact required further study to be better understood. Nevertheless, this study has highlighted the impact that negative annotations can have in semantic similarity applications.

6.1 Contributions

This work produced the following contributions:

- development of two semantic similarity measures: PolBMA for annotated entities and PolarResnik for single classes, that that into account both positive and negative annotations.
- creation of datasets populated with negative annotations in two relevant biomedical scopes: protein-protein interaction prediction and disease prediction.
- an evaluation on the impact of negative annotations in semantic similarity applications.

6.2 Future work

This dissertation determined that negative annotations impact semantic similarity, and highlighted the limitation of the lack of negative annotations in biomedical ontologies. The magnitude of the impact requires further study, and to do so the main points of study can include:

- Performing studies with a higher proportion of negative annotations. Thus can both stem from generating more automated negative annotations or perform an ablation on the positive annotations. The number of negative annotations should be theoretically higher than the number of positive annotations, but in this dissertation, they still accounted for one to three order of magnitude less than the positive annotations
- Improve negative annotation generation, so that these can be generated faster, more efficiently and in greater numbers.
- Application to other semantic similarity applications, such as gene-disease association prediction or gene function prediction.

Bibliography

Ashburner, M. et al. (Jan. 2000). "Gene ontology: Tool for the unification of biology". In: *The Gene Ontology Consortium. Nat Genet* 25, pp. 25–29.

Bodenreider, Olivier and Robert Stevens (Oct. 2006). "Bio-ontologies: Current trends and future directions". In: *Briefings in bioinformatics* 7, pp. 256–74. DOI: 10.1093/bib/bbl027.

Boyd, Kendrick, Kevin H. Eng, and David Page (2013). "Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals". In: *ECML/PKDD*.

Consortium, The et al. (Dec. 2020). "The Gene Ontology resource: Enriching a GOld mine". In: *Nucleic Acids Research* 49. DOI: 10.1093/nar/gkaa1113.

Cortes, Corinna and Mehryar Mohri (2003). "AUC optimization vs. error rate minimization". In: *Advances in neural information processing systems* 16, pp. 313–320.

Couto, Francisco et al. (Feb. 2006). "GOAnnotator: Linking protein GO annotations to evidence text". In: *Journal of biomedical discovery and collaboration* 1, p. 19. DOI: 10.1186/1747-5333-1-19.

Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram (2019). "Designing Disease Prediction Model Using Machine Learning Approach". In: *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1211–1215. DOI: 10.1109/ICCMC.2019.8819782.

Davis, Jesse and Mark Goadrich (2006). "The Relationship between Precision-Recall and ROC Curves". In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 233–240. ISBN: 1595933832. DOI: 10.1145/1143844.1143874. URL: https://doi.org/10.1145/1143844.1143874.

Dessimoz, Christophe and Nives Škunca (Jan. 2017). *The Gene Ontology Handbook*. Vol. 1446. ISBN: 978-1-4939-3741-7. DOI: 10.1007/978-1-4939-3743-1.

Fu, Guangyuan et al. (June 2016). "NegGOA: negative GO annotations selection using ontology structure". In: *Bioinformatics* 32.19, pp. 2996–3004. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw366. eprint: https://academic.oup.com/bioinformatics/article-pdf/32/19/2996/25072589/btw366.pdf. URL: https://doi.org/10.1093/bioinformatics/btw366.

Gonzalez, Mileidy W. and Maricel G. Kann (Dec. 2012). "Chapter 4: Protein Interactions and Disease". In: *PLOS Computational Biology* 8.12, pp. 1–11. DOI: 10.1371/journal.pcbi.1002819. URL: https://doi.org/10.1371/journal.pcbi.1002819.

Harispe, Sébastien et al. (May 2015). *Semantic Similarity from Natural Language and Ontology Analysis*. Vol. 8. DOI: 10.2200/S00639ED1V01Y201504HLT027.

Hart, G. Traver, Arun K. Ramani, and Edward M. Marcotte (2006). "How complete are current yeast and human protein-interaction networks?" In: *Genome Biology* 7, pp. 120–120.

Hill, David et al. (Feb. 2008). "Gene Ontology annotations: What they mean and where they come from". In: *BMC bioinformatics* 9 Suppl 5, S2. DOI: 10.1186/1471-2105-9-S5-S2.

Hoehndorf, Robert, Paul Schofield, and Georgios Gkoutos (Apr. 2015). "The role of ontologies in biological and biomedical research: A functional perspective". In: *Briefings in bioinformatics* 16. DOI: 10.1093/bib/bbv011.

Hossin, Mohammad and Sulaiman M.N (Mar. 2015). "A Review on Evaluation Metrics for Data Classification Evaluations". In: *International Journal of Data Mining Knowledge Management Process* 5, pp. 01–11. DOI: 10.5121/ijdkp.2015.5201.

Jain, Shobhit and Gary Bader (Nov. 2010). "An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology". In: *BMC bioinformatics* 11, p. 562. DOI: 10.1186/1471-2105-11-562.

Köhler, Sebastian, Michael Gargano, et al. (Dec. 2020). "The Human Phenotype Ontology in 2021". In: *Nucleic Acids Research* 49.D1, pp. D1207–D1217. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1043. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D1207/35364524/gkaa1043.pdf. URL: https://doi.org/10.1093/nar/gkaa1043.

Köhler, Sebastian, Marcel Schulz, et al. (Oct. 2009). "Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies". In: *American journal of human genetics* 85, pp. 457–64. DOI: 10.1016/j.ajhg.2009.09.003.

Kohli, Pahulpreet Singh and Shriya Arora (2018). "Application of Machine Learning in Disease Prediction". In: *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–4. DOI: 10.1109/CCAA.2018.8777449.

Masino, Aaron et al. (July 2014). "Clinical phenotype-based gene prioritization: An initial study using semantic similarity and the human phenotype ontology". In: *BMC bioinformatics* 15, p. 248. DOI: 10.1186/1471-2105-15-248.

McDowall, Mark D., Michelle S. Scott, and Geoffrey J. Barton (Nov. 2008). "PIPs: human protein–protein interaction prediction database". In: *Nucleic Acids Research* 37.suppl$_1$, pp. D651–D656. ISSN: 0305-1048. DOI: 10.1093/nar/gkn870. eprint: https://academic.oup.com/nar/article-pdf/37/suppl_1/D651/3330421/gkn870.pdf. URL: https://doi.org/10.1093/nar/gkn870.

Mohamed, Thahir, Jaime Carbonell, and Madhavi Ganapathiraju (Jan. 2010). "Active learning for human protein-protein interaction prediction". In: *BMC bioinformatics* 11 Suppl 1, S57. DOI: 10.1186/1471-2105-11-S1-S57.

Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava (2019). "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques". In: *IEEE Access* 7, pp. 81542–81554. DOI: 10.1109/ACCESS.2019.2923707.

Palaniappan, Sellappan and Rafiah Awang (2008). "Intelligent heart disease prediction system using data mining techniques". In: *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pp. 108–115. DOI: 10.1109/AICCSA.2008.4493524.

Pesquita, Catia, Daniel Faria, Hugo Bastos, et al. (Feb. 2008). "Metrics for GO based protein semantic similarity: A systematic evaluation". In: *BMC bioinformatics* 9 Suppl 5, S4. DOI: 10.1186/1471-2105-9-S5-S4.

Pesquita, Catia, Daniel Faria, Andre O Falcao, et al. (2009). "Semantic similarity in biomedical ontologies". In: *PLoS computational biology* 5.7, e1000443.

Pesquita, Catia, Daniel Faria, André O. Falcão, et al. (July 2009). "Semantic Similarity in Biomedical Ontologies". In: *PLOS Computational Biology* 5.7, pp. 1–12. DOI: 10.1371/journal.pcbi.1000443. URL: https://doi.org/10.1371/journal.pcbi.1000443.

Resnik, Philip (1995). "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In: *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453.

Rubin, Daniel L., Nigam H. Shah, and Natalya F. Noy (Dec. 2007). "Biomedical ontologies: a functional perspective". In: *Briefings in Bioinformatics* 9.1, pp. 75–90. ISSN: 1467-5463. DOI: 10.1093/bib/bbm059. eprint: https://academic.oup.com/bib/article-pdf/9/1/75/827967/bbm059.pdf. URL: https://doi.org/10.1093/bib/bbm059.

Seco, Nuno, Tony Veale, and Jer Hayes (Jan. 2004). "An Intrinsic Information Content Metric for Semantic Similarity in WordNet." In: vol. 16, pp. 1089–1090.

Su, Wanhua, Yan Yuan, and Mu Zhu (2015). "A Relationship between the Average Precision and the Area Under the ROC Curve". In: *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. ICTIR '15. Northampton, Massachusetts, USA: Association for Computing Machinery, pp. 349–352. ISBN: 9781450338332. DOI: 10.1145/2808194.2809481. URL: https://doi.org/10.1145/2808194.2809481.

Teixeira, David Carriço (2019). "Understanding ALS patients using Semantic Similarity". MA thesis. Lisboa, Portugal: Faculdade de Ciências da Universidade de Lisboa.

Teng, Zhixia et al. (Apr. 2013). "Measuring gene functional similarity based on group-wise comparison of GO terms". In: *Bioinformatics* 29.11, pp. 1424–1432. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt160. eprint: https://academic.oup.com/bioinformatics/article-pdf/29/11/1424/587849/btt160.pdf. URL: https://doi.org/10.1093/bioinformatics/btt160.

Vesztrocy, Alex Warwick and Christophe Dessimoz (2020). "Benchmarking gene ontology function predictions using negative annotations". In: *Bioinformatics* 36, pp. i210–i218.

Wu, Xiaomei et al. (May 2013). "Improving the Measurement of Semantic Similarity between Gene Ontology Terms and Gene Products: Insights from an Edge- and IC-Based Hybrid Method". In: *PLOS ONE* 8.5, pp. 1–11. DOI: 10.1371/journal.pone.0066745. URL: https://doi.org/10.1371/journal.pone.0066745.

Youngs, Noah et al. (June 2014). "Negative Example Selection for Protein Function Prediction: The NoGO Database". In: *PLOS Computational Biology* 10.6, pp. 1–12. DOI: 10.1371/journal.pcbi.1003644. URL: https://doi.org/10.1371/journal.pcbi.1003644.

Zhou, Naihui et al. (Nov. 2019). "The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens". In: *Genome Biology* 20. DOI: 10.1186/s13059-019-1835-8.