AN ONTOLOGICAL APPROACH TO QUANTIFY DISTANCE BETWEEN HEREDITARY DISEASE MODULES ON THE INTERACTOME.

DEPARTMENT OF COMPUTER SCIENCE

Royal Holloway, University of London

A Thesis submitted for the degree of Doctor of Philosophy

Horacio José Caniza Vierci

December 2015

Declaration of Authorship

I, Horacio José Caniza Vierci, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

_____

Signed

_____

Date

Acknowledgements

It is definitely a strange feeling having to write the acknowledgements for a PhD thesis. As I write them I find myself looking back, weighing the work I have done, searching for value.

I would like to thank my wife, Anahí, for standing by my side. We sat together through most chapters of this thesis and many lines of code. She was the strength on days where things were tiring, and the calm when things where overwhelming. Having the opportunity to come home to her after a long day is one of life's great pleasures. I will always remember our small house at 6 Englefield Close as not being big enough to hold all the love and happiness we brought into it.

I would like to thank my parents, José María and Margarita. They gave me the greatest gift any son could want: the gift of peace. They allowed me to finish my journey knowing that I was supported for whatever I might need. They pushed me when pushing was needed, and nurtured my incessant questioning while I was growing up. They are the reason I decided that this journey was possible, and I will always be thankful to them.

These past 4 years were busy years, I have worked many hours and I have learned a lot. All was done with gusto and the late nights were followed by long days with little ill effects. This is due in part to my supervisor, Professor Alberto Paccanaro. The word "supervisor" is, I feel, a bit removed from the relationship we have developed; Alberto is a great Teacher and a great friend. I am lucky to have been able to learn from someone who knows so much and is so willing to pass that knowledge on.

Among the many things I have learned during my PhD, one lesson will not expire: the best company one could possibly keep is that of smarter people. I was lucky to work among the members of the PaccanaroLab. Dr. Haixuan Yang, Dr. Prajwal Bhat, Dr. Emilio Ferrara, Dr. Beatrix Horvath and Juan Cáceres are extremely

smart and kind people, who took time to help me find the way. I am happy to have been able to spend more time with my good friend Dr. Alfonso E. Romero. We shared discussions on science and life, and shared many long conversations over quite a few pints.

I feel the responsibility as a Paraguayan to thank Prof. Paccanaro for his efforts to help strengthen the scientific community in Paraguay. With no bonds other than friends who remember him fondly as a good man, he took it upon himself to help the small country I call home. He took more trips than I can count and worked longer hours than he should have, exerting tremendous effort aimed at a foreign country. This was done without hope for retribution and with a drive that, if it were more common, the world would be a far better place.

Finally, we do not speak English at home, so I feel a bit of Spanish is needed.

A mi esposa Anahí. A mis padres José María y Margarita. A mis hermanos Silvia, Susana y Joaquín. A mis sobrinos Martín, Emilio, Francisco, Joaquín y Victoria.

Abstract

For about 30% of hereditary diseases no disease gene is currently known. Very little if anything at all is known about the molecular basis of these orphan diseases. In this Thesis I present an ontological method that accurately quantifies similarity between heritable diseases modules in the interactome, which can be used to help pinpoint the location of the perturbation causing the orphan diseases . This method, based on the MeSH ontologies, effectively brings together the existing information about diseases that is scattered across the vast corpus of biomedical literature.

I prove that sets of MeSH terms provide a highly descriptive representation of heritable disease and that the structure of MeSH provides a natural way of combining individual MeSH vocabularies. I also show that the measure can be used effectively in the prediction of candidate disease genes. The effective use of the vast information available allows the measure to be applicable for orphan diseases: the measure can help pinpoint the location of their molecular perturbations. More generally, the measure enables the transfer of knowledge between similar diseases, providing hypotheses for disease genes and even suggestions for drug repositioning.

I have validated the method through a machine learning approach to show the predictive power of the measure. Further to the numerical evaluation, I have curated a highly illustrative set of examples for the literature showcasing the accuracy of the method. Lastly, I show that the measure is effective for the prediction of candidate disease genes. I have developed a web application to query more than 28.5 million relationships between 7,574 hereditary diseases (96% of OMIM) based on the similarity measure.

During my PhD I have also developed GOssTo and GOssToWeb a console and web application to compute semantic similarities in the Gene Ontology. GOssTo was integrated into a disease gene prediction pipeline that showed the advantages of using functional similarities to improve the predictions.

Contents

List of Tables

List of Figures

Chapter 1

Biomedical Ontologies

The need for standardised vocabularies in the natural sciences arose as early as the 17Th century with Linnaeus's taxonomy. In the biomedical field, this need was recognised as early as the 19Th century with the precursors of International Classification of Diseases (ICD+) [1]. While the main aim at the time was to obtain a uniform vocabulary as a means to classify living things and to enable statistical analysis of the incidence of diseases [1], these vocabularies were extended and refined over the past decades. The new ontologies and classification systems, although built on the same principles changed to reflect the exponential growth of data [32].

The modern biomedical ontology, such as the Gene Ontology (GO) and the Disease Ontology (DO), not only provides this standard nomenclature, but they have become a subject of study in their own right.

In this chapter, I will explore the biomedical ontologies more relevant to my work, analysing their driving principles, evolution through time and their ontological structure. I will also present methods to determine similarity of genes based on the ontological structure of the Gene Ontology (GO).

1.1 Ontologies: A shared domain language

The term ontology, in Computer Science, is related to the philosophical concept of Ontology. Philosophically "Ontology" is the study of nature, the basic categories of life, existence and all their relationships. In Computer Science, and particularly in knowledge representation, the concept is, expectedly, narrower. The concept of "existence" in Computer Science, is reduced to the world view that can be effectively represented [92], thereby reducing the scope of the term.

In Computer Science, an "Ontology" describes shared knowledge of a particular domain [67]. Broadly, it is a description of a system, its parts and relationships, that is shared between a group of people. That is, it represents a specific world view [67]. Formally, an ontology is defined as a formal representation of a shared conceptualisation [67].

An ontology can, therefore, represent a specific domain, such as the organisation of a company or the structure of the army. A few examples stand out in, such as BabelNet [77], WordNet [74] and Umbel [97].

1.2 Biomedical ontologies

The need for wider, more specific standardised vocabularies in biology became apparent with the advent of large-scale functional analysis of proteins. The experimental validation of the extent of the functional conservation of proteins in orthologues [59] revealed the need for a cross-organism ontologies to describe genes and their products accurately. The exponential amount of information being added required standardised means to better use this information, driving Molecular Biology data analysis to the forefront "Big Data" [99].

Larger databases, such as Online Mendelian Inheritance in Man (OMIM) highlight

the relevance of the biomedical ontologies. OMIM's main focus is the relationship between the genes and their resulting disease phenotypes, in particular for those genes that have Mendelian inheritance patterns [4]. The database contains over 15,000 genes and more than 7,500 human diseases, of which, nearly 70% have associated some genetic background.

Perhaps the most significant difficulty when systematically analysing databases such as OMIM is their lack of structure. While comprehensive, OMIM is mainly aimed at medical practitioners, and it has extensive and detailed free-text disease descriptions that are inadequate for automated analysis. It is characterised by a loose structure, represented by entries referencing one another, consists of relations that are neither semantically defined nor abundant enough to be exploited in a systemic analysis. For example *Saethre-chotzen syndrome* (`MIM:101400`) refers to *Muenke syndrome* (`MIM:602849`) describing it as having "similar overlapping phenotype". The same entry for (`MIM:101400`) also refers to *Craniosynostosis* (`MIM:123100`), but in this case the referred entry contains information about structural changes in the cytogenetic location, and finally, it refers to *Cephalopolysyndactyly Syndrome* (`MIM:175700`), to indicate that it "appears to be located" in the same cytogenetic region [5]. The ontologies provide the structure that, when used appropriately as I will show, provide a tool for the large scale analysis of databases such as OMIM.

The ontologies analysed in this section are the most relevant for my work.

1.2.1 Gene Ontology (GO)

The Gene Ontology (GO) is a community effort to manually create a standard nomenclature for genes and gene products [59, 96]. It is based on the Open Biological and Biomedical Ontologies (OBO) [15] concepts, and originally the project included the model organisms *D. melanogaster*, *M. musculus* and *S. cerevisiae*, growing to include over 30 organisms in the current releases [3].

Gene Ontology (GO) is based on the Open Biological and Biomedical Ontologies (OBO) [15] concepts and it is organised into three ontologies. The *Molecular Function* ontology is composed of terms that describe activities at molecular level *e.g. Lactase activity* (`GO:0000016`). The *Biological Process* ontology is composed of terms that describe tasks carried out by genes and gene products, either independently or as part of a protein complex *e.g. Regulation of DNA recombination* (`GO:0000018`). And finally, the *Cellular Component* ontology describes the components of a cell, such as the *Polarisome* (`GO:0000133`).

1.2.2 The ontological structure

The ontological structure of each ontology in GO is defined by the relationships between its terms. In GO, 8 possible relationships link the terms.

is_a

The *is_a* relationship, occurring 71,177 times, defines the main structure of GO. It defines instances, in the class sense, of the terms. That is, *Reproduction* $\overrightarrow{is\_a}$ *Biological Process* implies that *Reproduction* is an instance of or a type of *Biological Process*.

part_of and *has_part*

The *part_of* relationship is the second most common relationship, occurring 8,573 times in GO. This relationship defines part-whole relationships. That is, if $A \overrightarrow{part\_of} B$, then A is a constituent part of B and the existence of A implies the existence of B [3]. For example, *Transcription factor activity, protein binding* (`GO:0000988`) $\overrightarrow{part\_of}$ *Regulation of nucleic acid-templated transcription* (`GO:1903506`).

The *has_part*, while not the complement of *part_of*, defines a logical reciprocal

[3]. It is used to indicate a relationship between two terms where one is a necessary part of the other, from the perspective of the parent. Following an *has_part* link implies an increase in specificity. For example, *Transcription factor activity, protein binding* (GO:0000988) $\xrightarrow{has\_part}$ *Protein binding* (GO:0005515), indicating that the transcription factor activity is always composed of protein binding. The *has_part* relationship occurs 710 times in GO.

regulates and the sub relationships *negatively_regulates* and *positively_regulates*

These relationships link processes where one of the processes directly affects the other [3], that is, *A regulates B* implies that *A* necessarily regulates *B*. The processes are either positively or negatively regulated, however, the relationship *regulates* is used when not enough information is available to accurately qualify the nature of the regulation. For example, *Regulation of mitotic cell cycle* (GO:0007346) $\xrightarrow{regulates}$ *Mitotic cell cycle* (GO:0000278). The *negatively_regulates* relationship occurs 2,857 times, *positively_regulates* 2,828 times and *regulates* 3,286 times.

happens_during and *occurs_in*

These relationships are special in the sense that, to the best of my knowledge, there is no specific definition of them in the GO documentation [3]. I analysed 10 previous releases of GO dating from 2006 to 2015, and the first occurrence of both *happens_during* and *occurs_in* is in 2015. From the existing *occurs_in* relationships in GO I was to infer that these relationships link together the Biological Process and Cellular Component ontologies thereby describing the physical location where a process takes place.

These relationships are rare, with *happens_during* being used only once relating the terms *Uterine Wall Breakdown* (GO:0042704) and *Menstruation* (GO:0042703).

The *occurs_in* relationship is used a total of 178 times.

Table 1.1 shows the topological characteristics of the GO ontologies. Since the majority (89.7%) of the relationships are represented in the *is_a* and *part_of* relationships, the depth and average number of children were calculated exclusively based on these two relationships. The maximum depth of the ontologies is determined by the length of the shortest path.

Ontology	Terms	Leaves	Max.depth	Avg.children
Biological Process	28,247	11,342	13	2.09
Molecular Function	10,944	7,729	11	1.21
Cellular Component	3,809	2,463	9	1.87

Table 1.1: *Topological characteristics of the three ontologies in the Gene Ontology*

Figures 1.1, 1.2, 1.3 show the number of nodes in each level of the ontology. This figure contrasts the number of leaves (red bars) and non-leaf (blue) terms in each level of the Biological Process ontology.

Not all ontologies GO grew from its inception. In particular, from the first release in 2006, the Biological Process ontology remained fixed at 13 levels, while the Molecular Function grew from 10 levels to the current 11 and the Cellular Component ontology grew from 8 to 9.

The growth, however, is not only reflected in the inclusion of new terms but also in the refinement of the ontological structure through the removal of obsolete terms. I explored all releases from 2006 to 2015 of GO, analysing growth and maintenance of the existing ontological structure. Figure 1.4 shows the number of terms in each yearly release of GO compared to the number of obsolete terms in that year. With a rate of growth of 128% since its inception (from $12,346$ in 2006 to $28,247$ in 2015) the growth process greatly exceeds the removal of obsolete terms, which grew by 89% from $1,011$ in 2006 to $1,918$ in 2015. This reflects the highly directed process that guides GO's growth: terms are added whenever the need arises [59, 96].

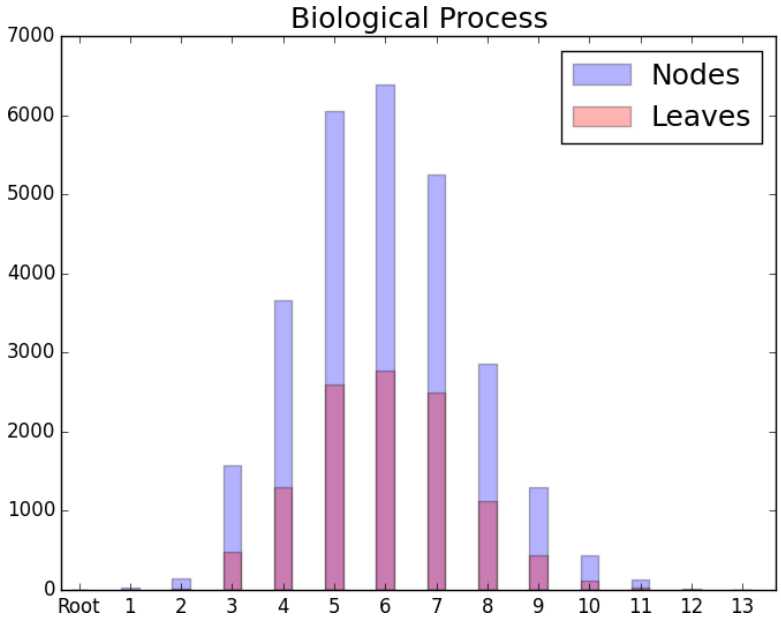

Figure 1.1: Distribution of nodes and leaves per level in the Biological Process ontology. The blue bars show the number of descriptors per level in the Biological Process ontology of GO and the red bars the number of leaves in each level.

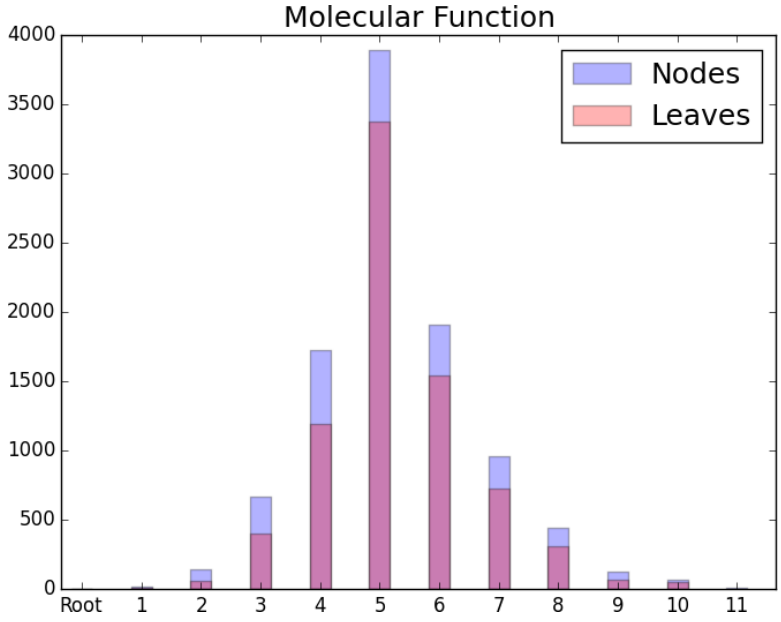

Figure 1.2: Distribution of nodes and leaves per level in the Molecular Function ontology. The blue bars show the number of descriptors per level in the Molecular Function ontology of GO and the red bars the number of leaves in each level.

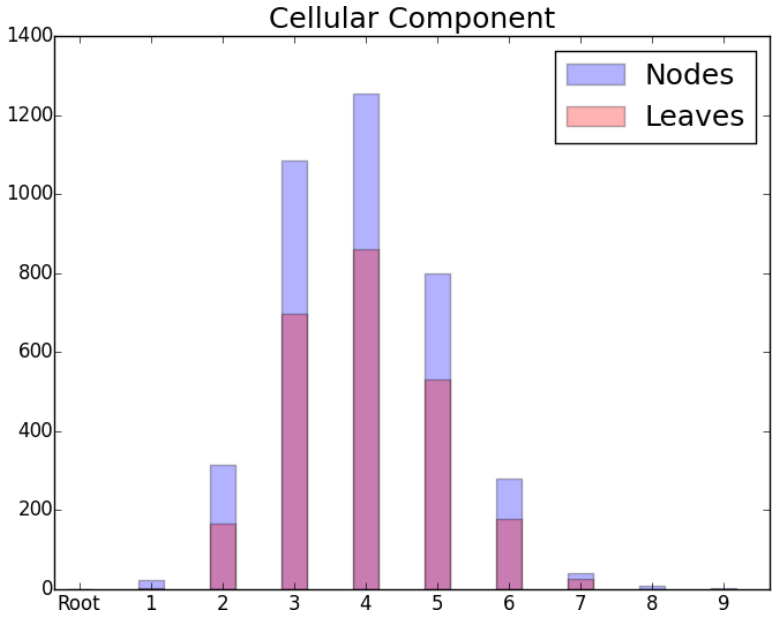

Figure 1.3: Distribution of nodes and leaves per level in the Cellular Component ontology. The blue bars show the number of descriptors per level in the Cellular Component ontology of GO and the red bars the number of leaves in each level.

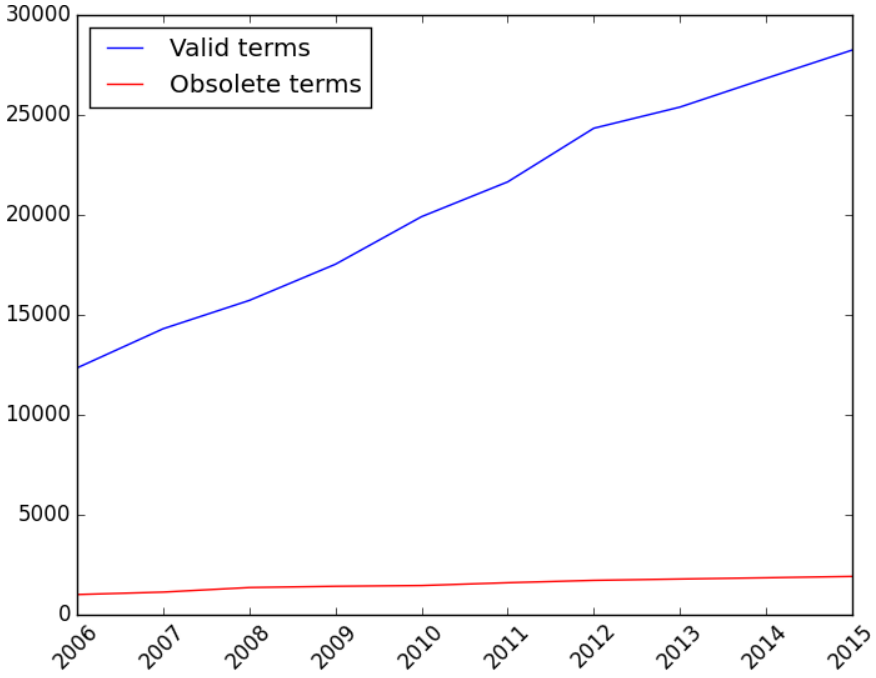

Figure 1.4: Growth of the Biological Process ontology. The blue curve shows the number of terms in the ontology and the red curve the number of obsolete terms.

In an attempt to analyse the refinement of the GO annotations I extracted all experimental annotations (*i.e.* with evidence codes EXP, IDA, IPI, IMP, IGI and IEP [96]) for *A. thaliana*, *H. sapiens*,*M. musculus*, *S. cerevisiae* and *C. elegans* from UniProt GOA [23]. For these organisms, the usage of terms is shown in table 1.2

Organism	Total terms	Mean usage of terms
A. thaliana	4,143	10.94
H. sapiens	8,885	12.62
M. musculus	9,479	18.76
S. cerevisiae	4,449	9.24
C. elegans	2,375	8.87

Table 1.2: Use of GO terms in the experimental annotations of some model organisms. The total terms correspond to the annotations with the experimental evidence codes EXP, IDA, IPI, IMP, IGI and IEP.

In all organisms, the most common term used is *Protein Binding* (`GO:0005515`), a term at level 2 in the Molecular Function ontology, used $2,557$ times in *C. elegans*, $6,572$ in *A. thaliana*, $10,874$ in *S. cerevisiae*, $22,492$ in *M. musculus* and $35,409$ in *H. sapiens*. This term is followed the equally general terms *Nucleus* (`GO:0005634`), *Cytoplasm* (`GO:0005737`) and *Mitochondrion* (`GO:0005739`). This skew in the use of the ontology reflects the issues associated to the overall low reproducibility of experiments [79, 35]: few experiments are generally repeated resulting in many genes with very general annotations.

1.2.3 Disease Ontology (DO)

The Disease Ontology (DO) is a community-driven, manually created resource to provide a uniform nomenclature for human diseases [57]. The centralised repository provided by the DO enables a precise identification of human diseases facilitating the sharing of information as well as large scale computational analysis. The DO is based on the OBO [15] concepts and links to databases such as Medical Subject Headings

(MeSH), ICD+ and OMIM through cross-referencing.

1.2.4 The ontological structure

DO defines a total of 6,590 non-obsolete terms, of which 5,776 (72%) are leaves. The ontological structure is defined exclusively through 6,940 *is_a* relationships and terms have, on average, 1.06 children. As is the case in GO, the *is_a* relationship in the Disease Ontology does not define instances of diseases, but rather reflects a type-subtype relationship. This can be seen in figure 1.5 which shows a branch of the ontology, starting from a leaf at the deepest level of the ontology.

Figure 1.6 shows the number of nodes in each level of the ontology where each bar corresponds to a level in the ontology. The blue part of each bar indicates the number of descriptors in the corresponding level and the red component of the bars the number of leaves. Again, as was the case in GO the majority of the terms in DO are in the mid levels of the ontology. Since the upper levels of the ontology are general (*e.g.* Endocrine system disease), the growth is mostly concentrated in the mid levels.

Older versions of DO are not available for download. I was, therefore, unable to analyse the growth and evolution of the ontology through time.

1.3 International Classification of Diseases (ICD+)

ICD+ is an ontology designed to monitor the incidence and prevalence of diseases. It is designed and maintained by the World Health Organisation, and is arguably, the first ever biological ontology dating as far back as the late 1800's [1]. ICD+ provides a standard vocabulary in order to enable the comparison and analysis of diseases and health issues across WHO member countries.

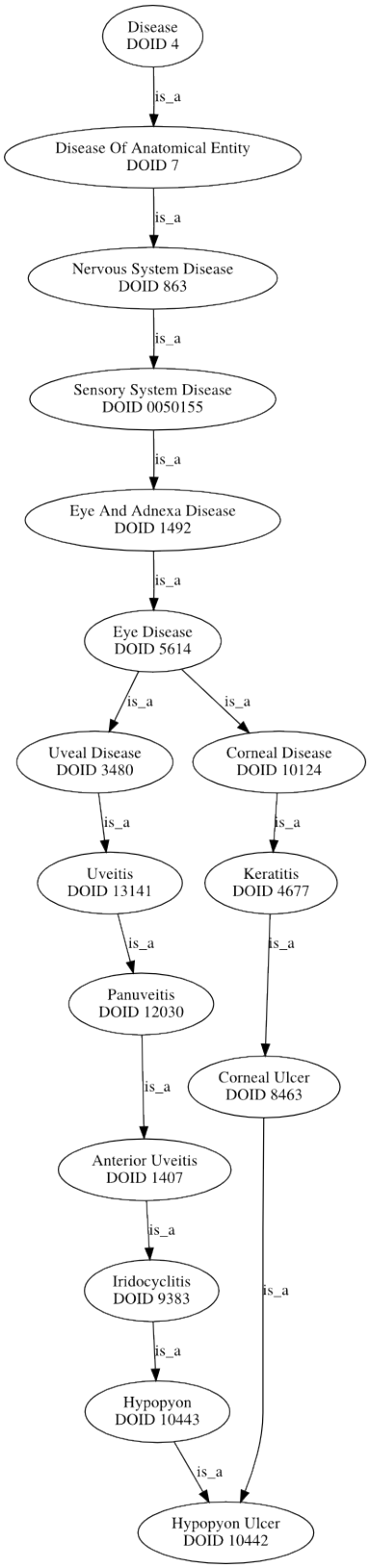

Figure 1.5: Example from the DO. The figure shows all paths to the root from a leaf in the DO.

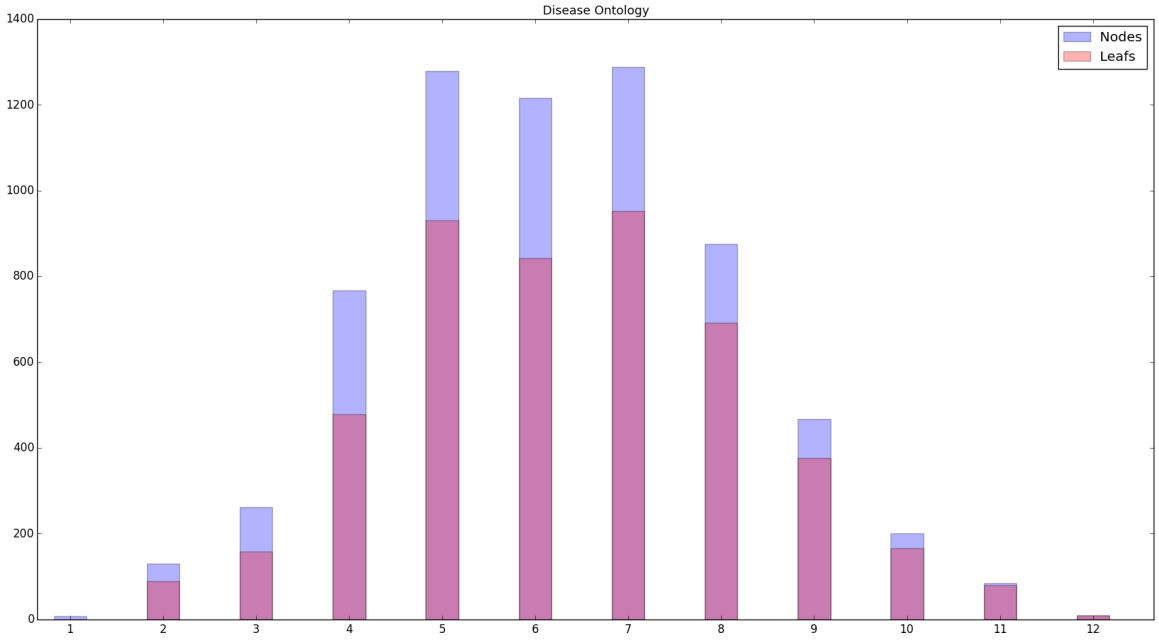

Figure 1.6: *Distribution of nodes and leaves per level in the DO.* The blue bars show the descriptors per level in DO and the red bars the number of leaves in each level.

1.3.1 The ontological structure

There are a total of 12,131 nodes linked through implicit *is_a* relationships. Of these nodes, 10,557 are leaves (87%). ICD+ is, compared to other ontologies shallow having a maximum depth of 3. The mean number of children is 4.92, compared to GOs and DOs (2.09 and 1.06 respectively) which reflects the shallow ontological structure and the large number of terms.

It is important to note, that the aim of ICD+ makes the ontology itself less complex. ICD+ need not specify the different nodes with the amount of specificity that, for example, GO needs. Since the aim of ICD+ is to classify diagnosis, the specific enough needs to ensure that individual terms can be used in groups to identify a diagnosis. While every term in the ICD+ describes a disease, in the DO this is not the case. In DO a disease is described by a collection of terms, and therefore, a more detailed ontology is required.

Each node is assigned a tree number with the following format: `<Root><TwoDigits>{"."<SingleDigit>"."}+` . Every `"."` indicates a new level and the `<TwoDigits>` to the left and to the right of the `"."` indicates a descriptors coordinate in that specific level. For example, the Tree Number (`A00`) (which corresponds to the ICD+ term *Cholera*) has three children (`A00.0`) (*Cholera due to Vibrio Cholerae 01, biovar cholerae*), (`AA0.1`) (*Cholera due to Vibrio Cholerae 01, biovar eltor*) and (`A00.9`) (*Cholera unspecified*) Figure 1.7 shows the full path from a leaf at the deepest level of the ontology to the root.

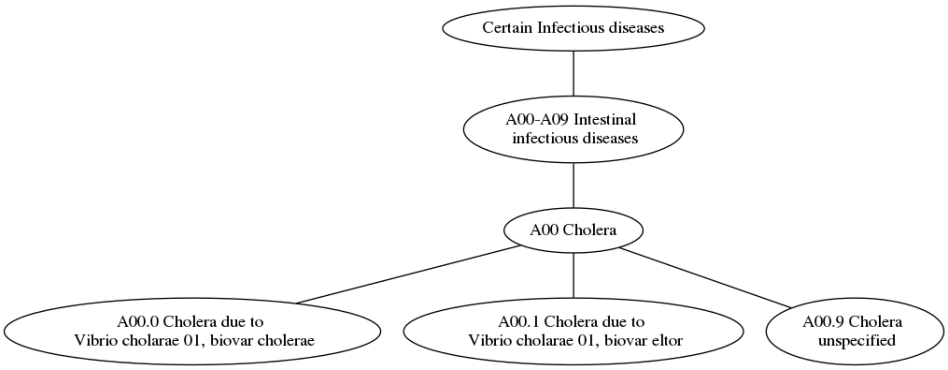

Figure 1.7: Example from ICD+. The figure shows all paths to the root from a leaf in ICD+.

Figure 1.8 shows the number of nodes in each level of the ontology in ICD+. The blue bars show the number descriptors per level in the ontology and the red bars the number of leaves in each level. The ontology is very particular, as the vast majority of leaves (95%) are located in the last level of the ontology. That is, the ontology was build to provide a unique vocabulary to help in the comparison of diagnostics, and is therefore designed to very specific. This is mainly due to the nature of the domain the ontology is designed to describe.

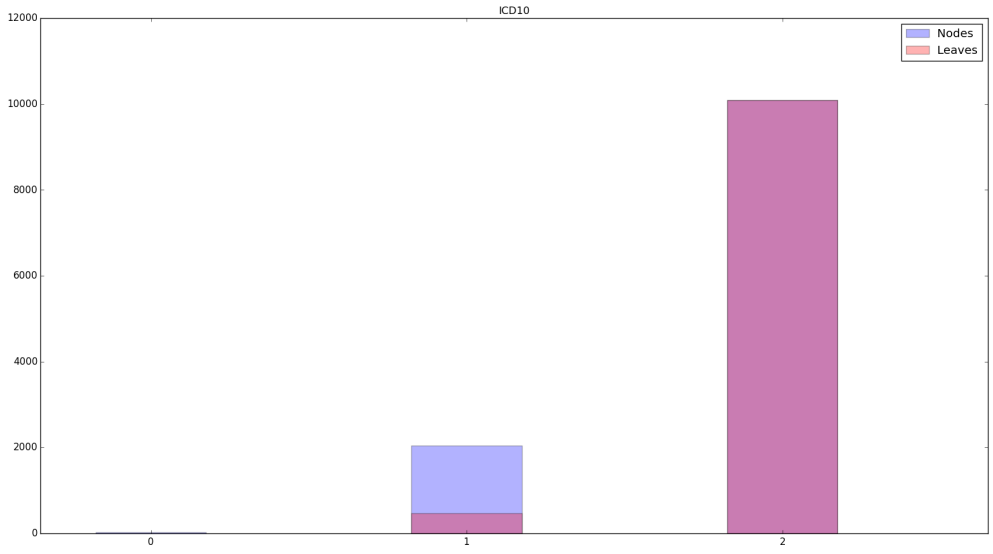

Figure 1.8: Distribution of nodes and leaves per level. The blue bars show the descriptors per level in ICD+ and the red bars the number of leaves in each level.

1.4 Human Phenotype Ontology (HPO)

The Human Phenotype Ontology (HPO) focuses on the phenotypic abnormalities, based on the Open Biological Ontologies [15] concepts. In the Human Phenotype Ontology (HPO) each term describes a single disease phenotype [81]. The construction of the HPO differs from that of GO and DO in that it is a result of text-mining of OMIM to produce the relevant disease phenotypes [81]. The complex construction procedure requires further detail.

To obtain the terms that will be used to construct the ontology, the authors extracted all available Clinical Synopsis (CS) descriptions from the diseases in OMIM [4]. These CS fields contain a list of known phenotypes associated to the diseases. All phenotypes appearing more than once in OMIM will be included in the HPO. For example, the phenotype *Dementia* is associated to more than a hundred diseases in OMIM and results in the term (`HP:0000726`), while *Presenile and senile dementia* is

exclusively associated to *Alzheimer's Disease* (`MIM:104300`) and therefore does not have a term in the ontology. Every putative term obtained through this process of text mining of the CS fields was manually curated, exploiting the authors' expertise in human genetics [81]. The list of terms was further expanded by performing string matching analysis of terms already in the ontology and those initially discarded. The relationships between the terms were manually constructed [81].

1.4.1 The ontological structure

The ontological structure is defined through *is_a* relationships. A total of $11,324$ nodes, of which $7,290$ (64,3%) are leafs, are connected through $11,423$ links. Figure 1.9 shows a specific branch built from a leaf in the HPO.

Figure 1.10 shows the number of nodes in each level of the ontology. The blue bars show the descriptors per level in the ontology and the red bars the number of leaves in each level.

Over the course of the 31 releases, the ontology maintained the 13 levels from 2012 through to 2014. I explored the releases available from 2012 to 2014 to compare the growth to the maintenance and removal of obsolete terms. Figure 1.11 shows the number of terms in each yearly release of the HPO compared to the number of obsolete terms in that release. The constant number of obsolete terms and the little increase (in some cases even decrease) in number of terms compared to GO reflects the construction procedure of the HPO: fewer terms need to be deprecated as most terms were already manually extracted from a comprehensive database such as OMIM.

1.5 Medical Subject Headings (MeSH)

MeSH is a controlled vocabulary created and maintained by the National Library of Medicine [98] in the United States of America. It is organised into several hierarchical

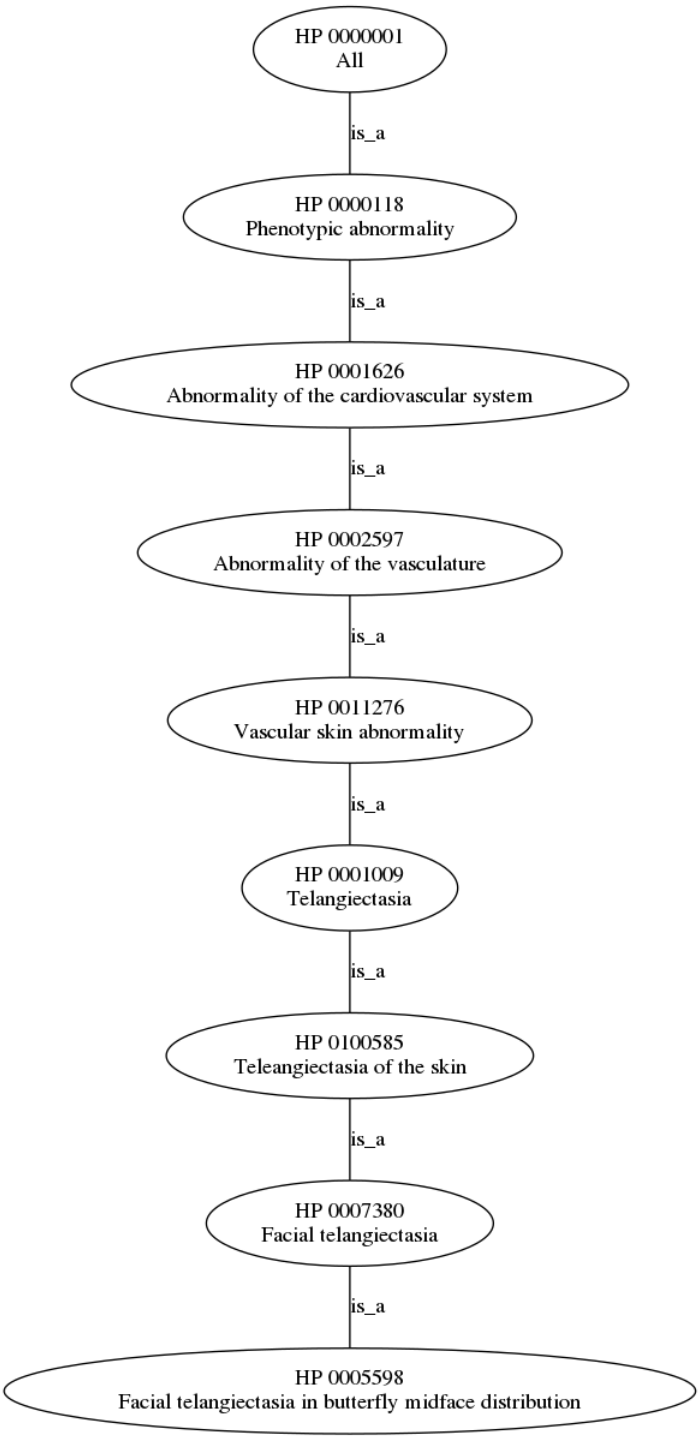

Figure 1.9: Example from the HPO. The figure shows all paths to the root from a leaf in the HPO.

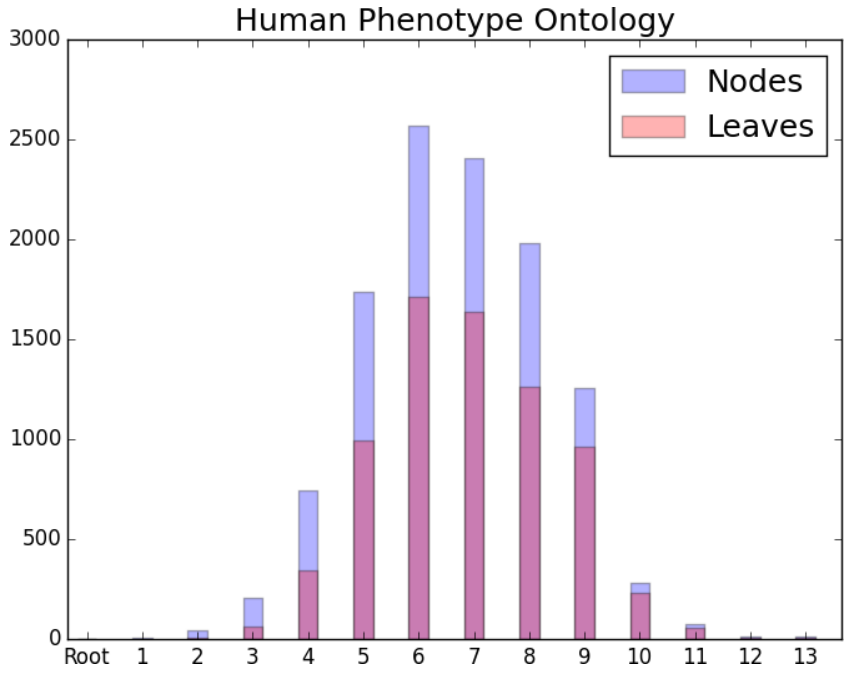

Figure 1.10: Distribution of nodes and leaves per level. The blue bars show the descriptors per level in the HPO and the red bars the number of leaves in each level.

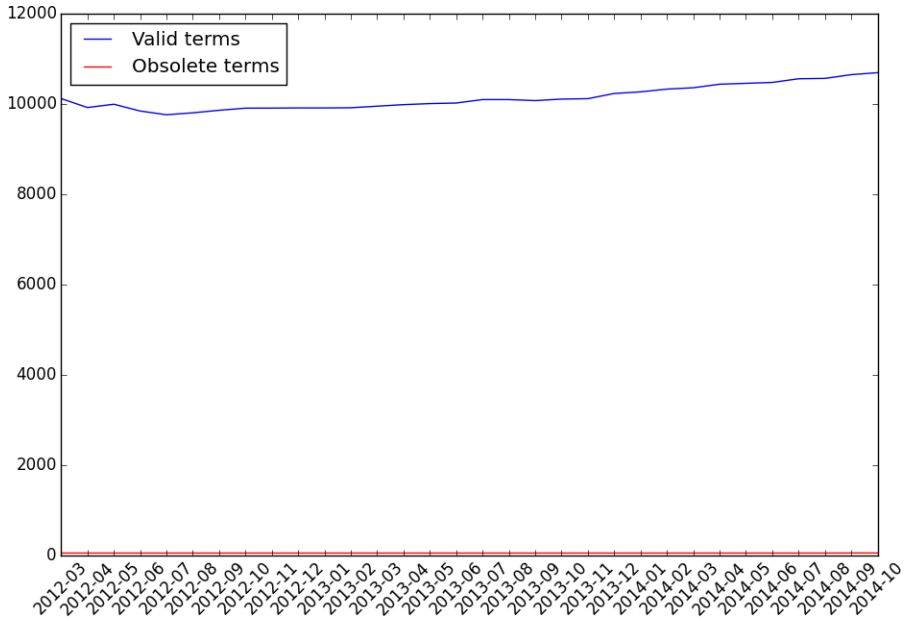

Figure 1.11: *Growth of the Human Phenotype Ontology.* The blue curve shows the number of terms in the ontology and the red curve the number of obsolete terms.

structures that allow searching and indexing of publications in the PubMed/MEDLINE library at different levels of specificity. Usually MeSH is identified with the descriptors [64, 106], however, there are two additional record types: Qualifiers and Supplementary Concept Records.

Descriptors

The descriptors are organised into 16 individual but overlapping ontologies, each of which represents a specific subject area: Anatomy [A], Organisms [B], Diseases [C], Chemicals and Drugs [D], Analytical, Diagnostic and Therapeutic Techniques and Equipment [E], Psychiatry and Psychology [F], Phenomena and Processes [G], Disciplines and Occupations [H], Anthropology, Education, Sociology and Social Phenomena [I], Technology, Industry, Agriculture [J], Humanities [K], Information Science [L], Named Groups [M], Health Care [N],Publication Characteristics [V] and Geographicals [Z]. These categories are not meant to be an exhaustive classification of the subject they represent, but rather a hierarchy of terms needed for the classification and indexing of the publications in PubMed. It is important to note that the descriptors in the Publication Characteristics [V] and Geographicals [Z] do not describe the *content* of the publications, but rather the *publications* themselves.

Subheadings or Qualifier

The Subheadings or Qualifier are records used for indexing and cataloguing PubMed entries alongside the Descriptors. They are organised into a smaller taxonomy of 88 descriptors. Effectively, the subheadings group together descriptors into a coherent topic related to the publication [98].

Supplementary Concept Records

The Supplementary Concept Records are used to index chemical compounds, drugs and other concepts. These records are not structured in an ontology, but are linked to one or more descriptors.

1.5.1 The ontological structure

The ontological structure of the descriptors is defined by the tree number associated to the descriptor. While the relationships are not explicitly specified in MeSH, I can identify two types: *is_a* relationships and *part_of* relationships. Each node is assigned a tree numbers have the following format: `<OntologyName> <TwoDigits> {"."<ThreeDigits>"."}+` . Every `"."` indicates a new level and the `<ThreeDigits>` to the left and to the right of the `"."` indicates a descriptor's coordinate in that specific level. For example, the Tree Number `G01.595.560.107` (which corresponds to the MeSH term *Acceleration* (`D000054`)) is shown in figure 1.12 Notice the recursive construction of the ontology, where all nodes at each level, share the prefix corresponding to their depth. For example, the nodes at level 3 Coriolis Force, Rotation and Acceleration have identical tree numbers up to the 3rd position, namely, `G01.595.560`.

Table 1.3 shows a few topological features of the MeSH ontologies.

In figure 1.13 I shows the number of nodes in each level of the ontology in MeSH. The blue bars show the descriptors per level in each MeSH ontology and the red bars the number of leaves in each level. The ontologies are all "wider" in the middle, that is, the majority of terms are in the middle levels of the ontology. It is important to note that in the case of MeSH the high-level leaves do not necessarily indicate a poorly constructed ontologies, but rather a well constructed an understood underlying taxonomy. This can be seen in an example from the Anatomy [A] ontology, in which a node at level 3 (from the 11 possible) can already represent a highly specific concept,

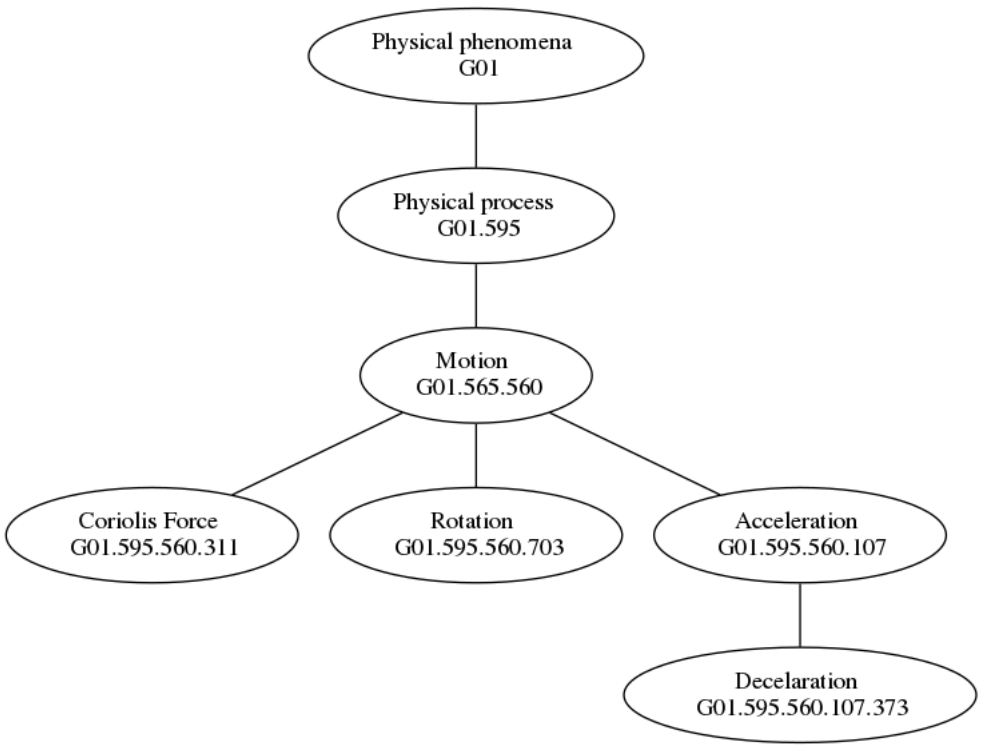

Figure 1.12: Example from MeSH. A small subset extracted from MeSH.

for example, *Adrenergic Neurons* (`D059331`).

I analysed the evolution of the MeSH ontologies by comparing the 2013 and 2014 releases. In figure 1.14 the red bars show the distribution of descriptors per level in 2013 MeSH and the blue bars the distribution of descriptor per level in the 2014 MeSH. The most noticeable change occurred in the Humanities [K] ontology, where a new level of specific nodes was added.

1.6 Semantic similarity of terms in an ontology

Semantic similarity measures are metrics to quantify similarity between objects in a given context [72]. This context is provided by well-structured controlled vocabularies

Ontology	Descriptors	Leaves	Max.depth	Avg.children
Anatomy [A]	1,703	768	11	3.08
Organisms [B]	3,670	894	12	3.53
Diseases [C]	4,621	1,271	10	3.40
Chemicals and Drugs [D]	9,280	1.450	11	3.41
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	2,727	887	10	3.83
Psychiatry and Psychology [F]	965	487	7	3.41
Phenomena and Processes [G]	1,978	748	10	3.73
Disciplines and Occupations [H]	388	280	7	3.52
Anthropology, Education, Sociology and Social Phenomena [I]	561	322	9	3.79
Technology, Industry, Agriculture [J]	513	218	10	4.09
Humanities [K]	191	145	7	3.17
Information Science [L]	415	213	9	3.89
Named Groups [M]	225	160	7	3.05
Health Care [N]	1,597	540	9	3.55
Publication Characteristics [V]	155	112	4	5.41
Geographicals [Z]	392	340	7	5.81

Table 1.3: The 16 MeSH ontologies. The number of terms in each ontology corresponds to the 2014 version of MeSH.

in the form of ontologies. For example, when trying to determine functional similarity between genes, the GO Biological Process ontology can provide the appropriate context [42].

Considering that GO [59] is perhaps the best known biomedical ontology, the following discussion will focus on GO without loss of generality.

1.6.1 The True Path Rule

The ontological structure allows the definition of a *True-Path*. This True-Path means that the entire path from a term to the root of the ontology must be true an consistent [42], and therefore, these ancestor terms must be valid annotations for the gene (in

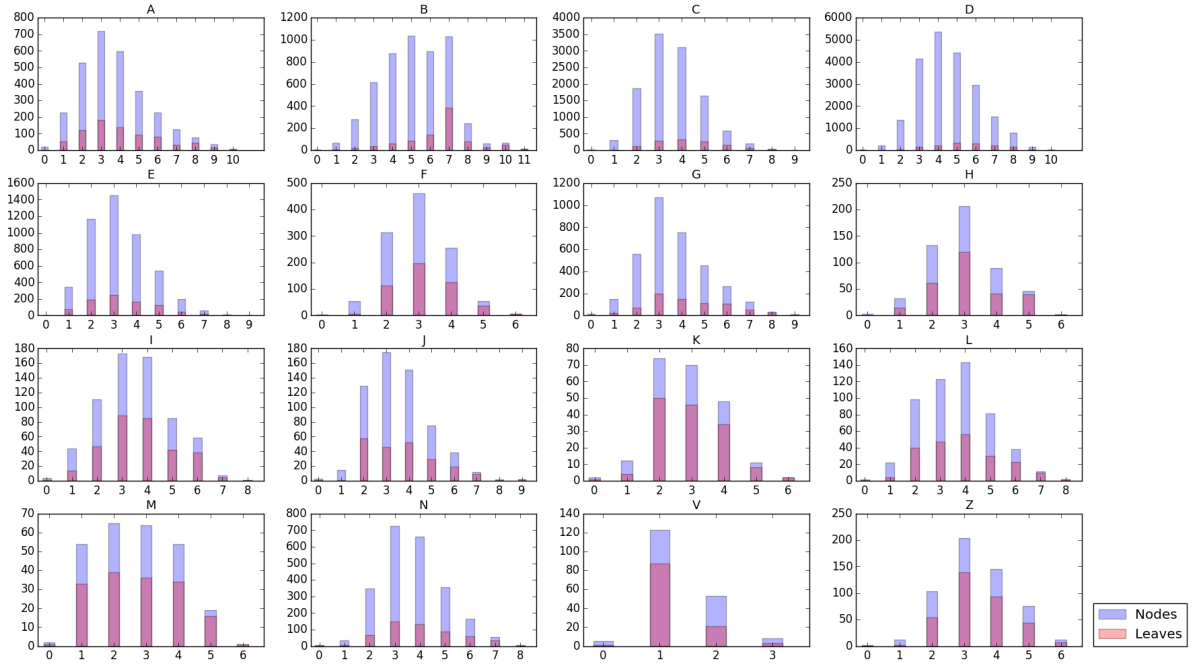

Figure 1.13: Distribution of nodes and leaves per level for each of the 16 MeSH ontologies. The x-axis shows the different levels in each ontology and the y-axis the number of descriptors in that level. The blue bars show the descriptors per level in MeSH and the red bars the number of leaves in each level.

the case of GO). Consider the example shown in 1.15, in this figure, a gene annotated with the function *Cell growth function*, is also a *single-organism cellular process*.

The True-Path rule results in an annotated Directed Acyclic Graph (DAG) of the ontology composed of all the terms that are on the path of to the root of the originally annotating terms. Figure 1.16 shows an example from annotations belonging to the *BRCA2* gene in Human. The red nodes indicate the direct annotations, and the white nodes the annotated subgraph constructed following the *is_a* relationship. In the context of GO, the terms in figure 1.16 are the only ones that are valid descriptions of the gene.

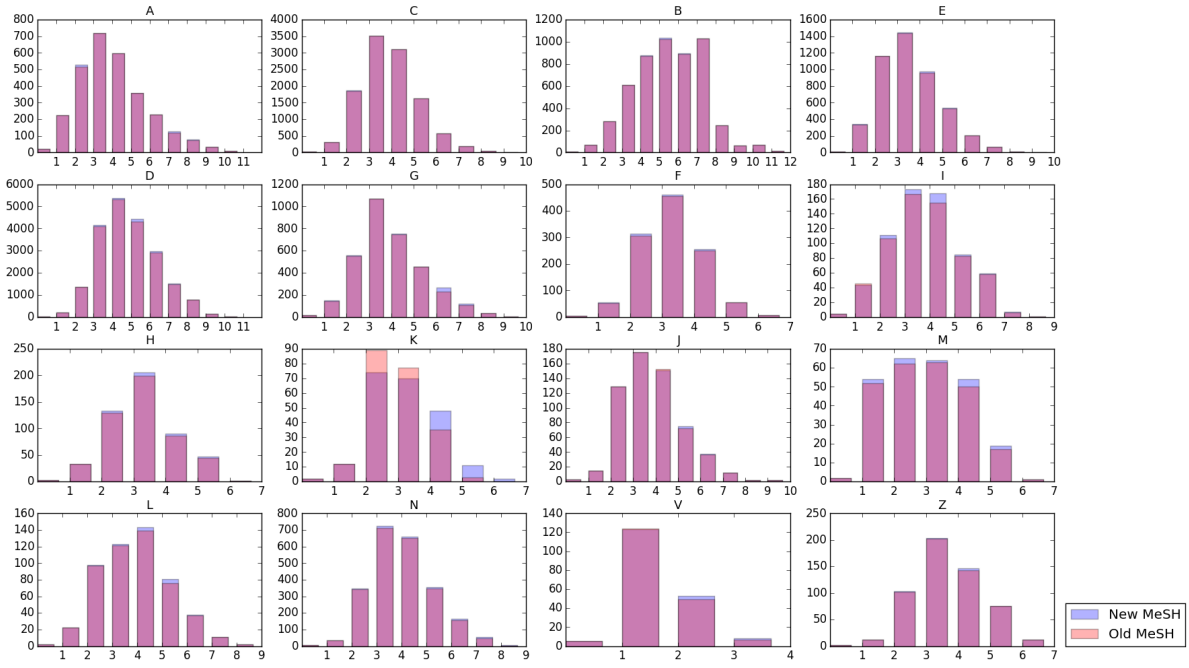

Figure 1.14: *Perspective on the evolution of MeSH.* The x-axis shows the different levels in each ontology and the y-axis the number of descriptors in that level. The red bars show the descriptor per level in 2013 MeSH, the blue bars the descriptors per level in 2014 MeSH.

1.6.2 Topological similarity measures

Based on these annotated DAG structures resulting from the True Path Rule, "simple" topological measures are possible. Consider the topological distance between the ontology nodes, where similarity between two nodes is inversely proportional to the length of the shortest path available [72]. This measure fails to account for the variable conceptual distance represented by the different links, as well as for the variability in detail of the various ontologies [72]. For example, consider the terms and relations *Regulation of cell morphogenesis* (`GO:0022604`) $\overrightarrow{is\_a}$ *Regulation of anatomical structure morphogenesis* (`GO:0022603`) and *Cellular Process* (`GO:0009987`) $\overrightarrow{is\_a}$ *Biological Process* (`GO:0008150`) are joined by an identical link, the conceptual distance is not

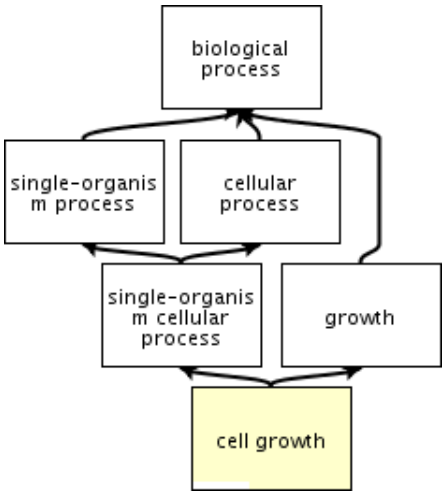

Figure 1.15: All genes with the *Cell growth function* have also the more general function *Growth*, and so on until the root. Image produced by QuickGo [78]

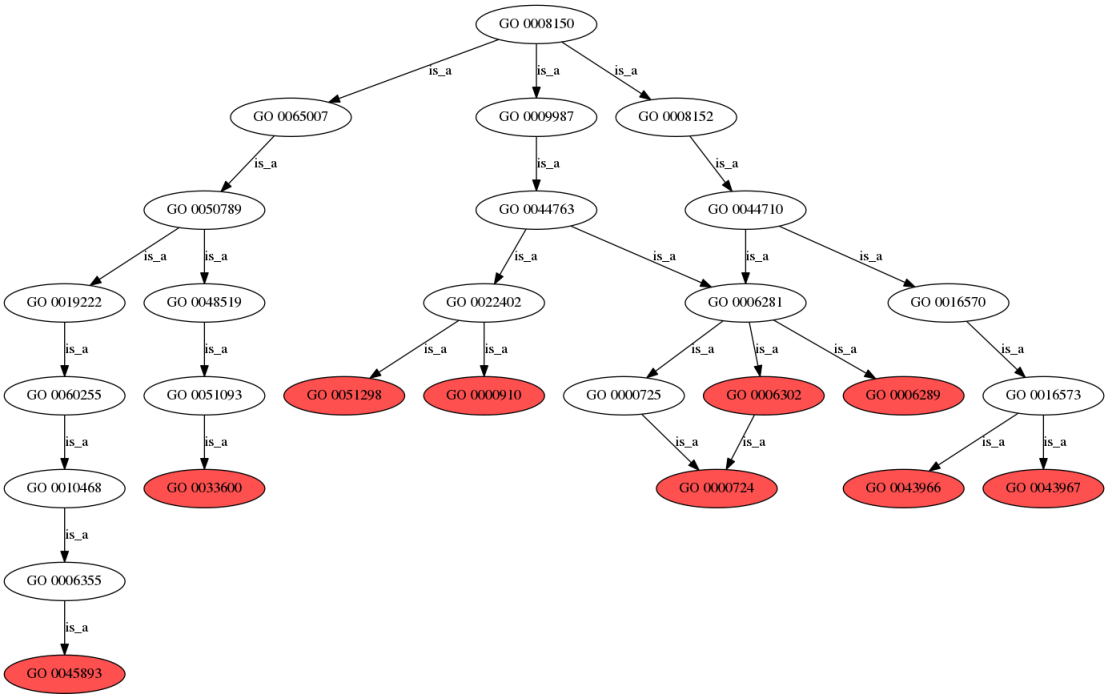

Figure 1.16: The red nodes correspond to all experimental annotations of the *BRCA2* gene in Human.

identical. While, as these simple topological measures show, the structure of the ontology is not strictly necessary to compute similarity between genes. However, as I prove through an in-depth analysis based on MeSH in chapter 5, its appropriate use results in more accurate similarity scores.

1.6.3 Ontological similarity measures

Several semantic similarity measures have been proposed to overcome the shortcomings of the simpler, topological similarity measures. The ones by Resnik [72], Jiang and Conrath [51] and Lin [25] are based on the information content of the lowest common ancestor of pairs of terms, and are often referred to as "term-based", due to the fact that they compute the similarity between individual terms [42]. Alternatively, simUI and simGIC [17] compare sets of terms rather than single terms using graph comparison approaches and are often referred to as "graph-based" [42]. These measures were chosen as a representative set, it is important to note however, that a few new measures have appeared [82, 34, 43] recently.

Term-wise measures

The term-based measures determine similarity of terms a and b based on the concept of information content of a term, defined as the negative logarithm of the probability of that term (calculated as the ratio between the number of diseases annotated by that term and the total number of annotated diseases) [72, 42]. Formally:

$$IC(c) = -log\left(p\left(c\right)\right) \tag{1.1}$$

Resnik[72] determines the similarity of two nodes in an ontology by calculating the information content of the Lowest Common Ancestor (LCA) of two given nodes. The Lowest Common Ancestor (LCA) is defined as the common ancestor of nodes a

and b with the highest information content. Formally, Resnik's measure is defined as:

$$sim\,(a,b) = -2 * IC\,(LCA\,(a,b)) \tag{1.2}$$

Further measures, which can be interpreted as normalisations of Resnik's measure [42] are proposed by Lin [25] and Jiang and Conrath [51]. Lin's measure is define as:

$$sim\,(a,b) = \frac{-2 * IC\,(LCA\,(a,b))}{IC\,(a) + IC\,(b)} \tag{1.3}$$

Lastly, Jiang and Conrath, define a distance measure, which can be converted to a similarity score. Formally, the measure is defined as follows:

$$dist\,(a,b) = 2 * IC\,(LCA\,(a,b)) - IC\,(a) - IC\,(b) \tag{1.4}$$

The conversion of this distance measures is straightforward.

$$sim\,(a,b) = f\left(1 - \frac{dist\,(a,b)}{max\,(dist\,(i,j))}\right) \forall\,(i,j) \tag{1.5}$$

While Resnik, Lin and Jiang and Conrath perform similarly on the GO [42], this is not the case for MeSH. An in-depth analysis is presented in chapter 5.

Since genes are annotated by sets of terms, the combination of the multiple scores into a single number that quantifies the similarity is required. To illustrate this, consider two genes a and b, annotated as follows: $G_a = \{t_1, t_2, t_4, t_6\}$ and $G_b = \{t_1, t_7\}$. For each pair of terms $\{(t_1,t_2),(t_1,t_7),(t_2,t_7),(t_4,t_1),(t_4,t_7),(t_6,t_1),(t_6,t_7)\}$, the similarity will be given by its LCA. Thus, genes a and b have up to 8 possible similarity scores based on their annotations. Several combinations are available at this point, among which are the arithmetic average of the scores, the maximum possible similarity between all pairs of terms [17], and weighted averages [33] of the scores. Whichever the choice, some compromise will have to be made. Choosing

the maximum possible similarity disregards the differences between the terms and choosing the arithmetic average disregards the similarities between the terms. To avoid the intrinsic problems of the term-based measures, Graph-based measures have been proposed.

Graph-based measures

Graph-based measures determine the similarity of the genes based on sets of terms rather than the individual terms. The fundamental graph-based measure, simUI [17] is based on the Jaccard coefficient of the annotated ontology, as defined in equation 1.6:

$$sim\left(i,j\right)=\frac{\left\|Terms\left(i\right)\cap Terms\left(j\right)\right\|}{\left\|Terms\left(i\right)\cup Terms\left(j\right)\right\|} \tag{1.6}$$

While this measure includes all terms annotating the genes, it fails to account for their specificity in light of the existing annotations.

To correctly account for the specificity of terms, Pesquita *et al.* [16] proposed simGIC, a modification of simUI. simGIC defines similarity as the quotient of the sum of the information content (IC) of the common terms between two genes and the sum of the information content of all terms used to annotate the genes. Formally:

$$sim\left(i,j\right)=\frac{\sum_{t\in Terms(i)\cap Terms(j)}IC\left(t\right)}{\sum_{t\in Terms(i)\cup Terms(j)}IC\left(t\right)} \tag{1.7}$$

simGIC has been shown to perform better than simUI [17], and these measures were overall shown to be better than the term-wise measures in GO [17, 42].

1.6.4 Zero similarity between sets of terms

A situation that needs to be considered arises when calculating semantic similarities in an ontology with term-based similarity measures. The following analysis will focus

on Resnik's [72] similarity measure as a representative of the term-based similarity measures.

If the LCA happens to be the root node, which according to the true-path rule annotates every gene in the organism, the similarity between the two genes will be equal to zero (as the information content of the root is zero). Figure 1.17 presents a toy example, in which only four nodes of the Gene Ontology annotate any genes, namely, the ones coloured blue and red. The coloured lines, blue and red respectively, represent the True-Path from each node of the corresponding colour.

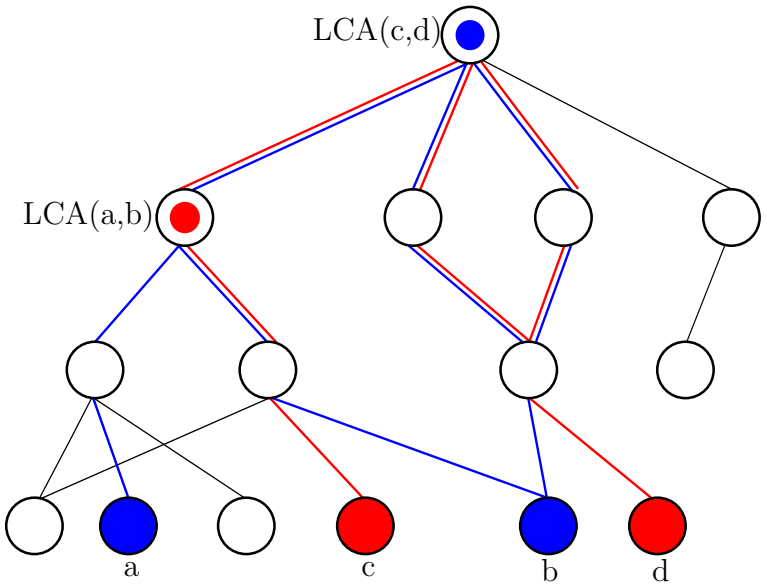

Figure 1.17: *Toy example of a set of terms with zero similarity.* A small set of terms exemplifies situations in which the similarity of two terms can be zero.

By following the blue lines the LCA for nodes labelled a and b is reached. The node labelled $LCA\,(a,b)$ will not annotate all the genes. In fact, the genes annotated by the red node d are not annotated by the node $LCA\,(a,b)$. To verify this suffices with following red line from node d to the root: this path does not include node $LCA\,(a,b)$ and therefore, the fraction of genes annotated by it will definitely be different than 1 leading to non-zero information content.

Conversely, by following the red lines the LCA of nodes c and d is reached. This node, represented by the node with the label $LCA\,(c,d)$, happens to be the root of this toy ontology. The fact that this node is the root means that, due to the true path rule, it annotates all genes (a,b,c,d). This means the information content of node $LCA\,(c,d)$ will be zero. While this example might seem contrived, this particular situation arises when considering that most organisms available in UniProt GOA have very few experimental annotations.

Lin's measure will also be zero whenever the LCA coincides with the root of the ontology, according to 1.3. Jiang and Conrath's measure requires further analysis. The distance measure in 1.4 will be properly defined, resulting in low distance, as given by the sum of the information content of the nodes being analysed. In cases where the LCA is the root, the similarity of two terms will be dependent exclusively on their own information content, as defined in equation 1.5. We can, therefore, understand the factor $2*(LCA\,(a,b))$ as being a damping factor, that will reduce the similarity according to the distance of the pair of terms to their LCA.

Consider a basic example, were the number of genes annotated by the root is 10, the number of genes annotated by nodes a and b are 2 and 3 respectively. If the root is chosen the similarity will be defined as follows by the various measures:

- Resnik: $-2*log\,(p\,(1))=0$
- Lin: $\frac{-2*log(p(1))}{log(0.2)+(0.3)}=0$
- Jiang: $dist\,(a,b)=2*log\,(p\,(1))-log\,(0.2)-log\,(0.3)=1,21.$

  To convert the distance into a similarity, the maximum value in the distance matrix has to be determined. For the purposes of this exercise, let the maximum value be $1,13$, corresponding to an LCA annotating 9 terms in the ontology. $sim\,(a,b)=f\left(1-\frac{1,22}{1,13}\right)=1.08$

A graphical illustration is shown in figure 1.18.

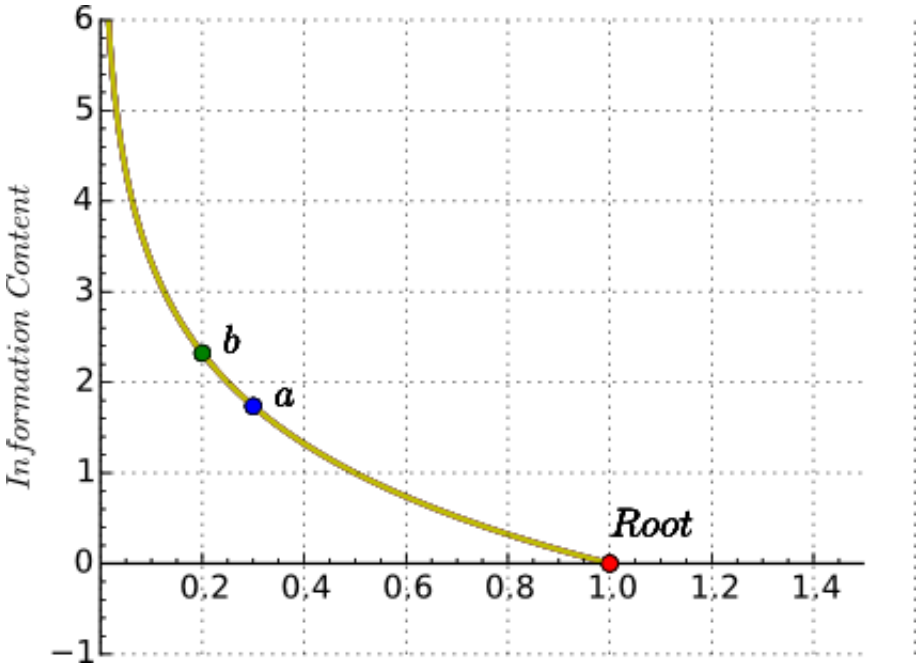

Figure 1.18: Zero information content. The red circle labelled Root corresponds to the information content of the root of GO. The green and blue circles labelled a and b, respectively, correspond to the information content of the terms a and b in GO.

1.7 Improving GO semantic similarities

An important recent development has been the introduction of the Random Walk Contribution by Yang *et al.* , developed in our Lab, which greatly improves the semantic similarity measures presented in § 1.6 [42].

The authors note that the GO annotations are constantly changing, evolving to reflect the new knowledge that becomes available. This change introduces an inherent uncertainty that has, so far, not been considered appropriately. In addition existing similarity measures disregard the ontological structure that spans below the terms that are being compared. Yang *et al.* [42] argue that the uncertainty in the annotations and the structure below the terms has an impact on the semantic similarity measures, and therefore need to be appropriately considered.

To illustrate the relevance of the ontology below the terms, consider figure 1.19 (redrawn from [42]) where, without loss of generality, all leaf terms annotate non-overlapping proteins. The pairs of nodes (A, B) and (C, D) have an identical structure above. However, since the terms (C, D) share the child Z, they would ideally be more similar than terms (A, B), a fact that traditional similarity measures would ignore.

Consider once again figure 1.19, but this time, note how the proteins annotated by each term, shown in parentheses, are only completely defined for the term A. Assuming that leaf nodes annotate non-overlapping genes, all other annotations could be specified further.

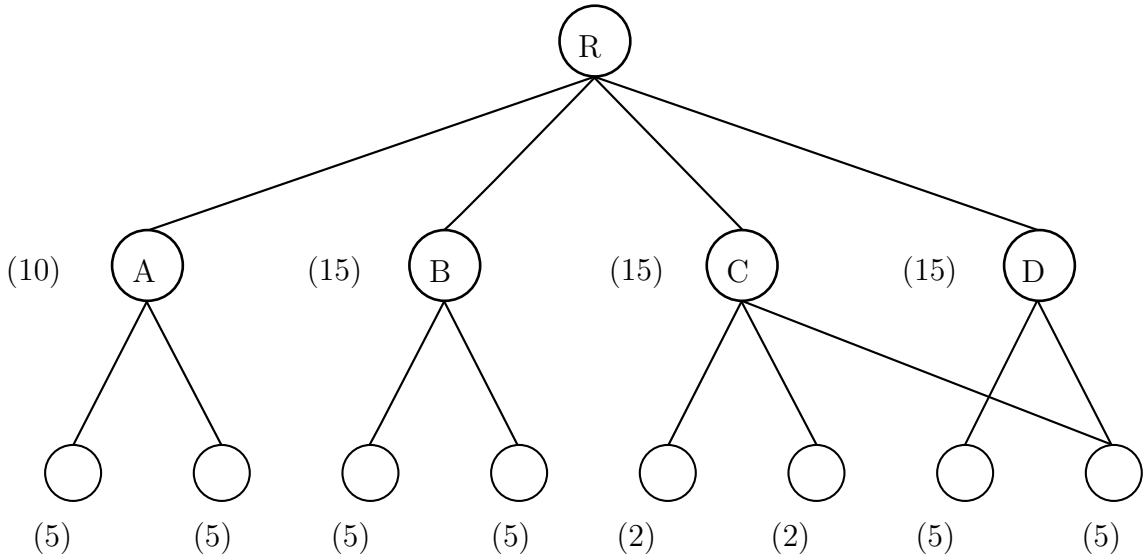

Figure 1.19: *The relevance of the ontology below the terms.* This figure was reproduced from [42]. The ontology above A, B and C, D is identical, however, terms C, D share a child. *The uncertainty in the annotations.* The annotations in node A are fully specified, while the annotations of term B can still be specified further.

The authors propose the Random Walk Contribution as an "add-on" for existing similarity measures (the Host Similarity Measure (HSM)) in order to extend them and appropriately handle the uncertainty in annotation as well as the ontological structure below the terms.

1.7.1 The method

According to the true-path rule, every protein annotated by a term in an ontology is also annotated by all its ancestors [42]. This means that every protein annotated by term F and by term G in the toy example presented in figure 1.20 (redrawn from [42] are also annotated by term D. The similarity of term D and some other term X could therefore be estimated by the pairwise similarities of G, X and F, X [42]. Viewing similarity from this perspective is relevant considering that these are leaf terms and therefore the HSM's are accurate for terms F and G [42]. To estimate the weights for the similarities F, X and G, X the authors define a downward random walk which will begin at the root of the ontology [42].

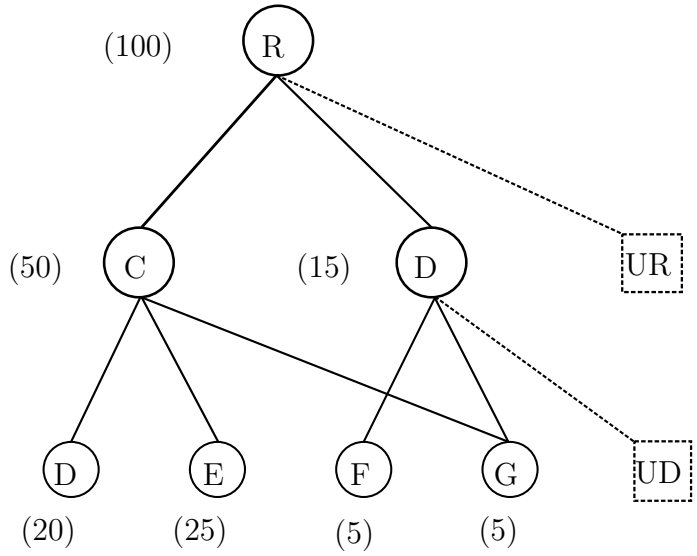

Figure 1.20: *Illustration of the ISM method.* The figure shows the inclusion of a fictitious node to account for the uncertainty in the annotations.

To define the transition probabilities, the authors account for the uncertainty in the annotations. Consider the case of node D in figure 1.20, in which there are proteins that could be specified further, possibly being annotated in the future by a node that does not currently exists (labelled U_D in the figure). The transition probability to

node U_D is defined by equation 1.8, where N_v^* is the number of proteins annotated by term v and not to any of is children, N_v is the total number of proteins annotated by node v.

$$P\left(v \to U_v\right) = \frac{N_v^*}{N_v} \tag{1.8}$$

Once the probability of the partially annotated proteins being annotated by a currently nonexistent node is determined, the remaining transition probabilities can be defined by equation 1.9.

$$P\left(v \to c\right) = \left(1 - P\left(v \to U_v\right)\right) \frac{N_c}{\sum_{u:\exists v \to u} N_u} \tag{1.9}$$

The random walk follows the transition probabilities defined for the nodes by equations 1.8 and 1.9. W_t^v denotes the probability of the walker being in node v at time t, with $W_0^R = 1$, where R is the root, and zero for all other nodes at time $t = 0$. The probability of being in any node at time t is determined by the probabilities at $t-1$ and the transition probabilities. However, the probabilities are different for leaf nodes and non-leaf nodes. Unlike the non-leaf nodes, once the walker has arrived at a leaf, it cannot leave. The probability of being on a leaf node l at time $t+1$ is equal to the probability of arriving at this node from a parent, denoted by $v \to l$ plus the probability of being there at time t, and is given by equation 1.10 [42].

$$W_{t+1}^{v_0}\left(l\right) = W_t^{v_0}\left(l\right) + \sum_{v:\exists v \to l} W_t^{v_0}\left(v\right) P\left(v \to l\right) \tag{1.10}$$

Similarly, for all non-leaf nodes v, the probability of being at node v at time $t+1$ is given by the probability of arriving to node v from one of its parents. W_{t+1}^v is defined by equation 1.11

$$W_{t+1}^{v_0}\left(v\right) = \sum_{q:\exists q \to v} W_t^{v_0}\left(q\right) P\left(q \to v\right) \tag{1.11}$$

The Random Walk Contribution to the similarity of nodes v_0 and v_1 is given by the weighed HSM of all leaf node L, excluding the added fictitious nodes. Formally:

$$RWC\left(v_0, v_1\right) = \sum_{i,j \in L} W_{\infty}^{v_0}\left(i\right) W_{\infty}^{v_1}\left(j\right) HSM\left(i,j\right) \tag{1.12}$$

This Random Walk Contribution is then combined into an Integrated Similarity Measure (ISM) according to equation 1.13

$$ISM\left(v_0, v_1\right) = RWC\left(\left(v_0, v_1\right)\right) \tag{1.13}$$

As defined, the Random Walk Contribution takes into account the ontological structure beneath the terms, as well as the uncertainty in the structure and the annotations. The transition probabilities encode the uncertainty in the annotations, since for nodes with higher uncertainty (see node D), the likelihood of the walker stopping in one of the leaves (see nodes F or G) is reduced [42]. Sharing of descendants is also accounted for by the Random Walk Contribution, as more shared descendants implies a more similar stationary distribution [42].

The method was tested measuring the predictive power of the improved semantic similarity measures on protein-protein interaction data, sequence similarity data and gene expression data [42]. The authors show that the Random Walk Contribution has consistently improve the traditional similarity measures[42].

1.8 The Gene Ontology Semantic Similarity tool: an integrated tool for computing semantic similarities

Together with Dr. Alfonso E. Romero and Samuel Heron, I have developed Gene Ontology Semantic Similarity Tool (GOssTo) [39], the Gene Ontology semantic similarity Tool, a user-friendly web and standalone tool for calculating semantic similarities between gene products according to the Gene Ontology. GOssTo is bundled with six semantic similarity measures, including both term- and graph-based measures, and has extension capabilities to allow the user to add new similarities. GOssTo also implements Yang's *et al.* the Random Walk, extending all the implemented semantic similarities.

The standalone version of the software is developed with ease of use in mind. It is very fast, allowing the calculation of genome-wide semantic similarities on a genomic scale in a few minutes. The web interface of GOssToWeb provides access to all capabilities of GOssTo through a clean and simple web interface. Chapter 6 presents GOssTo and GOssToWeb in full detail.

Chapter 2

Network Medicine: a network view of diseases

The genotype-phenotype relationship is not a linear one [11, 66] with environmental factors, variable penetrance, variable expressivity and other complex phenomena obscuring the real link between a specific phenotype and the underlying genotype [65, 48]. Over the past decades, linkage analysis, Genome-wide Association Studies (GWAS) and, more recently linkage analysis coupled with whole-genome sequencing (henceforth *association studies*) have produced large amounts of genotype-phenotype associations [66, 48]. While these methods have proven to be successful [22] they do not provide a complete picture of the nature of the phenotype-genotype relationship, and have been found insufficient to account for the wide variability in phenotypes [22, 10]. Elucidating the underlying mechanisms driving human diseases requires a broader analysis [22, 10]. The main assumption of Systems Biology is that genes do not act alone, but rather as parts of larger, complex mechanisms [12, 65]. Faults in these complex mechanisms, that is *perturbations* in the biological networks, result in diseases [66, 54].

This chapter is organised as follows. First, an introduction of the relevant biological networks will be presented followed by a discussion on disease modules. Then, the problem of quantifying disease similarities will be introduced.

2.1 The biological networks

The observed behaviour of the complex system that is the cell is the result of wider interaction and interdependent activity of the biological molecules [12]. This understanding of the modular nature of cellular organisation [56] guides the development of the biological networks. These network abstractions are commonly depicted as graphs where the nodes represent physical entities such as genes and proteins, and the edges represent a variety of interactions (*e.g.* physical or functional) [66]. In general, the construction of the networks follows three approaches, namely high-throughput experiments, literature curation of existing data and computational predictions [66].

For the purposes of this work, Protein-protein interaction (PPI) networks are the most relevant and will therefore be analysed in more detail. Table 2.1 presents a brief breakdown on the most relevant biological networks [66, 105].

2.2 Protein-protein interaction networks

Proteins rarely perform their tasks in isolation [66] but rather as part of functional modules called complexes [49, 56]. Within these complexes, each protein has a specific function that contributes to the overall function of the module [91]. A PPI network represents all known physical interactions between the proteins of an organism by means of an undirected graph. Many approaches have been developed for experimentally discovering the physical interactions between proteins. Yeast two-hybrid (Y2H) and Affinity purification/Mass spectrometry (AP/MS) are the most widely used [49].

Network	Nodes	Edges	Construction	Datasets
Protein-protein interaction	Proteins	Physical interactions. Undirected.	Literature curation, high-throughput experiments, computational prediction [66]	HPRD[93], BioGRID[18], IntAct[84], STRING[28]
Metabolic	Biochemical metabolites	Reactions, Enzymes. Directed or undirected	Literature curation and prediction based on orthology	BIOCYC [75], metaTIGER [50], KEGG [61]
Gene regulatory	Transcription factors, DNA regulatory elements	Transcription factor - regulatory element relationship. Directed	Y1H, CHiP [66, 94], Gene Knockdown [13], Coexpression [6]	SysGenSIM [13], ARACNE [6], ChIP-Array [73].
Coexpression	Genes	Coexpression measure. Undirected.	Built by computing expression profile similarity measures on transcriptonics data such as the datasets in GEO [89]	GEO [89]

Table 2.1: Summary of biological networks

Yeast two-hybrid (Y2H) methods work by co-opting the transcription mechanism genetically modified yeast cells. The system is designed in a way as to ensure that a reporter gene will be transcribed only when the Bait and Prey proteins interact [101]. As shown in figure 2.1, the protein of interest, the Bait, is bound to a DNA binding domain, while the Prey protein is bound to a Transcriptional activation domain. As the transcription factor will only be functional when both the DNA binding domain and the Transcriptional activation domain are present the reporter gene will only be expressed in those cells in which the bait and prey interact [8].

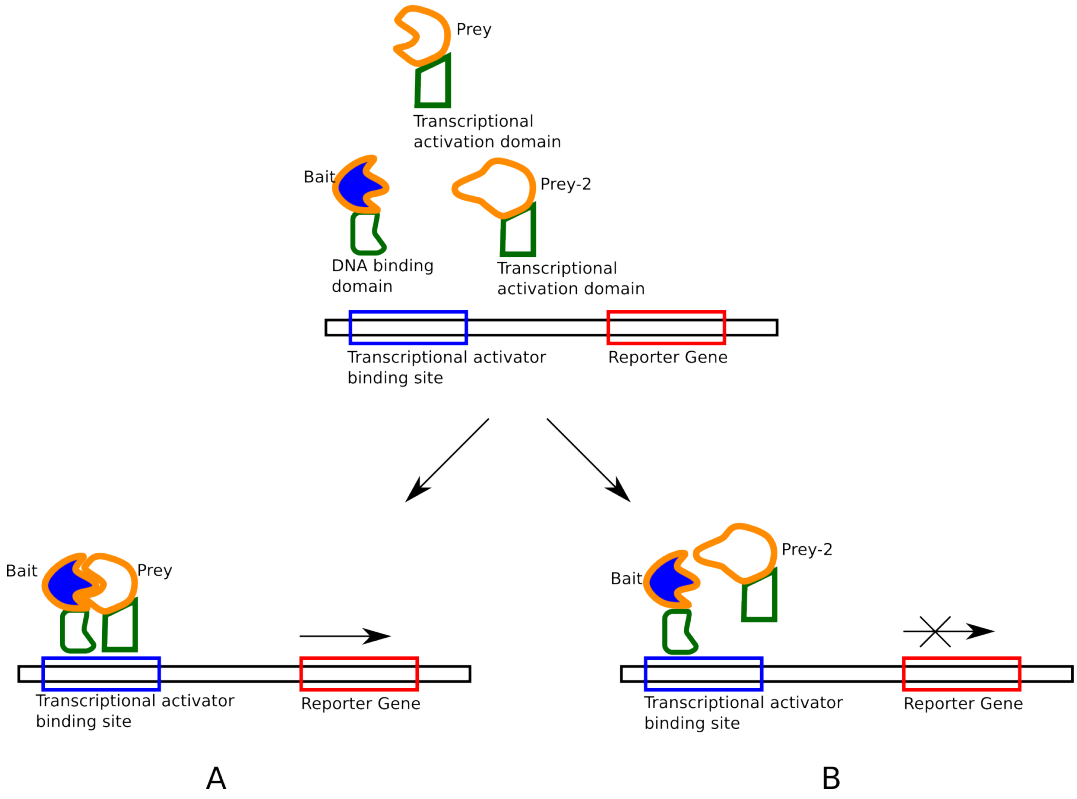

Figure 2.1: Yeast 2-hybrid. A) shows the case in which the Bait and Prey interact, with the resulting expression of the Reporter Gene. B) shows a case in which the Bait and Prey-2 protein do not interact, and the Reporter Gene not being expressed

Figure 2.2 presents an outline of a typical AP/MS experiment. The process begins by affixing a Bait protein to a matrix to inmobilise it (A). The Bait protein is passed through a protein mixture (B) where it will bind to its interacting partners (C). Through a series of washes (*i.e.* purification steps) the Bait is separated from its Preys (D), which are then analysed by Mass Spectrometry (E). AP/MS experiments are repeated for the same bait protein, resulting in weighted interactions of the bait-prey proteins [103].

In contrast to the binary interactions produced by Y2H methods, AP/MS approaches result in weighted co-complex information [103], that is, AP/MS can report

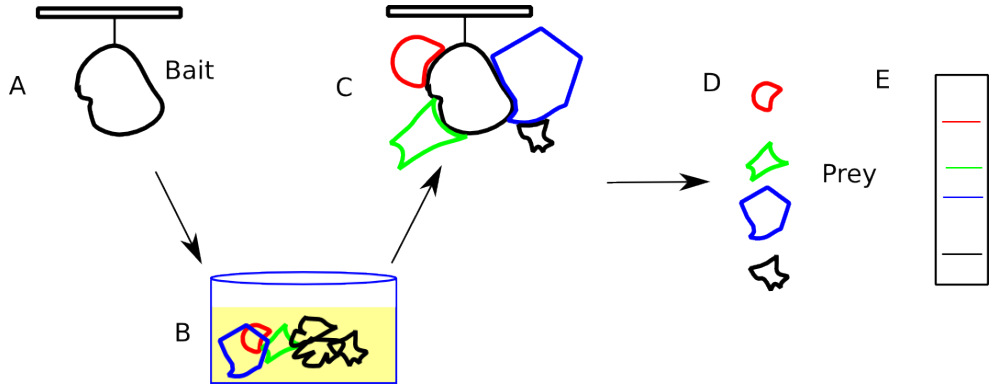

Figure 2.2: Affinity Purification / Mass Spectrometry. The bait protein is attached to an inmobilising matrix (A). The attached Bait protein is passed through a protein mixture where the interacting partners attach (B) (C). Through a series of purification steps the preys are separated (D) and analysed through mass spectrometry.

links between proteins that do not directly interact (notice the black protein in column D or figure 2.2). Y2H methods are not without fault, as they can fail to detect interactions that occur after post-translational modifications [8]. These methods are complementary, and a combination of methods might be required to obtain a high-quality, high-covering PPI network [8].

Several of the existing databases are the result of manual and automated literature curation by the authors, as is the case of the Mammalian Protein-Protein Interaction Database (MIPS) [70], the Molecular interaction database (MINT) [9], the Biomolecular Interaction Network Database (BIND) [36] and the Human Protein Reference Database (HPRD) [93]. IntAct [85] follows a more collaborative approach, where literature curation efforts are augmented by user-submitted interactions. In contrast, the High-quality interactomes (HINT) [45] and STRING [28] provide integrated PPI networks, constructed by integrating multiple databases (among which are MIPS, Human Protein Reference Database (HPRD), and others) and additionally, in the case of STRING, predicting interactions through machine learning approaches. In

addition to the construction of the networks substantial bioinformatics work is required to produce comprehensive and accessible datasets, in order to combine the individual experiments into comprehensive PPI datasets that are useful as systems biology tools.

While PPI networks are incomplete [46], some topological characteristics can readily be observed. The networks tend to be scale-free, that is, there are few very well connected nodes while most tend to have few connections [12]. This degree distribution is thought to be associated with gene duplication events which result in already well connected nodes gaining even more connections [12]. Interestingly, the scale-free nature of the networks does not seem to be associated to biases in the study of diseases [95].

These well-connected *Hub* nodes account for most of the essential genes [41]. This is indeed a property of scale free networks such as the Internet [68] where the removal of a central node has catastrophic consequences. Interestingly, only few disease genes encode hubs and the ones that do correspond to the minority of disease genes that are essential genes in the organism [54].

2.3 Human diseases and biological networks

For many diseases the disease status is determined by conditions that have to be met for the disease to be diagnosed [22]. This approach focuses on the observable pathophenotype [22] and is still a staple in modern medicine as evidenced by concepts such as the Medical algorithm [83]. However, complex diseases are not necessarily amenable to this type of analysis. In some diseases, for example *Phenylketonuria* (`MIM:261600`)) penetrance can be incomplete, causing only a subset of the individuals with a particular genotype to develop the phenotype. In other cases penetrance can vary with age, as is the case in late-onset diseases such as *ARMD1* (`MIM: 603075`).

A genotype can also have variable expressivity leading to a continuum of resulting phenotypes as in the case of *Marfan Syndrome* (`MIM: 154700`) [48, 27]. These complex scenarios reveal an underlying complexity that is not explained by linear disease-gene associations and therefore require a broader, more powerful analysis [22].

The systems view of diseases does not consider a disease to be a "whole" but rather a consequence of wider perturbations in the interactome – the "disease module" [63, 66, 54]. These perturbations have non-linear effects that result in the collection of phenotypes that co-occur to bring about the disease [22, 95]. This broader relationship between diseases and their causes have important implications on our understanding of human disease, as is evidenced in Goh's *et al.* Diseasome [54]

2.4 Relating diseases through biological networks

In their work, Goh *et al.* propose the construction of the "Diseasome", a global map of the disease-gene relationships. The Diseasome is represented as a bipartite graph in which one set of nodes represents the diseases and the other set of nodes the known disease genes. The links between the disease set and the gene set represent the known disease-gene associations found in Online Mendelian Inheritance in Man (OMIM). Based on the Diseasome, two complementary networks are constructed: the Human Disease Network (HDN) and the Disease Gene Network (DGN). In the HDN nodes represent diseases and two diseases are connected if they share common disease genes, while in the DGN nodes are genes and links connect genes associated to the same disease.

To illustrate the construction of the Diseasome and its complementary representations the HDN and the DGN, consider the disease-gene associations shown in table 2.2.

Figure 2.3 illustrates the Diseasome based on the disease-gene associations in

Disease	Gene
Familial Expansible Osteolysis (MIM:174810)	*TNFRSF11A*
Paget Disease of Bone (MIM:602080)	*SQSTM1,TNFRSF11A*
Osteopetrosis, Autosomal Recessive 7 (MIM:612301)	*TNFRSF11A*

Table 2.2: Example of disease-gene associations obtained from OMIM.

table 2.2. In the HDN, there are three nodes, one for each disease and since *TNFRSF11A* is associated to all three diseases, the network is fully connected. The link between (MIM:174810) and (MIM:602080) is wider, since these diseases share two disease genes. In the DGN there are two nodes, one for each gene, and one connection representing the fact that genes *TNFRSF11A* and *SQSTM1* are both associated to (MIM:602080). The complementary representations provided by the HDN and DGN provides a systems view of the interconnectedness of heritable diseases.

Some fundamental topological features of the disease modules are highlighted by Goh's *et al.* work. In particular, the authors show that, not only the genes associated to a disease tend to be functionally coherent, but genes associated with similar disorders are related to one another [54]. The interconnectedness of the diseases enables wider analyses to be carried out, identifying the extent of the perturbations that drive the diseases [95]. In particular, this understanding that the underlying causes of similar diseases must be somehow related allows us to explore disease similarity measures that would quantify the distance between the disease modules [64].

Figure 2.4 reproduces the figure presented by Goh *et al.* [54]. It is important to note that while the layout was designed manually, both the HDN and the DGN reveal the interconnectedness of human diseases. In the figure, each disease is coloured according to one of 22 disease classes, and it can readily be noted that diseases belong to the same class are tightly connected.

From a network medicine perspective quantifying disease similarity at molecular level would allow the transfer of knowledge between similar diseases [64], possibly

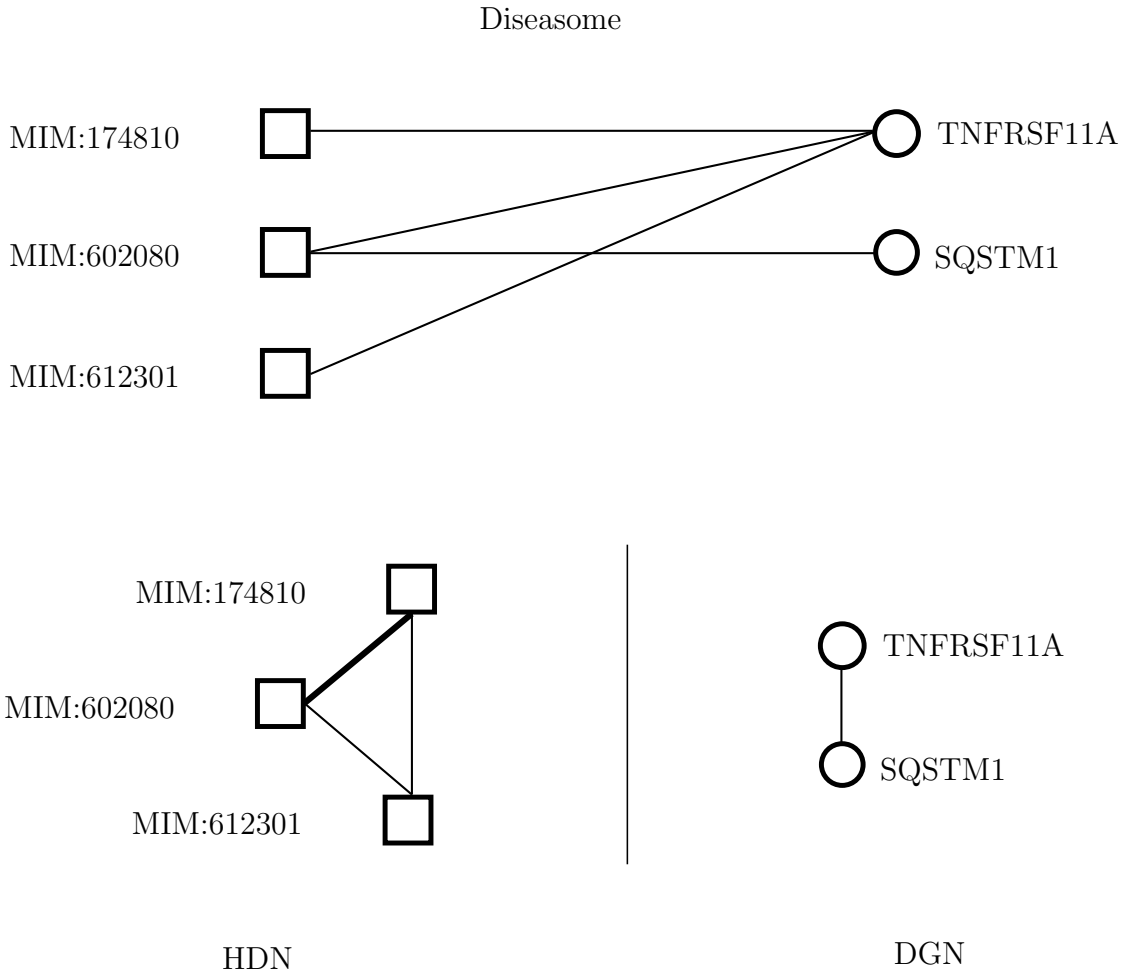

Figure 2.3: Example of the HDN and DGN projections of the bipartite diseasome. This example represents a subset of the diseasome based on data extracted directly from OMIM.

providing hypotheses for causal genes discovery and even suggestions for drug repositioning. The Diseasome follows a bottom-up approach, relating the diseases through their molecular basis. However, the lack of molecular-level information about the diseases (for about 30% of hereditary diseases in OMIM no disease gene is currently known) suggests that a wide covering and accurate measure should rely on a combination of the phenotype and genotype data in order to accurately quantify similarity between diseases.

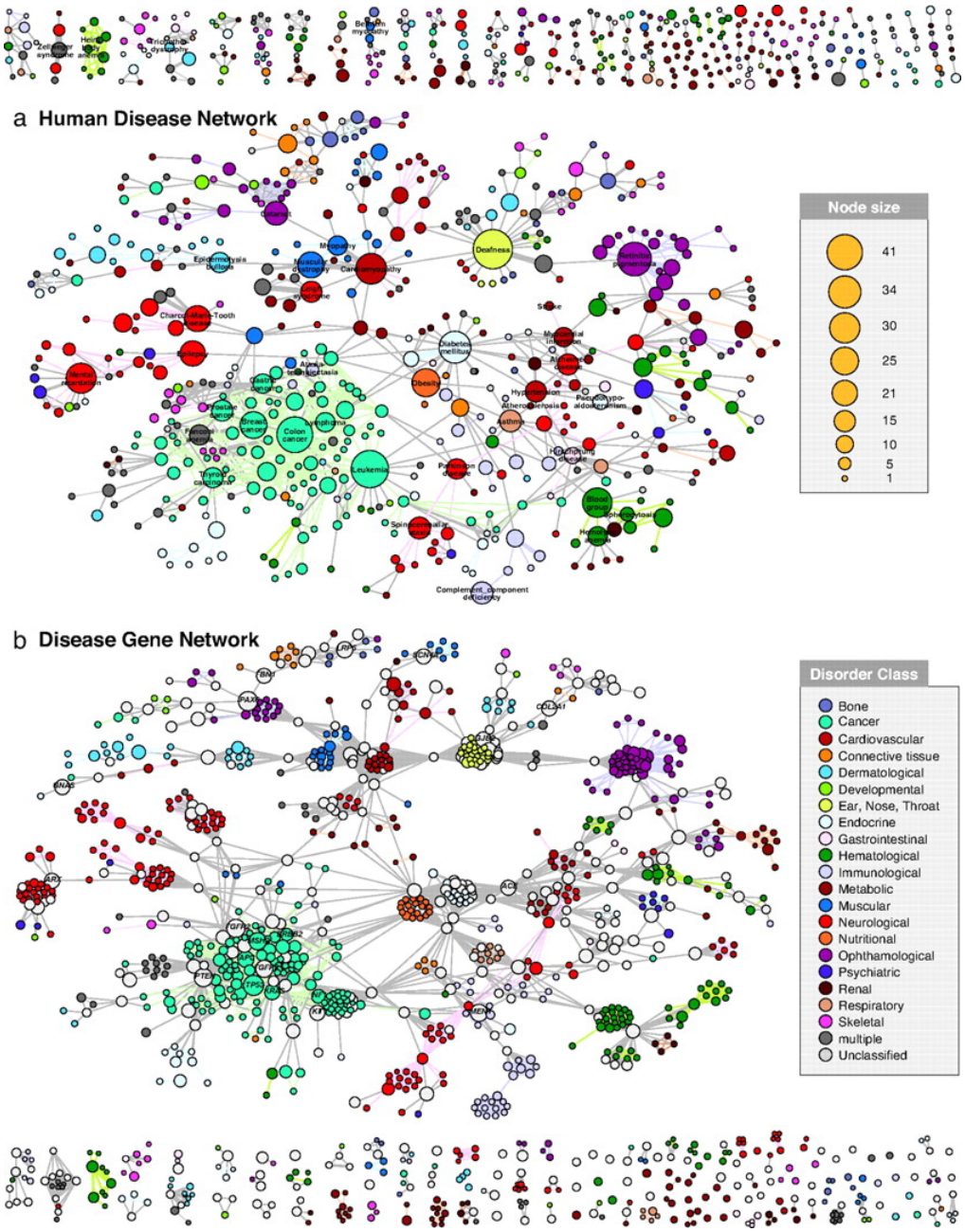

Figure 2.4: Visualisation of the HDN and DGN projections of the bipartite diseasome. Each disease is coloured based on its disease class. The figure is reproduced from Goh *et al.* [54].

The complexities of the genotype-phenotype relationship provides an argument for more comprehensive phenotype-based approaches to relating diseases. In particular, phenotypic information is more readily available and is more comprehensive than the existing molecular information [63].

Chapter 3

Existing methods for disease similarity

Few methods for quantifying disease similarity have recently appeared. These can be classified into two groups: phenotype and molecular data -based approaches. In the first group, the measures by van Driel *et al.* [64], Köhler *et al.* [81] and Zhou *et al.* [106] stand out. In the second group are the methods proposed by Goh *et al.* [54], Park *et al.* [86] and Mathur and Dinakarpandian [82].

In this chapter, I will provide an in-depth review of the relevant literature, analysing the motivation behind the existing approaches, identifying relevant features and shortcomings.

3.1 van Driel's *et al.* text mining analysis

van Driel *et al.* [64] present a measure based on text-mining analysis of the disease phenotype descriptions found in Online Mendelian Inheritance in Man (OMIM). The authors show that phenotype level similarity relates to the function of the genes associated to the diseases [64]. To verify the correlation between phenotype similarity

and protein level function, genotype distance correlation, van Driel *et al.* defined four molecular-level relationships. These relationships relate disease-genes through physical protein-protein interactions and three protein-protein similarities based on sequence, protein motifs and Gene Ontology (GO) function.

3.1.1 The method

van Driel's *et al.* fundamental building block is a "record" composed of the combination of the Clinical Synopsis (CS) and Text (TX) fields mined from OMIM. Due to the free-text nature of the entries, van Driel *et al.* implemented text analysis techniques to systematically retrieve the knowledge contained in them and produce a descriptive feature vector for each disease.

The records are mined for a predefined set of Medical Subject Headings (MeSH) terms which are used to construct feature vectors for every disease. The authors used the terms from the Anatomy [A] and Diseases [C] MeSH ontologies, which they filtered removing general terms such as "disease" and "syndrome". This resulted in $4,145$ MeSH terms to be used as features for the disease records. In order to construct the feature vector, the records are parsed counting the number of occurrences of each feature. In this way, every disease is represented by a $4,145$-dimensional feature vector, in which the occurrence count of each feature represents its relevance with respect to the disease.

The authors recognise the need to account for the hyponym - hypernym (*i.e.* the relationship between more specific and more general terms) relationship (see chapter 1). As the authors observe, whenever a term from the feature set appears in a record, its ancestors must also be considered [64]. For example, if a record references the feature (`Retina`), its hypernyms (*i.e.* ancestors) such as the term (`Eye`) must be relevant to the record as well.

To account for the hierarchical structure of MeSH, each term along the path to the

root starting from each feature will be considered relevant in relation to the feature found in the record. The relevance of feature c is given by:

$$r_c = r_{c,counted} + \frac{\sum r_{hypos}}{n_{hypo,c}} \tag{3.1}$$

Where r_{hypos} is the relevance of the hyponyms of feature c, $n_{hypo,c}$ is the number of hyponyms of feature c and $r_{c,counted}$ is the counted frequency of feature c in the record. r_c is calculated iteratively for each term starting at the deepest level of the MeSH ontology.

The features extracted from the MeSH ontologies are not equally informative. The increasing specificity of the terms as the ontology is traversed down results in very general terms used to describe diseases together with very specific terms. For example, the term *Chromosomes* (`D002875`) is used $4,294$ times across OMIM while *Afibrinogenemia* (`D000347`) appears only twice. This imbalance is corrected by the authors using the inverse document frequency measure (tdf-idf), thereby accounting for the variable frequency of the terms. The weight w_c of term c is given by the base 2 logarithm of the fraction of disease records that are described a specific term. Formally:

$$w_c = r_c * \log_2 \frac{N}{n_c} \tag{3.2}$$

Where N is the total number of OMIM records, and n_c is the number of times a MeSH feature c is used to annotate records in OMIM. The inverse document frequency reduces the relevance of a very common term while increasing the relevance of more infrequent terms that could provide more information. Thus, a rare term such as *Afibrinogenemia* (`D000347`) would have a weight given by $w_{D000347} = \log_2 \frac{7,812}{2} = 11.93$ while the more common term *Chromosomes* (`D030342`) would have a weight of $w_{D030342} = 0.86$.

The weight of a term is further refined to account for the variability of length of

the OMIM records.

$$weight_c = \frac{1}{2}\left(1 + \frac{w_c}{r_{mf}}\right) \tag{3.3}$$

In this way the weight of a feature $weight_c$ in a record is a function of the concept's frequency r_c divided by the frequency of the most occurring feature in that record, given by r_{mf}. Equations 3.1, 3.2 and 3.3 are applied in order to each disease's feature vector.

The pairwise similarity of the records is given by the cosine of the angle between each corrected feature vector, given by equation 3.4.

$$s\left(x, y\right) = \frac{\sum_{i=1}^{l} x_i y_i}{\sqrt{\sum_{i=1}^{l} x_i^2}\sqrt{\sum_{i=1}^{l} y_i^2}} \tag{3.4}$$

To illustrate van Driel's *et al.* method, I present a toy example based on a set of $1,000$ diseases and a feature set of 5 features A, B, C, D, E. For a a given disease the frequencies of the features are given by $r_{c,counted} = \langle 3, 4, 0, 1, 0\rangle$. According to van Driel's *et al.* method, equations 3.1, 3.2 and 3.3 are applied in order to every diseases' feature vector.

Based on the structure of the ontology in figure 3.1, vector $r_{c,counted}$ will be transformed by equation 3.1 recursively into

$$r = \langle 3_{r_A}, 4_{r_B}, 3.5_{r_C}, 1_{r_D}, 2.25_{r_E}\rangle$$

Since leaf terms A, B, D have no hyponyms, their relevance is given by their counted frequencies r_A, r_B and r_D respectively ($\frac{\sum r_{hypos}}{n_{hypo,c}}$ in equation 3.1 equals zero). For non-leaf terms their relevance is determined by their own frequencies and the relevance and number of their hyponyms. For C, $\sum r_{hypos}$ is the sum of the relevances of A and B, namely 7, and $n_{hypo,c}$ is 2. The relevance of C is then given by $r_C = 0 + \frac{7}{3} = 3.5$.

To weigh the relevances according to the number of times each feature is used to

describe the $N = 1,000$ diseases in this example, we need to define the total number of disease records each feature is used to describe:

$$n_{feature} = \langle 100_{n_A}, 500_{n_B}, 20_{n_C}, 3_{n_D}, 1000_{n_E} \rangle$$

According to equation 3.2, the weights are defined as follows:

$$w_{feature} = \langle 9.96_{r_A}, 4_{r_B}, 31.36_{r_C}, 8.38_{r_D}, 0_{r_E} \rangle$$

The weight of A is given by $w_A = 3 * log_2 \frac{1000}{100} = 9.96$

Finally, equation 3.3, accounts for the variable length of the records:

$$weight_{feature} = \langle 1.74_{weight_A}, 1_{weight_B}, 4.42_{weight_C}, 1.54_{weight_D}, 0.5_{weight_E} \rangle$$

For feature A, the final weight $weight_A$ is determined by the quotient of weight w_A of the term and the frequency of the term most used to describe diseases, namely B: $weight_A = \frac{1}{2} * \left(1 + \frac{w_A}{4}\right) = 1.74$

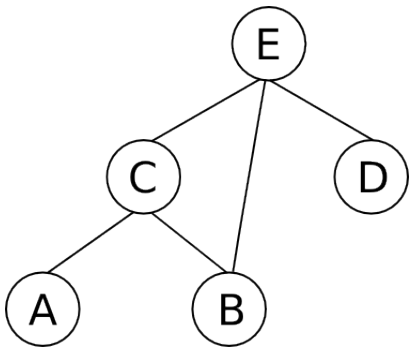

Figure 3.1: The final weights for the 5 features in the ontology are $weight_{feature} = \langle 1.74_{weight_A}, 1_{weight_B}, 4.42_{weight_C}, 1.54_{weight_D}, 0.5_{weight_E} \rangle$.

3.2 Köhler's *et al.* Human Phenotype Ontology

Köhler *et al.* observe that while OMIM is a broad and detailed compendium of heritable diseases, it is not suitable for automated analysis. The main obstacle resides in the lack of a centralised vocabulary that could be used to describe the OMIM diseases. This is particularly relevant considering that synonyms are not taken into account in the construction of OMIM (*e.g. muscle atrophy* and *muscular atrophy* are used to describe the same concept).

The authors propose the construction of the Human Phenotype Ontology (HPO), a wide covering, purpose specific ontology aimed at describing the phenotypic abnormalities of the diseases in OMIM. The HPO is constructed through a process of automated and manual curation by the authors [81]. Each term in the ontology represents a phenotypical abnormality and terms are related to one another through an *is_a* relationship. For details on the construction of the HPO, refer to chapter 1 § 1.4.

The authors provide annotations for OMIM, which enables disease similarities to be calculated based on the term-based similarities of their annotations.

3.2.1 The method

The method proposed by Köhler *et al.* measures similarity of two OMIM diseases based on the semantic similarity of terms in the HPO. The similarity of two terms t_1, t_2 is defined by the Resnik [72] similarity of the terms, that is, the information content of their common ancestor with highest information content, that is:

$$TermSim\,(t_1, t_2) = max_{a \in A(t_1,t_2)} - \log p\,(a) \tag{3.5}$$

where $A = (t_1, t_2)$ is the set of ancestors common to t_1 and t_2.

Based on this term-wise similarity, disease similarities are calculated as follows:

$$sim\left(d_1, d_2\right) = avg\left[\sum_{s \in d_1} max_{t \in d_2} TermSim\left(t_1, t_2\right)\right] \tag{3.6}$$

where s are terms annotating disease d_1 and t are terms annotating disease d_2.

Since this similarity score is not symmetric, a transformation is applied:

$$HPOsim\left(d_1, d_2\right) = \frac{1}{2} sim\left(d_1, d_2\right) + \frac{1}{2} sim\left(d_2, d_1\right) \tag{3.7}$$

3.3 Zhou's *et al.* Disease Symptom Network

Zhou *et al.* [106] propose the construction of the Human Symptoms Disease Network (HSDN), a network that reflects the dynamics of heritable diseases from the perspective of the physical manifestations that characterise them. In the HSDN, nodes are diseases and the links represent similarities calculated based on the co-occurrence of symptoms.

In order to build the HSDN, Zhou *et al.* mine PubMed analysing the co-occurrence of a symptom and a disease. The co-occurrence is compiled into a feature vector that characterises each disease based the frequency of its symptoms. Pairwise similarity between diseases is obtained by calculating the cosine of the angle between the feature vectors and then conserving only statistically significant scores.

3.3.1 The method

Zhou *et al.* define the *symptoms* by filtering the Diseases [C] ontology in MeSH, extracting terms which describe clinical manifestation of diseases contained in the sub ontology *Signs and Symptoms* [C23.888]. To define *diseases*, a similar procedure is followed, whereby the Diseases [C] ontology is used excluding the *Signs and Symptoms*

[C23.888] sub ontology, the *Animal Diseases* [C22] sub ontology and some general terms like "Disease" and "Symptom". Through this process, the authors obtain $4,442$ diseases and 327 symptoms.

The main assumption is that disease terms and symptom terms that co-occur in a publication indicate a relation between the disease and the symptom. By mining PubMed, the authors obtain the set of MeSH terms associated to each publication, which enables the construction of a feature vector that describes the co-occurrence of diseases and symptoms in PubMed. For a given disease, its feature vector d_j is defined as follows:

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{n,j}) \tag{3.8}$$

where $w_{i,j}$ denotes the number of times a feature i coincides with a disease j in PubMed.

To account for the different specificity of the terms in MeSH, (*e.g.*) the tdf-idf was measured:

$$w_{i,j} = w_{i,j} * \log \frac{N}{n_i} \tag{3.9}$$

where N denotes the total number of diseases, and n_i the number of diseases which reference symptom i. The $w_{i,j}$ weights are further filtered by a Chi-squared test with a threshold P-value of 0.05 in order to preserve only significant disease-symptom associations.

The remaining significant vectors determine the symptoms most relevant to each disease, and an appropriate measure of similarity between them would determine similarity between the diseases. This is done following a similar approach to van Driel *et al.* [64], calculating the cosine of the angle of the feature vectors:

$$cos\,(d_1, d_2) = \frac{\sum_i d_{x,i} d_{y,i}}{\sqrt{\sum_i d_{x,i}^2}\sqrt{\sum_i d_{y,i}^2}} \tag{3.10}$$

3.4 Goh's *et al.* Human Disease Network

I preface this section by noting note that while Goh *et al.* [54] do not provide a similarity measure as such, I have developed a method to produce disease similarity scores based on the data presented by the authors in [54]. For a detailed discussion on the construction of the Diseasome, refer to chapter 2 § 2.4.

The authors combine the diseases in OMIM into 1,284 syndromes through automatic textual comparison of the names followed by manual curation. For example *Anaemia, hypochromic microcytic* (`MIM:206100`) and *Anaemia, hemolytic, Rh-null, regulator type* (`MIM:268150`) are combined into a single syndrome, *Anaemia*. Each syndrome is then categorised into one of 21 categories according to the physiological system it affects. The categories are as follows: *a*) Bone, *b*) Cancer, *c*) Cardiovascular, *d*) Connective tissue disorder, *e*) Dermatological, *f*) Developmental, *g*) Ear-Nose-Throat, *h*) Endocrine, *i*) Gastrointestinal, *j*) Haematological, *k*) Immunological, *l*) Metabolic, *m*) Multiple, *n*) Muscular, *o*) Neurological, *p*) Nutritional, *q*) Ophthalmological, *r*) Psychiatric, *s*) Renal, *t*) Respiratory and *u*) Skeletal. The authors labelled all remaining syndromes without an appropriate category such as *Alcoholism, susceptibility to* (`MIM:103780`) and *Antley-Bixler syndrome* (`MIM:207410`) as "Unclassified".

I constructed a binary disease similarity measure from the Diseasome through a mapping that considers all diseases in a category to be similar (similarity equal to 1) and all pairs of diseases with diseases appearing in different categories as dissimilar (similarity equal to 0). For example, *Sarcoidosis, early-onset* (`MIM:181000`) and *Neutrophil immunodeficiency syndrome* (`MIM:608203`) which were both categorised as "Immunological" by Goh *et al.* , are are therefore considered similar according to the mapping I propose. Following the work of Goh *et al.* , this binary similarity measure, albeit coarse, reflects the organisation of the interactome, whereby diseases that are

similar at phenotype level (*i.e.* affect the same physiological system), are the result of perturbations in nearby regions in the interactome [54]. The technical details of the construction of the disease similarity score based on the Diseasome are presented in chapter 4.3.

3.5 Park's *et al.* co-localisation of disease proteins

Park *et al.* [86] study the phenotypic similarities of heritable diseases analysing the subcellular location of the disease proteins. In their work, Park *et al.* prove that disease proteins from phenotypically similar diseases tend to be co-localised in the same cellular compartment. The author's main assumption is that proteins associated to phenotypically similar diseases are more likely to share a subcellular compartment. In order to verify this claim, Park *et al.* construct a disease similarity measure based on the location of the known disease proteins. Furthermore, diseases affecting the same physiological system show significant association with specific cellular compartments.

3.5.1 The method

Park *et al.* describe a disease in terms of the subcellular location of its disease proteins by measuring the association of the disease proteins to 10 predetermined subcellular compartments. These spatial profiles are compiled into the Disease-associated Protein and Subcellular Localization (DPL). In Park's *et al.* work, a disease corresponds to one of the $1,284$ syndromes defined by Goh *et al.* [54] (see § 3.4), and disease-protein associations are obtained from OMIM.

The authors use the GO Cellular Component ontology to annotate the disease related proteins in order to describe their subcellular localisation. To determine the appropriate GO term for the $1,171$ disease associated proteins, the authors use a combination of predicted subcellular locations (for 609 proteins) and annotations

provided by UniProt GOA (for $1,168$ proteins).

Based on these GO annotations, the authors calculate an association score defined by the Ochai coefficient, between a disease D_i and a subcellular location L_j:

$$OC\left(D_i, L_j\right) = \frac{P_{D_i \in L_j}}{\sqrt{P_{D_i} \times P_{L_j}}} \tag{3.11}$$

where $P_{D_i \in L}$ denotes the number of proteins associated to disease D_i which are located in subcellular compartment L_j; P_{D_i} and P_{L_j} are the number of proteins associated to disease D_i and the number of disease proteins located in location L_j respectively. The association score between a disease and a location is then defined by normalising the $OC\left(D_i, L_j\right)$ score as follows:

$$AS\left(D_i, L_j\right) = 100 \frac{OC\left(D_i, L_j\right)}{\sum_k OC\left(D_i, L_k\right)} \tag{3.12}$$

In this way, a disease's association score $AS\left(D_i, L_j\right)$, can be thought of as a vector of length $\|L\|$ that sums up to 100. This vector constitutes a "location profile" for a disease. The similarity between two diseases is given by the Pearson Correlation Coefficient of these location profiles [86]. Formally:

$$PCC_{ij} = \frac{N_l \sum_l AS_{il} AS_{jl} - \sum_l x_{il} \sum_l x_{jl}}{\sqrt{N_t \sum_l x_{il}^2 - \left(\sum_l x_{il}\right)^2} \sqrt{N_t \sum_l x_{jl}^2 - \left(\sum_l x_{jl}\right)^2}} \tag{3.13}$$

AS_{il} denotes the association score between disease i and subcellular localisation j.

3.6 Disease similarity based on functional similarity of disease proteins

A few methods have recently appeared linking the similarity of phenotypes to the functional similarity of disease proteins, such as the ones by Cheng *et al.* [55], Suthram *et al.* [88] and Mathur and Dinakarpandian [82]. These methods are conceptually similar, and in the following, Mathur and Dinakarpandian's method will be presented as an example.

Mathur and Dinakarpandian [82] present a measure based on functional similarity of the disease associated proteins, given by the semantic similarity of GO terms annotating the proteins. The method has two steps. First, the authors design a new semantic similarity measure and evaluate its performance by comparing it with existing measures. The similarities between disease-associated proteins are then used to determine similarity between the diseases.

The protein semantic similarity measure

The semantic similarity measure proposed by the authors considers the graph structure of the ontology and the co-occurrence of the annotations simultaneously, under the assumption that the continually changing structure in the GO will render existing similarity measures inaccurate [82].

The authors measure co-occurence of annotations for a pair of terms through the Jaccard coefficient of GO terms of two annotated gene products, namely:

$$sc\left(x,y\right)=\frac{n\left(x\cap y\right)}{n\left(x\cup y\right)} \tag{3.14}$$

where x, y are GO terms and $n\left(x\cap y\right)$ are the number of genes annotated with term x and y simultaneously.

The authors note that the Jaccard coefficient would be inaccurate as specific terms would have the same effect on the similarity as would very general terms. They account for this effect by weighting the *sc* with the average information content of the terms:

$$sim\left(x,y\right)=sc\left(x,y\right)Avg\left(IC\left(x\right),IC\left(y\right)\right) \tag{3.15}$$

$sim\left(x,y\right)$ quantifies similarity between two terms in an ontology accounting for the different specificity of the terms and relevance of the terms [82].

Conceptually, Mathur and Dinakarpandian's *sc* measure is similar of the Graph-based measure simUI (see § 1.6.3). The main difference lies in that While simUI measures gene similarity based on the GO terms used to annotate the genes, *sc* while the *sc* measures term similarity based on the genes the GO terms annotate.

To measure the similarity between two annotated objects, that is, between two sets of terms, the authors propose the following:

$$Mb\left(A,B\right)=\frac{1}{2}\left[\frac{\sum_{1\leq i\leq m}msim\left(T_{Ai},T_{B}\right)}{m}+\frac{\sum_{t\leq j\leq n}msim\left(T_{Bj},T_{A}\right)}{n}\right] \tag{3.16}$$

where A,B are genes, T_A,T_B are terms annotating A and B and m and n are the number of terms annotating A and B. $msim\left(T_{Ai},T_{B}\right)$ is the maximum semantic similarity between the ith term annotating A and the set of terms annotating B, as given by equation 3.15.

3.6.1 The method

Mathur and Dinakarpandian propose two methods to determine disease similarity based on the function of the disease proteins: Process-Similarity based (PSB) and Process-Identity based (PIB).

The PSB and PIB measures are a three stage process. The first stage removes

terms that are not statistically significant in describing the function of the disease proteins. The second stage eliminates very general terms, and finally, the similarity is calculated.

In the first stage, the authors consider only the over-represented GO terms associated to each disease (hypergeometric test, Benjamini-Hochberg correction) [82].

In the second stage, to reduce the effect of very general, but over-represented terms associated to the disease proteins, the authors normalise the information content of each GO term with respect to each disease it annotates:

$$NF = \frac{IC_{GO}\left(term\right)}{MaxIC_{GO}} * \frac{IC_{DIS}\left(term\right)}{MaxIC_{DIS}} \tag{3.17}$$

$IC_{GO}\left(term\right)$ is the information content of $term$, $MaxIC_{GO}$ is the maximum information content of any term in GO, $IC_{DIS}\left(term\right)$ is the information content of $term$ with respect to the GO terms annotating the disease and $MaxIC_{DIS}$ is information content of the most informative term annotating the disease. The normalising factor NF for $term$ is defined as the product of the relevance of $term$ with respect to all other terms in GO and the relevance of $term$ with respect to all other terms annotating a disease. $\frac{IC_{GO}(term)}{MaxIC_{GO}}$ determines how relevant $term$ is in the ontology, while $\frac{IC_{DIS}(term)}{MaxIC_{DIS}}$ measures the relevance of $term$ in the context of the GO annotations of the diseases' proteins. The normalising factor NF is then multiplied by the maximum semantic similarity between the ith term annotating A and the set of terms annotating B $msim\left(T_{Ai}, T_B\right)$ used in equation 3.16.

Finally, in the third stage, the similarity is calculated. The Process-Similarity based (PSB) measures similarity based on the common GO terms annotating the two diseases. For every pair of diseases, the similarity is given by the measure proposed by the authors in equation 3.16. The Process-Identity based (PIB) measure, determines disease similarity based only on the common terms. For each GO term shared by a

pair of diseases, its self-similarity is calculated using equations 3.14 and 3.15. These values are summed, and the resulting score used to quantify disease similarity.

3.7 Discussion

In general, methods that rely on molecular information to quantify disease similarity will have to manage the limited information available. The lack of known genes for 72.9% (approximately 5,844 diseases) of OMIM's 8,006 (as of June 2015) diseases presents the major impediment, resulting in invariably low covering methods.

Of the existing measures that focus on phenotypic information, there are some characteristics that should be noted. The discussed measures, namely van Driel *et al.* , Köhler *et al.* and Zhou *et al.* , focus exclusively on the disease symptoms, and while they are important for categorising the diseases, the diseases should be thought of as multi-dimensional entities of which the symptoms are but one. I will show (see Chapter 4.3) that an important aspect of OMIM has been overlooked thus far by the existing methods – OMIM is a collection of highly diverse information. The database is the result of a process of curation [4] of the existing bio-medical literature and as such does not provide any new knowledge, but rather constitutes a centralised repository of all that is known about a particular heritable disease [4]. The descriptions contained in OMIM are far richer than the mere symptoms or phenotypes. The entries contain descriptions on mechanisms of inheritance, ethnic and racial characteristics of the affected individuals (*e.g.* increased prevalence of *Familial Mediterranean Fever* (`MIM:249100`) in Sephardic Jews) pathogenesis of the diseases and even relationships between drugs and the disease (*e.g. Sudden infant death syndrome* (`MIM:272120`) and Beta-blockers) among other characteristics. I will show that including this information is, in fact, useful.

Interestingly, when analysing Zhou's *et al.* and van Driel's *et al.* method, I noted

that there is substantial overlap between the CS fields and the subset of terms from the *Signs and Symptoms* [C23.888] contained in the Clinical Synopsis fields of OMIM. Zhou's *et al.* HSDN could be viewed as a further refinement of van Driel's *et al.* method. For example, van Driel *et al.* matches the term Ecchymosis (`D004438`) for disease *Glanzmann Thrombasthenia* (`MIM:273800`), and this term also co-occur in the publication identified by PubMed ID 14233375 "Hemorrhagic Thrombocytic Dystrophy. A Discussion Of Nosology", and is thereby associated by Zhou *et al.* to the same disease. Further large-scale analysis is not possible, due the lack of mapping between the Chemicals and Drugs [D] ontology used by Zhou *et al.* and OMIM used by van Driel *et al.* .

Of note is a situation that arises in van Driel's *et al.* method. While the authors acknowledge the value of the ontological structure of MeSH, they fail fail to appropriately account for their Directed Acyclic Graph (DAG) structure of the ontologies. This DAG structure causes some issues in light of equation 3.1, which results in some terms having an undue effect on the relevance of their ancestors. Consider the toy example presented in section § 3.1, figure 3.1. The relevance of term A is determined by equation 3.1. This equation defines the relevance of the terms based on the relevance of its hyponyms. However, since C is a hyponym of A, and B is a hyponym of both A and C, the relevance of term A would disproportionately consider the relevance of B: through C and directly. Should the ontologies be defined strictly as trees, this situation would not arise.

I will show in chapter 4.3, that the appropriate use of the ontology is fundamental for disease similarity measures the accurately quantify molecular relatedness between heritable diseases.

Chapter 4

A network medicine approach for disease similarities

A graphical conceptualisation of the method I developed is shown in Figure 4.1. The method begins (Step 1.) by extracting the publications referenced by all diseases in Online Mendelian Inheritance in Man (OMIM). In Step 2. the Medical Subject Headings (MeSH) terms which describe the publications are collected from PubMed. Each disease is then annotated with the MeSH terms associated to the publications it references. These sets MeSH terms, are compared (Step 3.) using ontological semantic similarity measures which produce, for each pair of MeSH annotated diseases, a single non-negative real number. The similarity of the MeSH terms annotating the diseases quantifies the similarity between the diseases, which in turn, quantifies that quantifies their molecular relatedness (Step 4.)

In this chapter I will discuss each step of my disease similarity measure in detail.

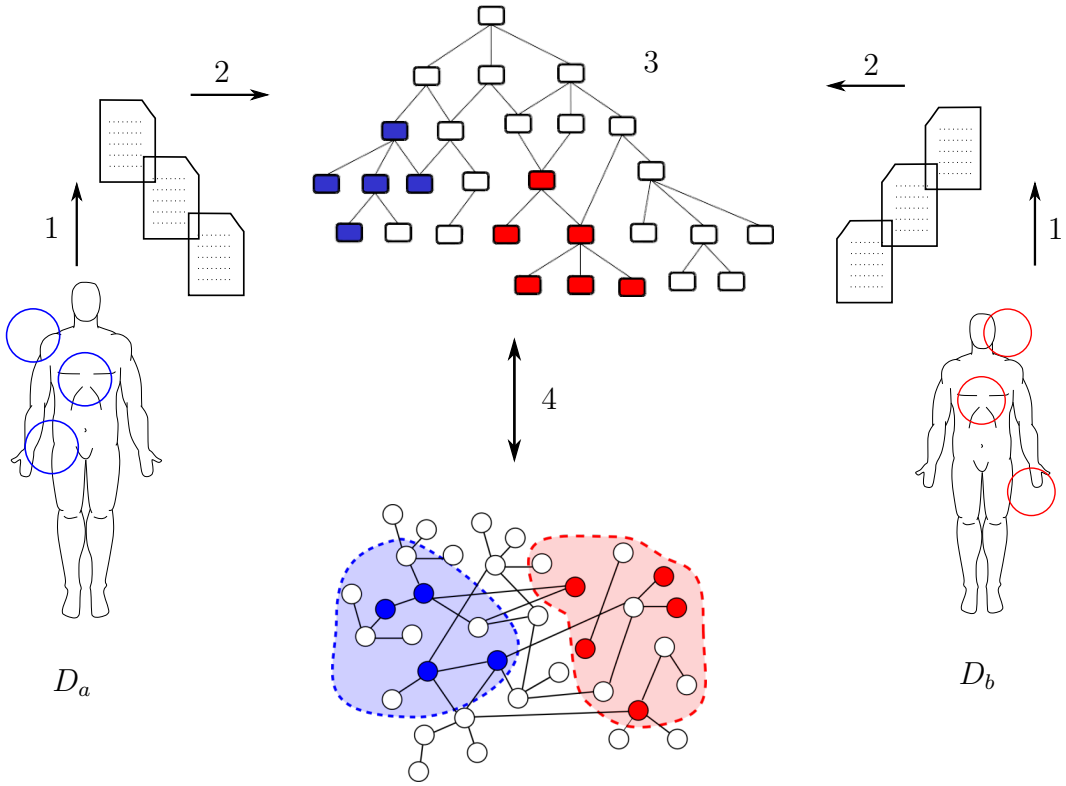

Figure 4.1: Outline of the method. The process starts with the mapping of OMIM diseases to PubMed publications (1). The MeSH terms for each publication are obtained from PubMed, mapping the OMIM disease onto the MeSH ontology (2). A semantic similarity measure quantifies the similarity between both sets of MeSH terms in the ontology (3). The resulting similarity score represents the molecular distance between the disease modules of diseases D_a and D_b (4).

4.1 Annotating OMIM diseases with MeSH terms

The assignment of MeSH terms to OMIM diseases, or the *annotation* of the diseases, consists in finding the set of MeSH terms that most accurately, unambiguously and concisely describe each disease. This concept is analogous to the Gene Ontology (GO) annotations of gene products [96].

The main assumption on which my disease similarity measure is based is that the MeSH terms used to describe the publications in PubMed will also be good descriptors

of the OMIM diseases. This assumption relies on the fact that the OMIM entries are compendia of the literature most relevant to the disease [4], and therefore the MeSH terms assigned to the referenced publication will also provide accurate and detailed descriptions for the diseases themselves [98]. This process of matching MeSH terms to OMIM diseases by proxy of the referenced publications is, considering the nature of the OMIM entries and MeSH annotations of PubMed publications, comparable to obtaining manual annotations of the OMIM disease.

Table 4.1 shows the number of diseases annotated by each MeSH ontology. An ontology is considered to annotate a disease if at least a term from that ontology is used to describe a disease.

Ontology	Annotated Diseases
Anatomy [A]	6,781
Organisms [B]	7,488
Diseases [C]	7,321
Chemicals and Drugs [D]	5,958
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	7,000
Psychiatry and Psychology [F]	3,271
Phenomena and Processes [G]	7,018
Disciplines and Occupations [H]	1,994
Anthropology, Education, Sociology and Social Phenomena [I]	1,903
Technology, Industry, Agriculture [J]	348
Humanities [K]	315
Information Science [L]	4,063
Named Groups [M]	6,775
Health Care [N]	4,257
Geographicals [Z]	2,834

Table 4.1: The 16 MeSH ontologies. The number of annotations is calculated by the diseases annotated with at least a term from the ontology.

4.2 MeSH based similarity of diseases

Having obtained the MeSH annotations for the OMIM diseases, a measure to compare these annotations is needed. I analysed and tested several well established semantic similarity measures namely the ones by Resnik [71], Pesquita [17], Jiang and Conrath [51], and Lin [25] as well as simpler topological measures. See Chapter 1 for details on the various semantic similarity measures tested. Resniks's [71] similarity measure proved to be superior in performance, and I have therefore chosen it to quantify similarity between OMIM diseases.

The method described so far may, and in some cases will, produce several similarity scores per pair of diseases. Since the ontologies are (conceptually) separate entities, the semantic similarity measure when used in each ontology separately will produce a score for every pair of diseases annotated with terms from the same ontology. This results in up to 16 scores for each pair of diseases. In order to obtain a single similarity score that characterises the molecular relatedness of the diseases, a combination of either the scores or the ontologies is required.

Since the MeSH ontologies are not mutually exclusive, I performed an analysis of the entire MeSH structures to verify the extent of the overlap. Figure 4.2 shows the amount of pairwise overlap between the different ontologies. In this figure, the colour of the links corresponds to the pairwise Jaccard coefficient between the ontologies and the width of the links correspond to the number of terms shared by the ontologies.

I noticed that, while the overlap between the ontologies connects most of them (of the 120 possible connections, 61 exist) it serves a more important purpose: the overlap establishes paths between the ontologies, even between those with no overlap. That is, it is possible to start from any ontology and visit all others (except Publication Characteristics [V], which will be discussed later) through the links resulting from the overlap. Based on this fact, I decided to combine the ontologies by adding a new

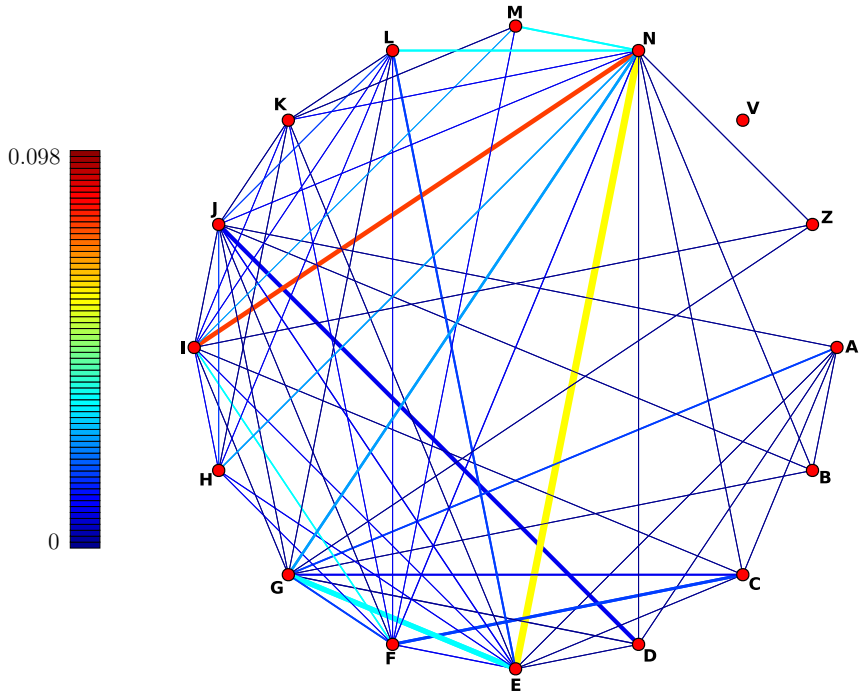

Figure 4.2: Overlap of the MeSH ontologies. Nodes represent MeSH ontologies and links are related to the amount of overlap between them. Link colours correspond to the Jaccard coefficient between the set of terms in each pair of ontologies. Link thicknesses correspond to the number of shared terms between ontologies and only strictly positive links are shown. MeSH Ontologies abbreviations: Anatomy [A], Organisms [B], Diseases [C], Chemicals and Drugs [D], Analytical, Diagnostic and Therapeutic Techniques and Equipment [E], Psychiatry and Psychology [F], Phenomena and Processes [G], Disciplines and Occupations [H], Anthropology, Education, Sociology and Social Phenomena [I], Humanities [K],Information Science [L], Named Groups [M], Health Care [N], Publication Characteristics [V], Geographicals [Z].

root node at level zero, connected to each of the ontologies root node through an *is_a* relationship. This results in a single, sweeping ontology that combines all areas of knowledge present in MeSH which, when analysed with an ontological semantic similarity measure, results in a single score for every pair of diseases. In Figure 4.3, the overlap is shown as nodes of different ontologies connected to one another. The red links between the A and F ontologies, establish a path between these otherwise disconnected ontologies.

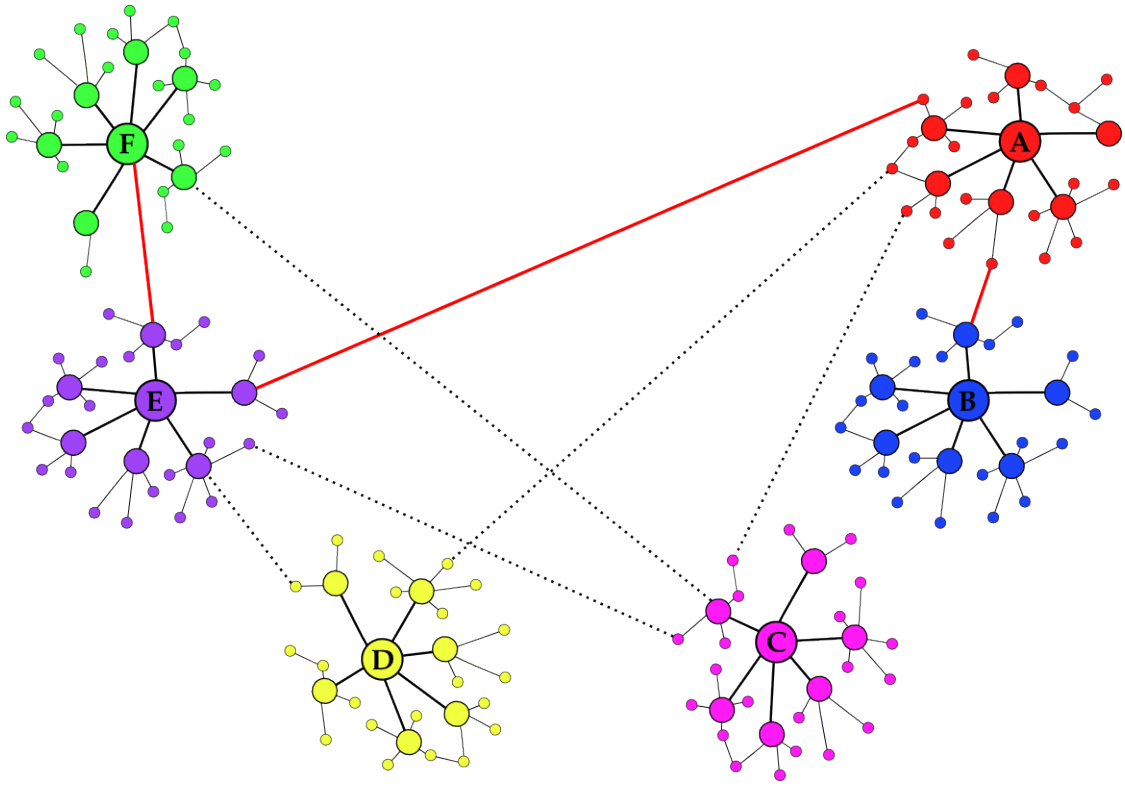

Figure 4.3: The overlap between the ontologies established paths between all of them. These paths (shown in red) allow the comparison of diseases annotated with terms from non-overlapping ontologies.

The fictitious root node is a requirement when using Resnik's semantic similarity measure, since otherwise some similarities would be undefined. If the fictitious root node did not exist, diseases annotated only with the root term of each MeSH ontology would not be comparable. In figure 4.4, terms t_1 and t_2 would not be comparable without the added root node **R**, they would lack a Lowest Common Ancestor (LCA) and their similarity would therefore not be defined.

The combination results in a single score for each pair of diseases for which a MeSH annotation can be produced. The scores are positive real unbounded numbers. No order preserving transformation with the aim of rescaling the scores was applied. Such a transformation would have no effect on performance and could make it data

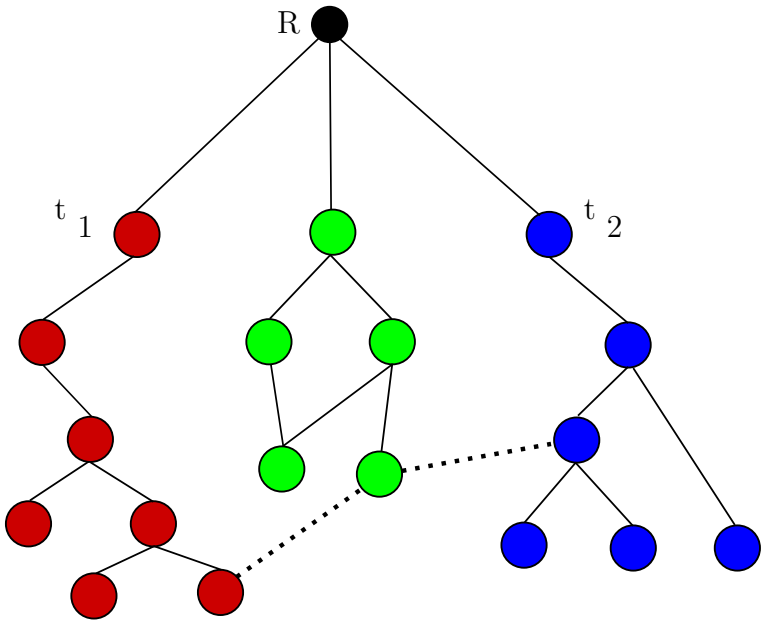

Figure 4.4: Nodes labelled t_1 and t_2 show the need for the fictitious root node labelled **R**. Should node **R** not exist, the similarity of nodes t_1 and t_2 would not be defined.

set-dependent or lead to misinterpretations (*e.g.* values constrained between zero and one might be wrongly interpreted as probabilities).

It is important to note that Publication Characteristics [V]does not annotate any diseases in OMIM and can therefore be omitted.

4.3 Evaluating disease similarity measures

The network medicine principles introduced in chapter 2 provide the rationale for the evaluation of disease similarity measures. For example, we know that the interactors of known disease-associated proteins tend to also be involved in the disease [95]. Thus, the evaluation follows the premise that similar diseases are close in the interactome [54] and a phenotype similarity measure should be able to accurately quantify the distance between disease modules on the interactome [64].

I use a machine learning approach to show that the measure I propose is capable

of accurately predicting molecular-level relationships between diseases. In addition to the numerical analysis, I curate a highly illustrative set of examples from recent medical literature. These examples showcase the method's capability to accurately quantifying phenotype similarity between diseases that are reported to be molecularly related. I will also show that the measure is able to accurately group diseases based on the affected physiological system. Lastly, by comparing the current version of OMIM with a two-year old release, I show the potential of the measure to produce candidate disease genes.

4.4 Definition and construction of the evaluation datasets

The evaluation of the similarity measures follows the approach used by van Driel *et al.* [64] who proposed to quantify the molecular level similarity between diseases using three relationships between their disease proteins. The first relationship determines molecular relatedness based on protein-protein interactions between disease proteins. The second relationship is based on the co-occurrence of Pfam-A signatures (*i.e.* families, domains, motifs or repeats), and it relates two diseases if any of their disease proteins share at least one of these signatures. The last relationship proposed by van Driel *et al.* is based on sequence similarity, and it relates two diseases whenever their disease proteins are similar in sequence.

Each criteria produces a binary dataset that relates diseases at a molecular level. The evaluation is thus reduced to a binary classification problem, where disease similarity scores are used to predict these binary relationships. The performance of the measure is evaluated by computing the area under the ROC curve (AUC). Finally, I include coverage in the evaluation, defined as the percentage of OMIM diseases for

which similarities can be computed.

For the results analysed in this work, the July 21 2014 release of OMIM was used. This release contains 23,611 records of which 7,812 correspond to Mendelian diseases. The diseases in OMIM reference a total 62,829 publications. The 2014 release of MeSH contains 27,149 terms, or which 13,220 were referenced by 62,393 of the publications referenced by the OMIM diseases.

The Pfam dataset

The first relationship proposed by van Driel *et al.* is based on the co-occurrence of Pfam-A signatures, that is families, domains, motifs or repeats. Pfam-A [76] is a database of curated protein families and as such provides a grouping of functionally related proteins that allows the association of diseases at molecular level based on structural characteristics of their disease proteins. Should proteins share structural features, then a mutation perturbing these features will result in similar phenotypes [64].

After manually verifying the content of the MeSH ontologies it became apparent that certain MeSH terms correspond to Pfam signatures. To avoid any bias in the evaluation, the MeSH ontologies were curated to extract all MeSH terms that describe Pfam-A signatures. This automatic curation process, followed by manual verification, resulted in the 113 descriptors shown in Table 4.2. Disease pairs in which a protein's Pfam signature matched any of the ones listed in Table 4.2 were excluded from the evaluation. After filtering, 33,660 pairs relating 2,647 OMIM diseases were evaluated.

The Protein-protein interaction dataset

The second relationship presented by van Driel *et al.* determines molecular relatedness based on protein-protein interactions between disease proteins. This is perhaps the most literal interpretation of the disease module, as it directly relates to evidence

MeSH term	Name	MeSH term	Name
D001081	Apyrase	D013049	Spectrin
D005294	Ferrochelatase	D002364	Casein
D050600	Snare	D051348	Tropomodulin
D043169	Endostatin	D002155	Calsequestrin
D004815	EGF	D005914	Globin
D014168	Transferrin	D005801	Homeobox
D015847	IL4	D014357	Trypsin
D017370	IL11	D009320	ANP16
D001119	Arginase	D035561	TFIIA
D018664	IL12	D000519	Melibiase
D016596	Vinculin	D046988	Proteasome
D064451	Hepcidin	D051152	Clusterin
D006466	Hemopexin	D053523	Amelogenin
D016547	Kinesin	D050683	Synaptobrevin
D001839	Bombesin	D003094	Collagen
D052116	Endomucin	D005293	Ferritin
D018793	IL13	D037282	Calreticulin
D018969	IGFBP	D025801	Ubiquitin
D016173	CSF-1	D056489	Nucleoplasmin
D052243	Resistin	D013884	Rhodanese
D005420	Flavoprotein	D019409	IL15
D053673	Glypican	D014216	TAN
D013004	Somatostatin	D008049	Lipase
D014559	Urocanase	D019922	Neuromodulin
D035581	TFIIB	D025481	6PF2K
D013879	Thioredoxin	D013947	Thymosin
D014598	Uteroglobin	D054477	Glutaredoxin
D014442	Tyrosinase	D064248	Geminin

Table 4.2: MeSH terms names matching Pfam-A families, domains, repeats or motifs.

that similar diseases tend to have interacting proteins [54].

According to the Protein-protein interaction (PPI) dataset two diseases are related if a physical interaction between any of their disease proteins is reported in Human Protein Reference Database (HPRD) [93]. This criterion resulted in $15,515$ disease pairs relating $2,512$ OMIM diseases.

The Sequence Similarity dataset

The last relationship proposed by van Driel *et al.* is based on sequence similarity, and it relates two diseases whenever any of their disease proteins are similar in sequence. This criteria is based on the observation that disease proteins associated to similar diseases tend to be functionally similar [42] and also tend to compress [54], enabling the use of sequence similarity measures as a proxy of functional associations between disease proteins.

The construction of the Sequence Similarity dataset is based on a Smith-Waterman local alignment of the sequences with a threshold e-value smaller or equal to $10^{-}6$. This criterion results in $37,486$ diseases pairs relating $2,817$ OMIM diseases.

Coverage

The evaluation also included coverage, defined as the fraction of OMIM diseases for which similarities can be computed. This is particularly important when considering that 27% $(5,519)$ of the diseases in OMIM have no known molecular basis as of June 2015. The coverage of a measure thus becomes an indicator of its capability to locate the module of an orphan disease on the interactome in the absence of molecular information.

My method covers $7,575$ OMIM diseases corresponding to 96.9% of the $7,812$ diseases in OMIM. The shortfall in coverage is due to the lack of MeSH annotations for some OMIM diseases, a situation that arises in two situations: *a*) either no publications were found to be associated to a particular OMIM disease, or, *b*) the publications associated to the diseases have no MeSH terms.

Those diseases with no associated publications either do not reference any publications (*e.g. Fragile Site 20p11* (`MIM: 136590`)), or the publications could not be retrieved through PubMed's API interface. In some exceptional cases, the PubMed

identifier referenced in OMIM was an invalid. This was the case of *Myofibrillar Myopathy* (`MIM:601419`) which at the time of querying OMIM, referenced the PubMed ID 10553984, which is non-existent. These cases were reported to the OMIM staff.

Out of the 62,829 PubMed identifiers available from OMIM, 62,393 are annotated with at least one MeSH term, the remaining 436 publications were either very new (*e.g. "Human CalDAG-GEFI gene (RASGRP2) mutation affects platelet function and causes severe bleeding"* - PubMed ID 24958846, published 2014) or very old (*e.g. "Some possible effects of nursing on the mammary gland tumour incidence in mice"* - PubMed ID 17793252, published 1936) and had no MeSH terms assigned at the time of querying the database. A few PubMed entries were found to be lacking MeSH annotations without apparent reason, such as as *"Neonatal Hyperinsulinism"* - PubMed ID 10322395, published 1999.

Comparison

The three evaluation datasets relate a similar number of diseases, however, the PPI dataset is comparatively sparser than the other two, as shown in table 4.3. This is due in part to the sparseness of the PPI networks themselves [66] and in part by the stringent evaluation criteria selected. However, it is important to note that the datasets overlap significantly.

			Overlap		
Dataset	Links	Nodes	Pfam	PPI	Sequence
Pfam	33,660	2,647	·	2,793	3,232
PPI	15,515	2,512	2,793	·	2,745
Sequence	37,486	2,817	3,232	2,745	·

Table 4.3: Topological characteristics of the evaluation datasets.

4.5 Numerical evaluation of the performance

In this section I will present the numerical evaluation of the similarity measure. First, the performance of the measure using terms from the individual ontologies, followed by the results of combining the ontologies. Lastly, I will compare the existing similarity measures with the one I propose.

4.5.1 Performance of the individual ontologies

Figures 4.5 to 4.19 present the performance of the proposed method in the MeSH ontologies that produced annotations. It is important to note that since the Publication Characteristics [V] ontology does not annotate any of the referenced publications in OMIM, it produces no similarity score.

In chapter 4 §4.1 I discussed the importance of the information contained in the MeSH ontologies that are not specifically related to diseases.

Importantly, the Chemicals and Drugs [D] ontology, is the best performing ontology, as can be seen in figure 4.8. This is remarkable specially when considering the highly specific sub ontology of the Diseases [C] ontology, *Pathological Conditions, Signs and Symptoms [C23]* (see figure 4.7) . The higher performance highlights the thorough nature of the descriptions in OMIM. The rich descriptions in OMIM are far wider and more detailed than the mere symptom.

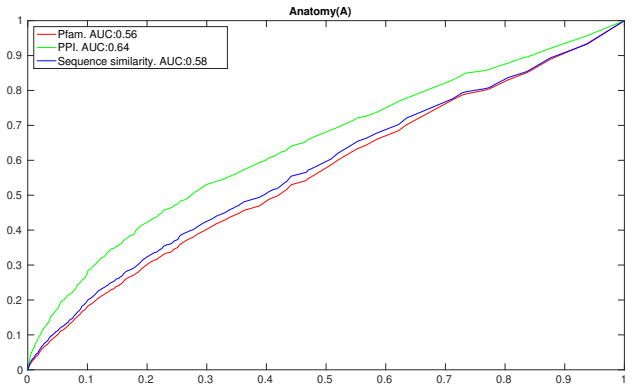

Figure 4.5: ROC curve [A] ontology

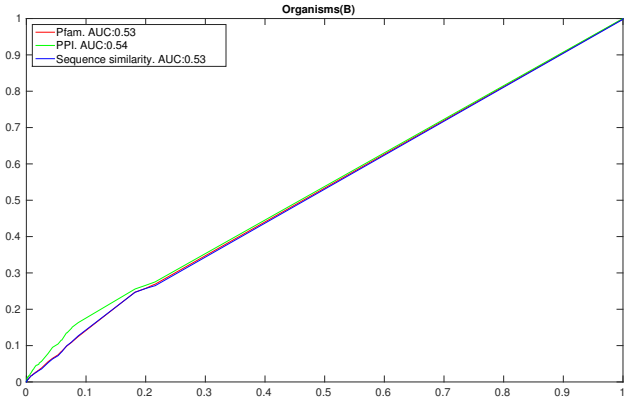

Figure 4.6: ROC curve [B] ontology

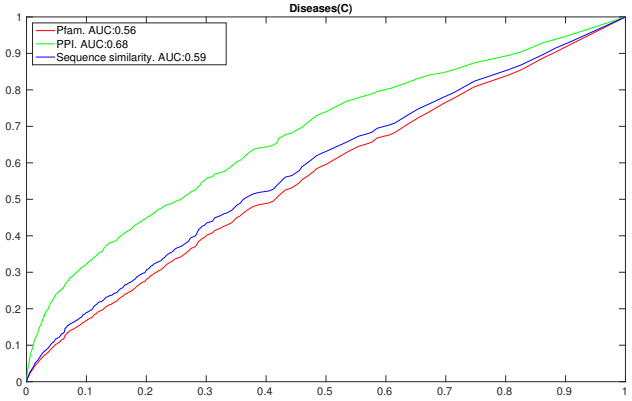

Figure 4.7: ROC curve [C] ontology

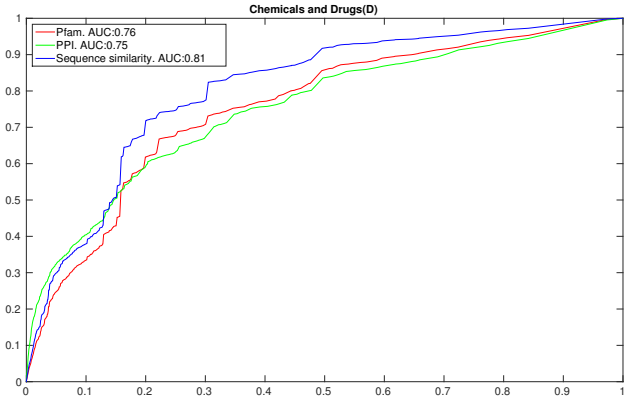

Figure 4.8: ROC curve [D] ontology

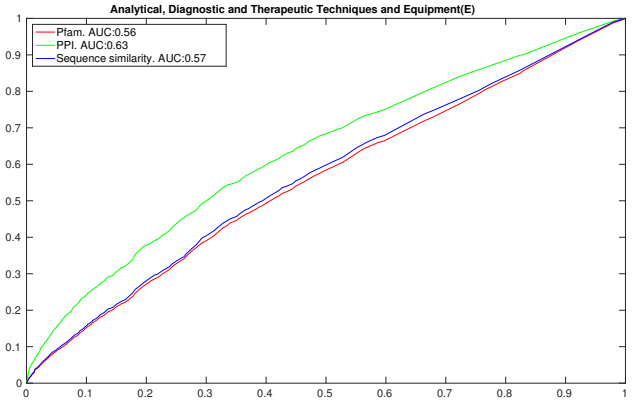

Figure 4.9: ROC curve [E] ontology

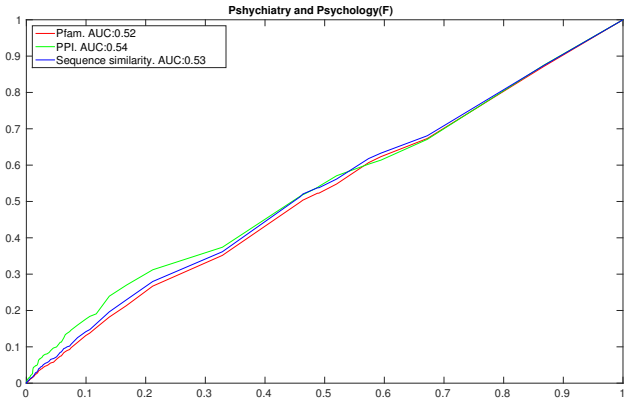

Figure 4.10: ROC curve [F] ontology

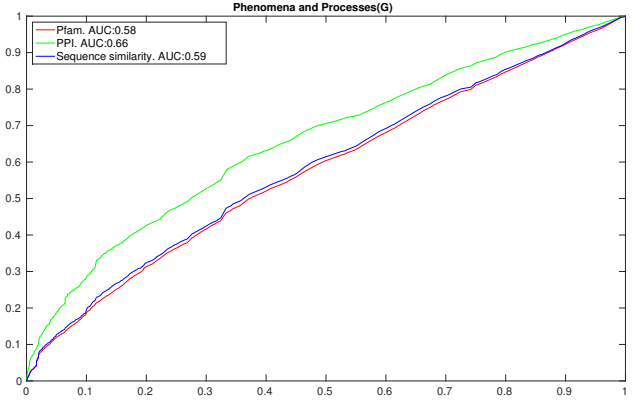

Figure 4.11: ROC curve [G] ontology

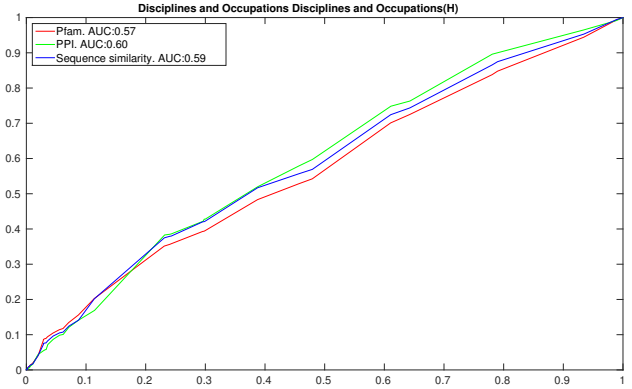

Figure 4.12: ROC curve [H] ontology

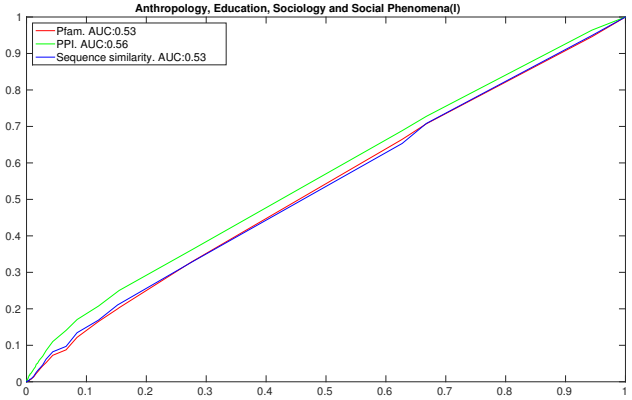

Figure 4.13: ROC curve [I] ontology

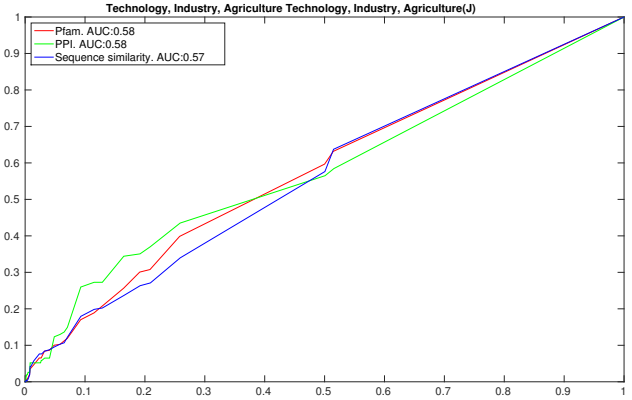

Figure 4.14: ROC curve [J] ontology

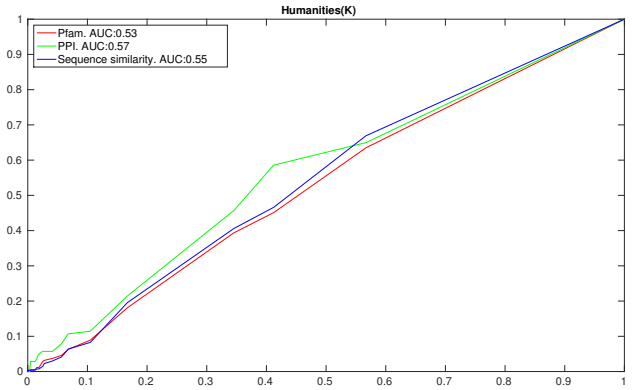

Figure 4.15: ROC curve [K] ontology

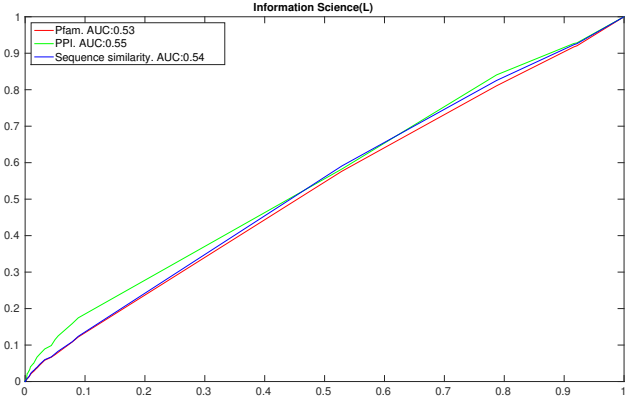

Figure 4.16: ROC curve [L] ontology

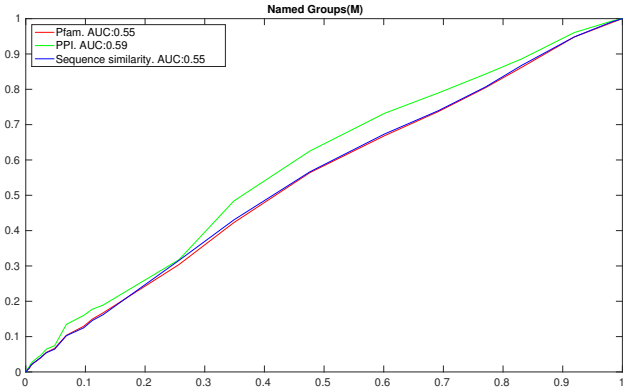

Figure 4.17: ROC curve [M] ontology

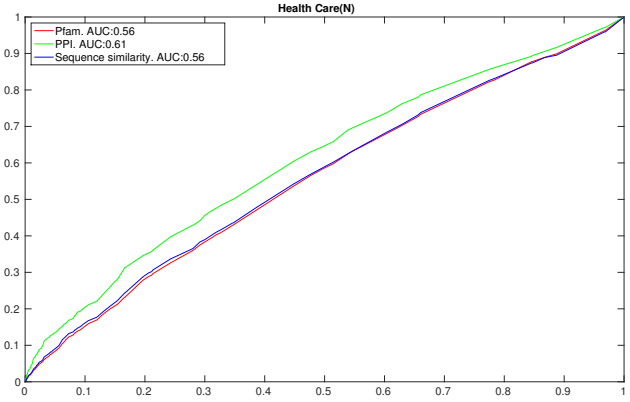

Figure 4.18: ROC curve [N] ontology

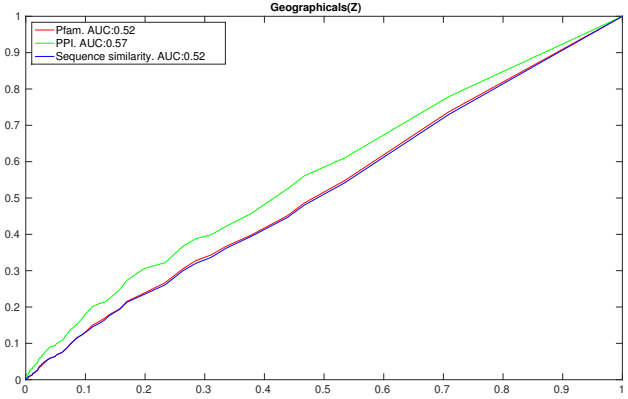

Figure 4.19: ROC curve [Z] ontology

4.5.2 Performance of the combined ontologies

While all ontologies in MeSH could be combined following the method discussed in § 4.1, I have chosen to discard the poor performing ontologies in order to keep the smallest possible set of terms needed to accurately define the similarities. Table 4.4 shows the coverage and AUC of the ROC curve for each ontology.

Ontology	Coverage	AUC Pfam	AUC PPI	AUC Sequence
Anatomy [A]	6,781	0.56	**0.64**	0.58
Organisms [B]	7,488	0.53	0.54	0.53
Diseases [C]	7,321	0.56	**0.68**	0.59
Chemicals and Drugs [D]	5,958	0.76	**0.75**	0.81
Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]	7,000	0.56	**0.63**	0.57
Psychiatry and Psychology [F]	3,271	0.52	0.54	0.53
Phenomena and Processes [G]	7,018	0.58	**0.66**	0.59
Disciplines and Occupations [H]	1,994	0.57	0.60	0.59
Anthropology, Education, Sociology and Social Phenomena [I]	1,903	0.53	0.56	0.53
Technology, Industry, Agriculture [J]	348	0.58	0.58	0.57
Humanities [K]	315	0.53	0.57	0.55
Information Science [L]	4,063	0.53	0.55	0.54
Named Groups [M]	6,775	0.55	0.59	0.55
Health Care [N]	4,257	0.56	0.61	0.56
Geographicals [Z]	2,834	0.52	0.57	0.52

Table 4.4: The 16 MeSH ontologies. The coverage of each ontology is calculated by the diseases annotated with at least one of its terms. The AUC is the Area Under the ROC curve. See figures 4.5 to 4.19 for a graphical representation of the ROC curves.

The results presented in figure 4.20 are obtained using the ontologies which had an AUC above 60% (shown in boldface in table 4.4 for the PPI dataset while maintaining a high coverage of OMIM diseases. The combined ontologies are: Anatomy [A], Diseases [C], Chemicals and Drugs [D], Analytical, Diagnostic and Therapeutic Techniques and Equipment [E] and Phenomena and Processes [G]. I have analysed the

combination various other subsets of ontologies and results were found to be similar as long as ontologies with high coverage were included.

The reason behind the choice of the PPI dataset as a decision criteria for the combination of the ontologies can be seen in figure 4.20, particularly when comparing the Sequence Similarity and PPI datasets. The PPI curve (blue), grows more sharply in the bottom left quadrant of the ROC plot, which means that the measure issues a positive classification when the evidence is strong, *i.e.* the measure is more conservative [90]. Conversely, in the Sequence Similarity dataset (green curve), the measure is more liberal, issuing positive classifications when the evidence is not as strong [90].

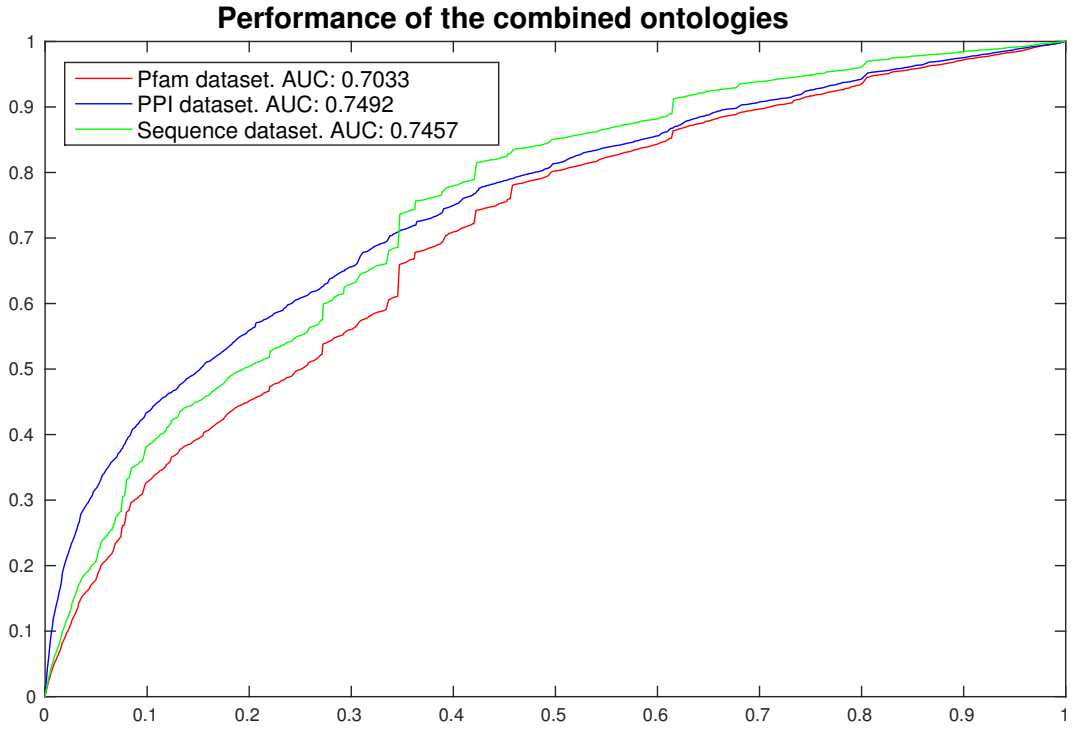

Figure 4.20: Performance evaluation of the semantic similarity method on the combined ontologies. Each ROC represents the predictive power of the semantic similarity method on the Pfam, PPI and Sequence dataset respectively. The combined ontologies are Anatomy [A],Diseases [C],Chemicals and Drugs [D],Analytical, Diagnostic and Therapeutic Techniques and Equipment [E] and Phenomena and Processes [G].

4.5.3 Comparing with the existing measures

If the following, I will detail the mappings built for each method, in order to compare them with the disease similarity method I introduce in this Thesis.

Goh *et al.*

Due OMIM's continuous update policy, it was impossible to retrieve the same dataset used by the authors at the time of publication. However, through a reverse engineering process, a total of 1,717 syndromes were retrieved from Goh's *et al.* syndromes.

In their curated morbidmap Goh *et al.* provide the mapping between diseases in OMIM and syndromes. While most entries were complete and it was possible to obtain the OMIM numbers comprising the syndrome, some entries were incomplete. To maximise the coverage of this reverse engineered similarity measure, a fuzzy string matching procedure followed by manual curation was performed.

The name each incomplete entry was compared to the names of all entries in a current morbidmap (21 July 2014) obtained from OMIM. A fuzzy string matching procedure matched these incomplete entries to all entries in the newer morbidmap whose Levenshtein string similarity ratio was higher than 0.9. If more than one entry satisfied the matching cut-off value, the entry was discarded. The matched entries were verified manually, thus obtaining 1,717 OMIM diseases extracted from the Goh's *et al.* curated morbidmap. Of these 1,717 OMIM diseases, 1,542 were directly matched and 175 where extracted through the string matching procedure.

From these 1,171 OMIM diseases, a binary similarity matrix was constructed, in which diseases were represented as either similar or not similar. Similarity between diseases is given by the physiological category of the syndrome from which the disease was extracted. If two diseases belong to syndromes with identical category, those diseases were considered similar.

Park *et al.*

Following the reverse engineering process I designed for Goh's *et al.* Diseasome, the syndromes were mapped onto their constituent OMIM diseases, each of which was assigned the subcellular profile of the syndrome it was extracted from. Since the authors provide the original Disease-associated Protein and Subcellular Localization (DPL), obtaining the correlation between the disease localisation profiles becomes a trivial exercise. This allows the construction of a square, real-numbered matrix bounded between 0 and 1 of similarities between the 1.177 OMIM diseases.

van Driel *et al.*

van Driel *et al.* [64] have made the similarity matrix between all diseases listed in OMIM at the time of publication in 2006 (5,132) available from their website `http://www.cmbi.ru.nl/MimMiner/suppl.html`. To produce updated similarity scores we contacted Prof. Han G. Brunner (Radboud University Nijmegen Medical Centre, Department of Human Genetics) who provided the original scripts used to produce the similarity matrix. The scripts required some fixes to adapt to newer versions of the libraries and operating system and after these fixes were applied they were used to compute up-to-date similarity scores.

To compare van Driel's *et al.* method there was no mapping required, as they provide a square similarity similarity matrix for all OMIM diseases.

Köhler *et al.*

Köhler *et al.* provide a matrix of similarity as defined in [81]. There is no need to map the dataset provide by the authors, as it already relates OMIM diseases to one another. The Human Phenotype Ontology (HPO) similarity dataset contains a total of $6,441$ OMIM diseases in the October 2014 release.

Zhou *et al.* and Mathur and Dinakarpandian

In the case of Zhou [106] *et al.* and Mathur and Dinakarpandian [82] I was unable to compare the performance of the similarity scores.

In the case of Mathur and Dinakarpandian, the authors did not have similarity scores available. From the description in the publication, I was unable to accurately replicate the method.

Zhou *et al.* provide a full similarity matrix, however, they do no provide a mapping between OMIM diseases and MeSH diseases. To the best of my knowledge, no such mapping exists, and therefore I decided to map the diseases myself. I first attempted to map the MeSH diseases to the OMIM diseases by approximately matching the MeSH disease names and OMIM disease names calculating the Levenshtein distance, and considering names with matching ratio higher than 90% to be identical. After manually verifying the mapping I concluded that the method produces an excessive number of false mappings. In a second attempt, I matched the MeSH disease terms to the OMIM diseases through the Disease Ontology (DO) [100]. Several DO entries contain a cross-reference field, which matches the DO term with corresponding entities in alternative databases. MeSH terms were matched to OMIM diseases based on co-occurrence in the same cross reference field in the DO. Unfortunately, this mapping had a very low coverage, resulting in only 454 OMIM diseases being mapped.

I am, therefore unable to present a comparison between either Zhou's *et al.* and Mathur's and Dinakarpandian's method.

Performance comparison

The similarity scores obtained through the mapping described before were evaluated using the tree relationships proposed by van Driel *et al.* [64]. Figure 4.21 presents a comparison between the proposed method and the approaches by Park *et al.* [86], Goh

[54] *et al.* , van Driel *et al.* [64] and Köhler *et al.* [81]. Both larger AUC values and larger coverage are better, and since these scores are all bound between 0 and 1, they are combined into a composite score to compare the methods' overall performance, following the approach presented in [91].

To show the performance of the method without analysing coverage, figures 4.22, 4.23 and 4.24 shows the ROC curves for each evaluated method.

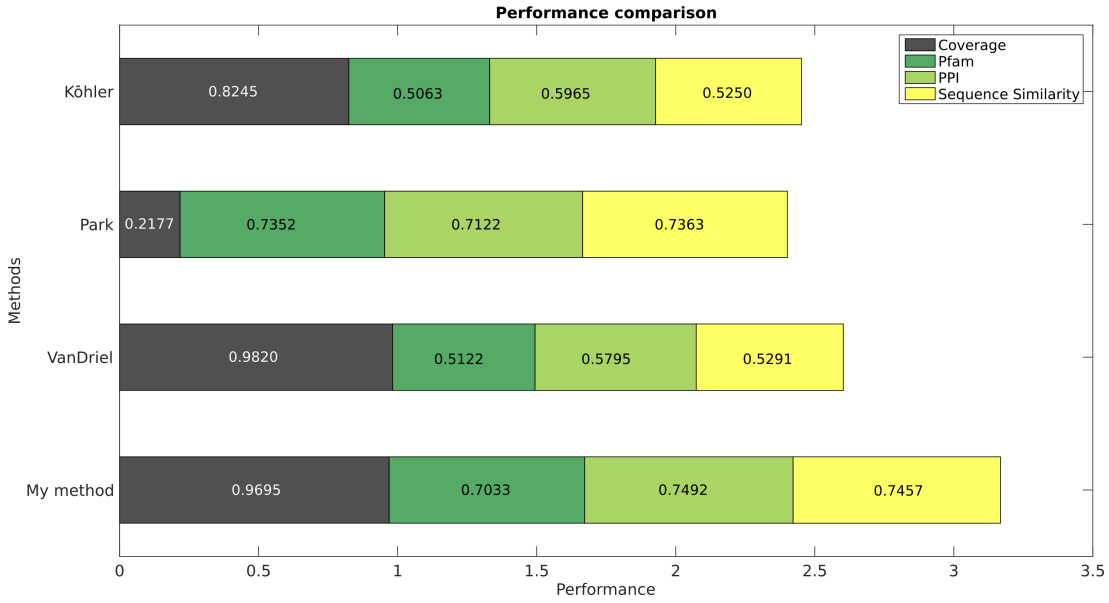

Figure 4.21: Performance Comparison. For each method, the grey bar quantifies its OMIM coverage, coloured bars quantify its performance measured by AUCs on the Pfam, PPI and Sequence Similarity datasets. The total length of each bar represents the overall performance of each method.

4.6 Verifying the correlation with molecular level similarity

To further assess the correlation of the similarity measure I propose with the molecular level similarity of the diseases, a contrast between the distribution of similarity scores

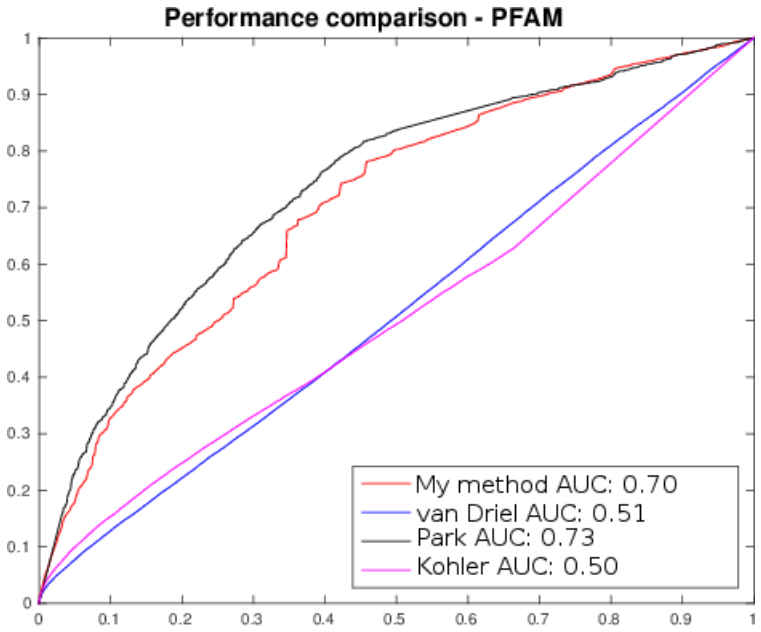

Figure 4.22: ROC plot of the performance of proposed method with the combined ontologies evaluated on the Pfam dataset

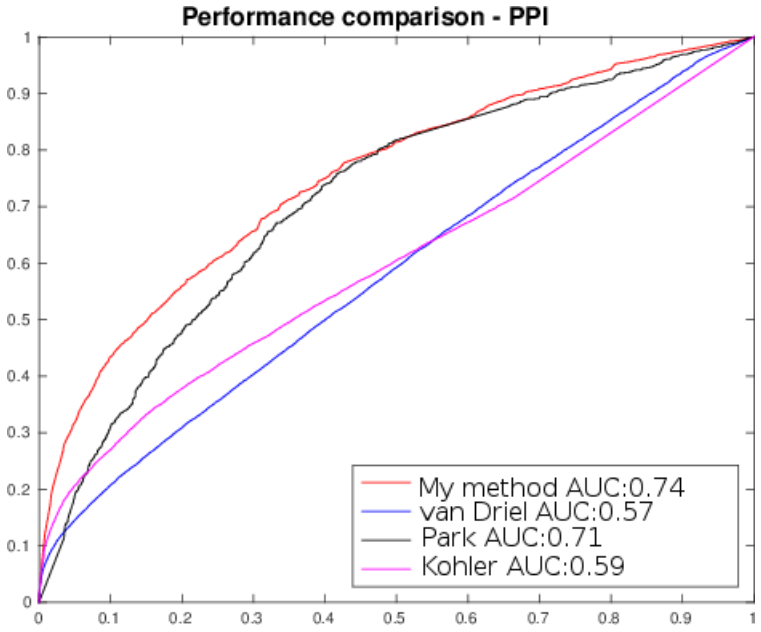

Figure 4.23: ROC plot of the performance of proposed method with the combined ontologies evaluated on the PPI dataset

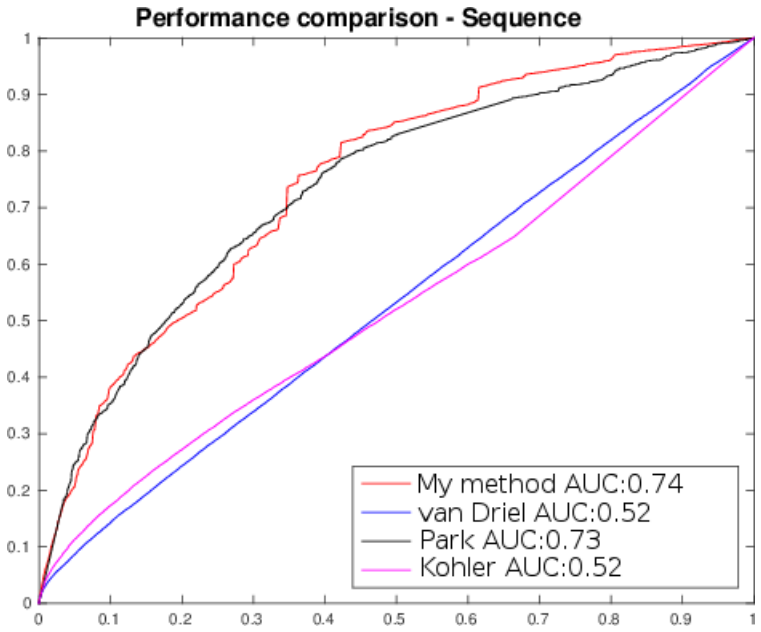

Figure 4.24: ROC plot of the performance of proposed method with the combined ontologies evaluated on the Sequence Similarity dataset

for all pairs of diseases with that of the subset of pairs sharing disease genes is shown in figure 4.25. The two distributions are very different (Student's t-test $P < 10^{-350}$). Interestingly, 90% of the pairs of diseases with shared genes have high-similarity scores (99th percentile or higher), indicating that high-similarity values are correlated with existing knowledge of relatedness at molecular level.

4.6.1 Assessing the measures ability to predict molecular similarity

For many disease pairs with high similarity score, it is readily verified that they are indeed similar at molecular level by analysing existing medical literature. For example, the score between *Budd-Chiari* (`MIM:600880`) syndrome and *Myeloproliferative disorder* (`MIM:131440`) is in the 97th percentile and genes associated to these diseases have in vivo verified first-level interactions (*JAK2* – *PDGFRB*) [93]. Furthermore, it

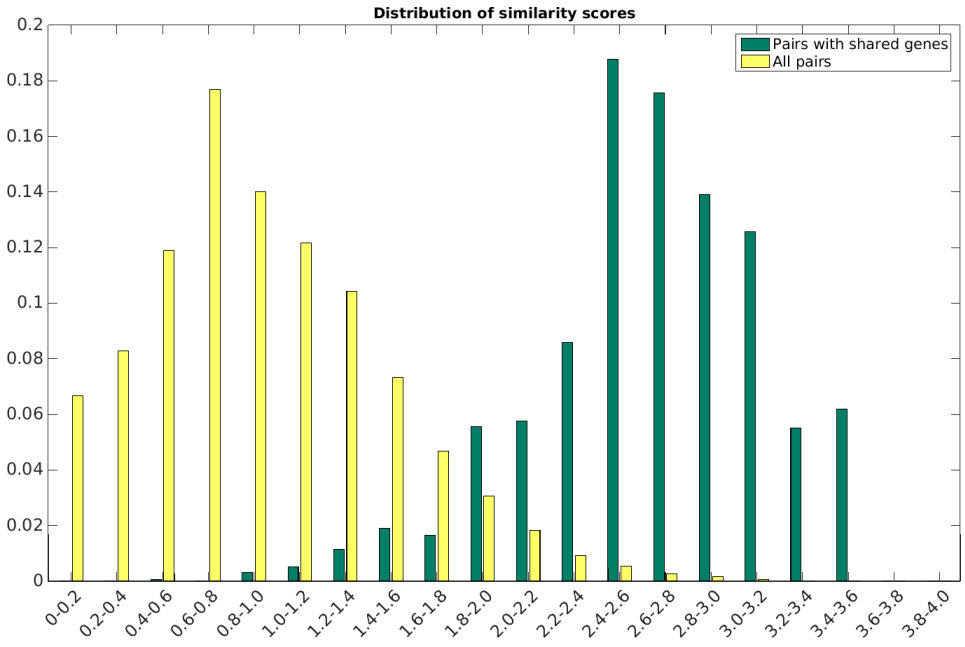

Figure 4.25: Distribution of similarity scores for all pairs of diseases (yellow bars) vs. distribution of similarity scores for disease pairs sharing one or more disease genes (green bars). 90% of the pairs of diseases with shared genes have scores in the 99th percentile or higher.

is known that these two diseases are causally related [40].

The score between *Breast Cancer* (`MIM:114480`) and *Noninsulin-dependent Diabetes (NDDIM)* (`MIM:125853`) lies in the 100th percentile, and several cancer related proteins are known to interact with NDDIM related proteins (*TP53* - *HNF4A*, *CDH1* - *PTPN14*, *CDH1* - *IRS1*) [93]. Moreover, there exists statistical evidence of increased risk of Breast Cancer in Women with type 2 diabetes [53].

The similarity scores between *Type I von Willebrand disease (VWD1)* (`MIM:193400`) and *pseudo von Willebrand disease (VWDP)* (`MIM:177820`), two bleeding disorders, lies in the 100th percentile. VWD1 is a consequence of exceptionally low levels of plasma von Willebrand Factor (VWF) [47], while VWDP is characterised by subtle mutations in the alpha subunit of the glycoprotein Ib (*GPIb*α) subunit, causing it to bond uncharacteristically to VWF [87].

4.6.2 Embedding diseases in 3d space

The measure can be used to produce a 3D graphical representation of human diseases automatically. Figure 4.26 top shows the embedding of diseases into 3D space obtained applying t-SNE [58] a recently developed dimensionality reduction technique.

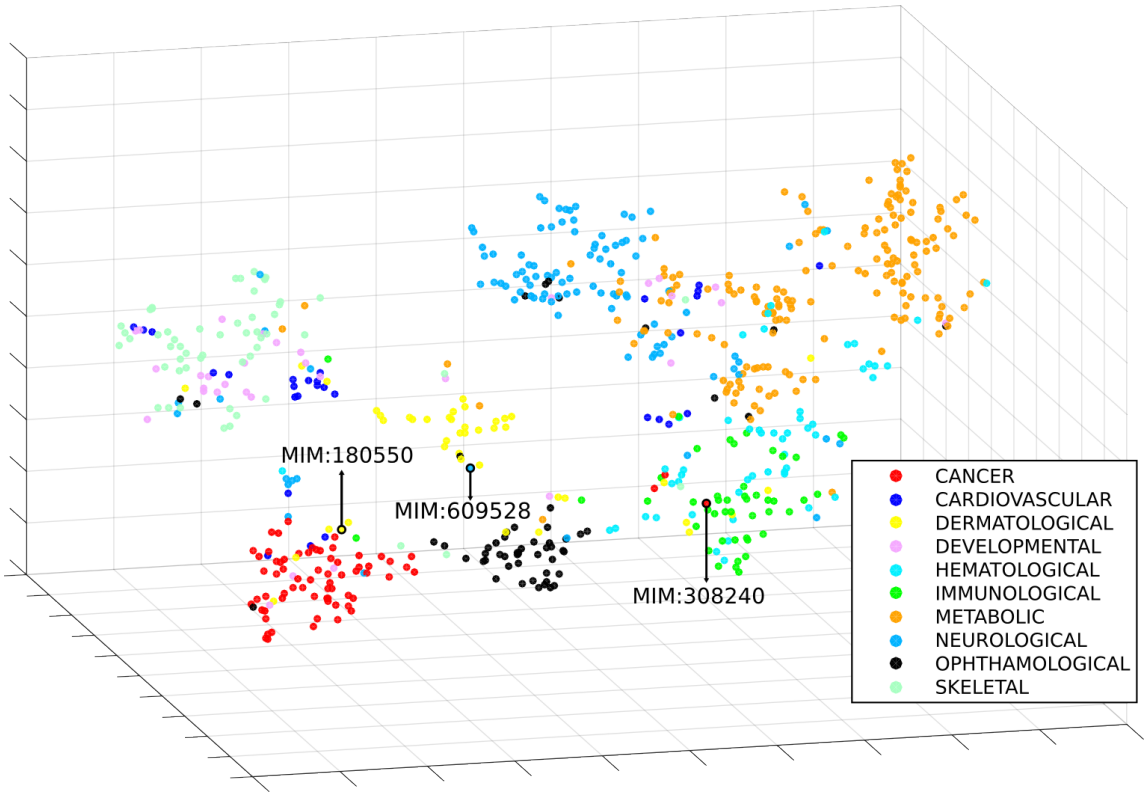

Figure 4.26: Embedding of hereditary diseases in 3D space using t-SNE. Each point represents an OMIM disease. Colours are assigned based on their disorder class according to Goh *et al.* [54]. Highlighted diseases belong to multiple phenotypic classes and are discussed in the main text. The figure shows the diseases belonging to the 10 most numerous disease classes in Goh *et al.* [54].

In the figure, each point corresponds to a disease and the distance between two diseases relates to their similarity according to the proposed method. Colouring each disease according to the disease classes identified in Goh *et al.* [54] reveals that diseases in the same class tend to be grouped together. The 10 most numerous classes are

shown in the figure (see Chapter 2 § 2.4) This is interesting, as Goh *et al.* show that these classes group diseases that are highly related at molecular level.

Notice how some diseases which, from a phenotypical perspective belong to multiple classes, are placed appropriately at the boundaries between them (see diseases pointed by arrows in Figure 4.26. For example the *Ring dermoid of Cornea* (`MIM:180550`), is located at the boundary between the Dermatological, Cancer and Ophthalmological classes. This disease is characterised by dermoids (growths with a skin-like structure) in the eye. In general, dermoids exhibit known hallmarks of cancer [24]. *Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome* (`MIM:609528`) is characterised by severe neurological impairment as well as keratoderma and late-onset ichthyosis. The embedding places this disease at the boundary between the Neurological and Dermatological classes. In other cases, diseases that belong to more than one class are placed closer to a class different from the one chosen by Goh *et al.* , but their position is overall appropriate when considering the diseases' characteristics. For example, *Lymphoproliferative syndrome, X-linked, 1* (`MIM:308240`), exhibits both immunological and cancer features. It is characterised by severe immunological dysregulation, and is related to several phenotypes (including lymphoma) and often occurs after an infection (Epstein-Barr virus). The embedding places this disease closer to immunological diseases than to the cancer group.

The clear grouping of diseases is made possible by the difference between average inter- and intra- class similarity values, visualised as a heat map in Figure 4.27. Note that pairs of classes with high average inter-class similarity contain diseases which are often related. For example, this can be the case for diseases in the immune and respiratory classes as it is known that an abnormal immune response can cause chronic respiratory diseases [60].

One important thing to note is that figure 4.26 is the result of a computational method that depends exclusively on the pairwise similarity between the diseases. I

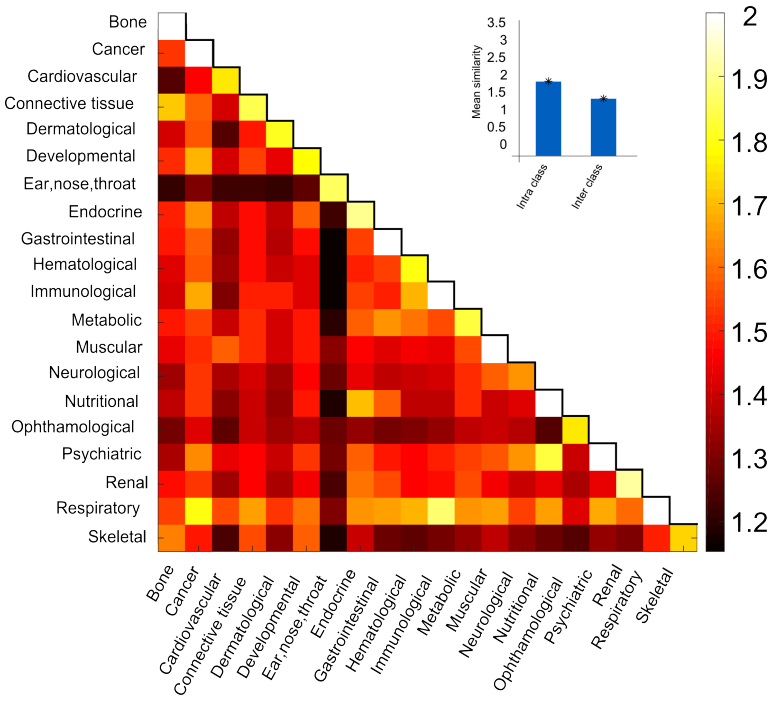

Figure 4.27: Each (x,y) tile represents, for the disease classes in Goh *et al.* the mean similarity of disease pairs where one disease belongs to class x and the other to class y. The values range from 1.15 (Gastrointestinal – Ear, nose, throat) to 2.71 (Nutritional-Nutritional). The colours range between the minimum mean similarity and 2, with all values above 2 (In the diagonal: 2.01 Bone, 2.05 Immunological, 2.06 Gastrointestinal, 2.07 Muscular, 2.1 Psychiatric, 2.2 Cancer, 2.5 Respiratory, 2.71 Nutritional) set to 2. Inset: the average intra-class similarity is significantly higher than the average inter-class similarity (t-test p-value $\leq 10^{-350}$).

did not intervene in any way to produce the figure. Nevertheless, comparing 4.26 with Goh's *et al.* diseasome (reproduced from [54] in figure 2.4) the grouping of the classes is remarkably similar.

4.7 Candidate disease genes prediction

Evidence in the literature proving the molecular relatedness of diseases with high similarity scores illustrates the measure's power in predicting the molecular relatedness between two diseases. However, the transfer of knowledge between diseases is perhaps the most important use of a disease similarity measure. That is, an accurate measure should also be able to help provide candidate disease genes.

To assess the effectiveness in providing candidate diseases "old" similarity scores were calculated using an older version of OMIM (downloaded on April 9th 2013). Based on these calculations, several pairs of diseases which had high similarity values according to this old data from 2013, have since been shown to be close on the interactome.

For example, the proposed method reported (using the 2013 version of OMIM) no disease genes for SHORT syndrome (`MIM:269880`), *Dermatofibrosarcoma protuberans* (`MIM:607907`) and Right Atrial Isomerism (`MIM:208530`). However, the similarity scores indicated SHORT syndrome to be very similar at molecular level to *Noninsulin-dependent Diabetes Mellitus* (`MIM:125853`) (99th percentile), thus suggesting that disease genes for SHORT syndrome could be located in the neighbourhood of Diabetes. This is indeed the case, as the new version of OMIM links SHORT syndrome to gene PIK3R1, which has a verified in vitro interaction with IRS1, a gene associated to Noninsulin-dependent diabetes. Interestingly the publication identifying the association of PIK3R1 to SHORT syndrome namely "PIK3R1 Mutations Cause Syndromic Insulin Resistance with Lipoatrophy" [19], was published in July 2013, postdating the OMIM data used. For a comparison of the referenced publications in the current version of OMIM and July 2013 version, please refer to B.

Similarly, the "old" similarity scores indicated Dermatofibrosarcoma to be very similar at molecular level to *Juvenile Myelomonocytic Leukaemia* (`MIM:607785`) (100th

percentile). The current version of *omim* shows an association between Leukaemia and the gene *PDGFRB*, which interacts with *PDGFB* a gene associated to Dermatofibrosarcoma.

Lastly, the "old" score between Right Atrial Isomerism and *Tetralogy of Fallot* (`MIM:187500`) is in the 100th percentile and now it has been shown that they share a disease gene (*GDF1*).

Chapter 5

Discussion on the factors that affect the performance of the disease similarity measures

In this chapter I will analyse the suitability of the various semantic similarity measures presented in Chapter 1 for calculating disease. I will show how the structure of the MeSH ontologies improves the accuracy of the disease similarity calculations. I will show that not all semantic similarity measures introduced in chapter 1 perform equally well, showing that the correct use of the ontology is essential for accurately quantifying disease similarity. Finally, I will show how the correct use of the MeSH terms to annotate the OMIM diseases has a significant impact on the performance of the similarity measures.

5.1 Using the MeSH ontological structure improves the accuracy of disease similarity calculations

To verify the importance of MeSH's ontological structure, I have implemented several overlap-based similarity measures, both based on the overlap of publications and the overlap of MeSH terms. Exploring the results obtained measuring the overlap of the publications associated to the diseases it becomes apparent that the publications are not enough to accurately quantify disease similarities. Conversely the overlap between the sets of MeSH terms annotating the diseases shows that exploiting the structure of the MeSH ontology is essential to accurately quantify similarity between diseases at molecular level.

The simpler, overlap-based similarities are given by several the following measures:

- Jaccard: The similarity of two diseases $sim\,(a,b)$ is given by the Jaccard coefficient of their respective annotation sets. Formally:

  $sim\,(a,b) = \frac{|Annot(a) \cap Annot(b)|}{|Annot(a) \cup Annot(b)|}$

- Dice:The similarity of two diseases $sim\,(a,b)$ is given by the Sørensen-Dice coefficient of their respective annotation sets. Formally:

  $sim\,(a,b) = \frac{2*|Annot(a) \cap Annot(b)|}{|Annot(a)||\cup Annot(b)|}$

- Overlap: The similarity of two diseases $sim\,(a,b)$ is given by

  $sim\,(a,b) = \frac{2*|Annot(a) \cap Annot(b)|}{min(|Annot(a)|,|Annot(b)|)}$

- Common: The similarity of two diseases $sim\,(a,b)$ is given by the size of he intersection of their annotations. Formally $sim\,(a,b) = |Annot\,(a) \cap (b)|$

5.1.1 Measuring the overlap of publications

Measuring the overlap of the publications rests on the idea that similar diseases are described in overlapping groups of publications. The size of the overlap would, therefore, accurately quantify similarity between the diseases. The performance of the simple measures is shown in Figure 5.1.

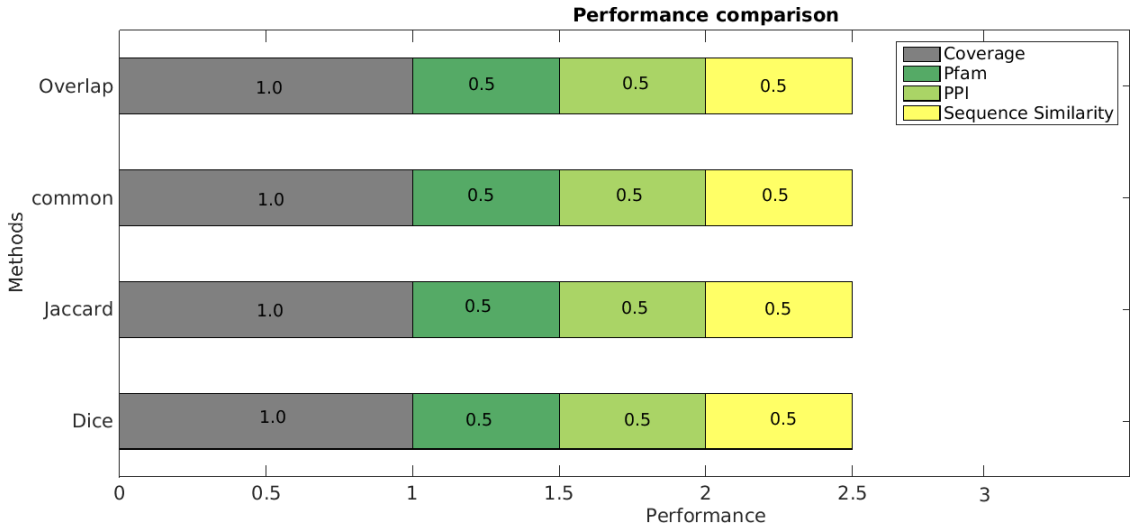

Figure 5.1: Performance of the simpler, overlap-based similarity measures. Each bar shows the combined AUC of the ROC curves on the Pfam, PPI and Sequence Similarity datasets. The pairwise disease similarities were calculated measuring the overlap of publications referenced by the OMIM diseases.

The coverage of this simple overlap measures is high, since most diseases in OMIM reference at least one publication. However, the number of pairs of diseases with at least one common publication is extremely low. From the $28,686,525$ possible pairs, only $8,757$ $(0,03\%)$ pairs of diseases share at least one publication. Furthermore, these pairs include only $4,114$ diseases, meaning that only about 48% of the diseases in OMIM are represented. The extremely low number of positive overlap pairs is the result of a very coarse measure that is only able to produce similarity between highly similar diseases. As is to be expected, maximum similarity is given to pairs

of diseases which represent variations of one syndrome, such as *Cowden Syndrome 5* (`MIM:615108`) and *Cowden Syndrome 6* (`MIM:614109`).

The reason so few diseases share a publication can be traced back to the bias that exists in the study of diseases [80]. Figure 5.2 illustrates this imbalance in the study of diseases, showing the number of publications each OMIM entry references. The majority of diseases (76%) references fewer than 10 publications and 99% of the OMIM records references fewer than 100 publications. The best referenced record is *Methemoglobinemia, Beta-Globin Type, Included* (`MIM:141900`) with 1,094 publications followed by *Methemoglobinemia, Alpha-Globin Type Included* (`MIM:141800`) with 387.

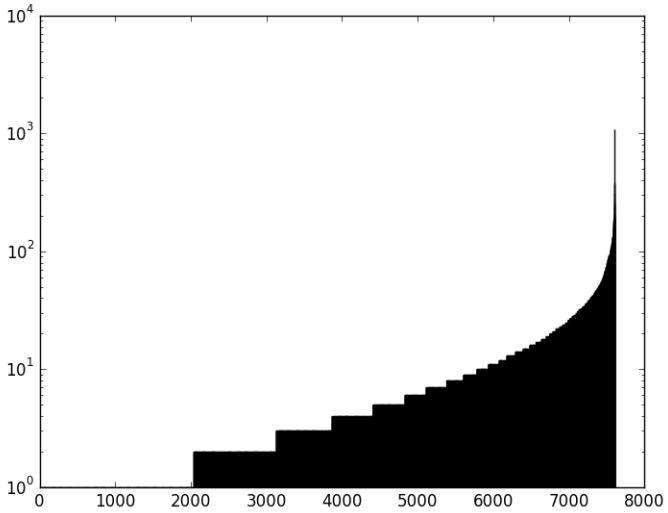

Figure 5.2: Number of referenced publications. The figure shows the number of publications (Y-axis) each OMIM disease (X-axis) references in increasing number of referenced publications. The Y-axis ranges from 1 to 10^3 in log scale. The disease with the most annotations references $1,094$ publications.

Importantly, the number of referenced publications does not necessarily correlate with the prevalence of the various diseases. While I have not performed a large-scale analysis a few examples are illustrative of this imbalance. Sickle-cell disease,

Methemoglobinemia, Beta-Globin Type, Included (`MIM:141900`), the diseases with the most references, affects around 100,000 United States citizens [20]. In contrast there are 29 OMIM diseases which have the word "heart" in the name. Collectively, these diseases reference a total of 585 publications, even though heart disease is much more prevalent than Sickle cell disease [26] and the leading cause of death in the developed world in general [104] and among the top in the UK (leading cause for men, second for women,[69]).

5.1.2 Measuring the overlap of MeSH terms

Based on the assumption that similar diseases would share a significant fraction of MeSH annotations, I measured the overlap of annotations using the simple similarity measures. Figure 5.3 shows a comparison of the performance.

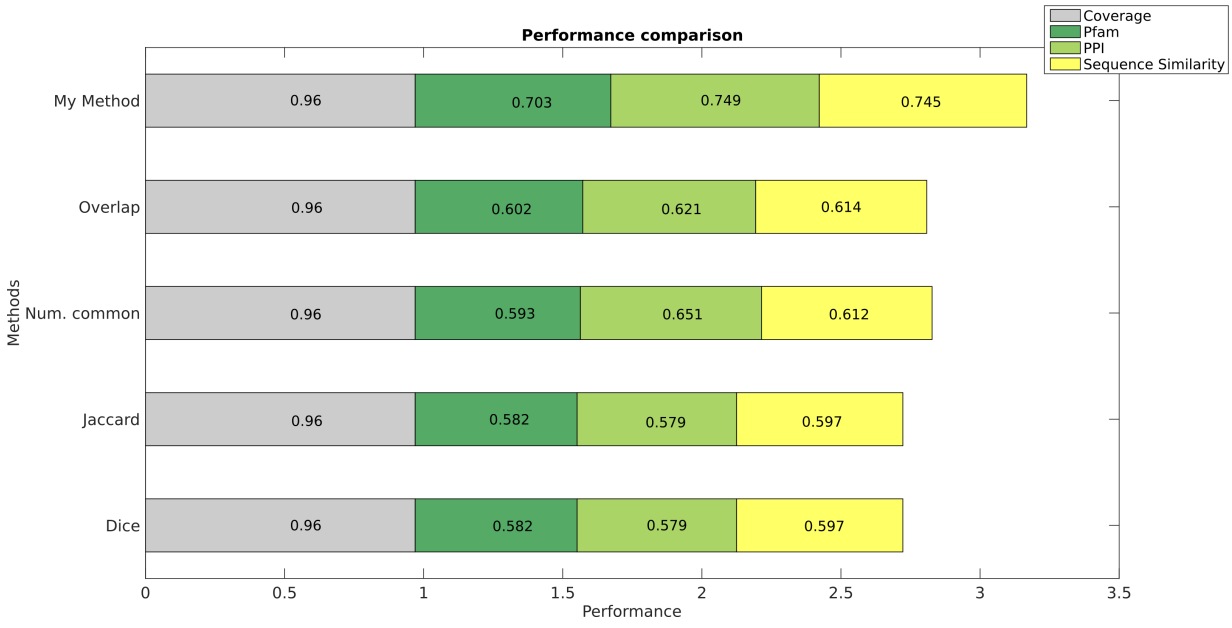

Figure 5.3: Performance of the simple similarity measures. The similarity was calculated using the various overlap measures of MeSH terms.

Due to the fact that the simple similarity measures do not consider the ontological

structure, the scores they produce do not depend on the specificity of the annotations. This results in measures with a reduced capability to discriminate between diseases with high-quality annotations and those annotated with several general but overlapping terms. As an example, 5 overlapping but very general terms (*e.g.* Elements - (D004602)) are as good as 5 very specific overlapping terms (*e.g.* Argon - (D001128)). Additionally, since the true path rule does not apply, terms following the path to the root of the ontology are not considered. This situation is similar to the one faced by van Driel *et al.* in [64]. As the authors noted, it is important to consider the relevance of the hypernyms of the actual terms found in a record [64].

These simple measures are generally coarse and are unable to discriminate between two pairs of *slightly* similar diseases. A coarse measure will heavily penalise dissimilar diseases and lesser-known diseases, as the information available for them might not be as broad or detailed. The nature of the domain, where information is uncertain and scarce, a very conservative measure would not be appropriate. Figures 5.4, 5.5 and 5.6 show the ROC curves for each dataset. Notice how these curves contrast sharply with those presented in figure 4.20, particularly in the region of $(0.3, 0.6)$ in the X and Y axis, where the curve is almost linear, showing the measure's reduced ability to accurately distinguish between the slightly similar instances [90].

It is important to note that the lack of overlap when analysing the overlap of both publications and MeSH terms, is not a definite measure of dis-similarity between diseases. It might simply reflect the lack of information, but a value of "zero" would at the same time represent a lack of information as well as dis-similarity.

In contrast, the ontology-based similarity measures (Resnik, Lin, Jiang and Conrath, simUI and simGIC, see 1) are only able to produce a similarity score of zero when the root of the ontology is chosen as the Lowest Common ancestor or, equivalently, when two diseases have no common terms along the path to the root. Conceptually, such a situation would arise only when two diseases are annotated with terms that

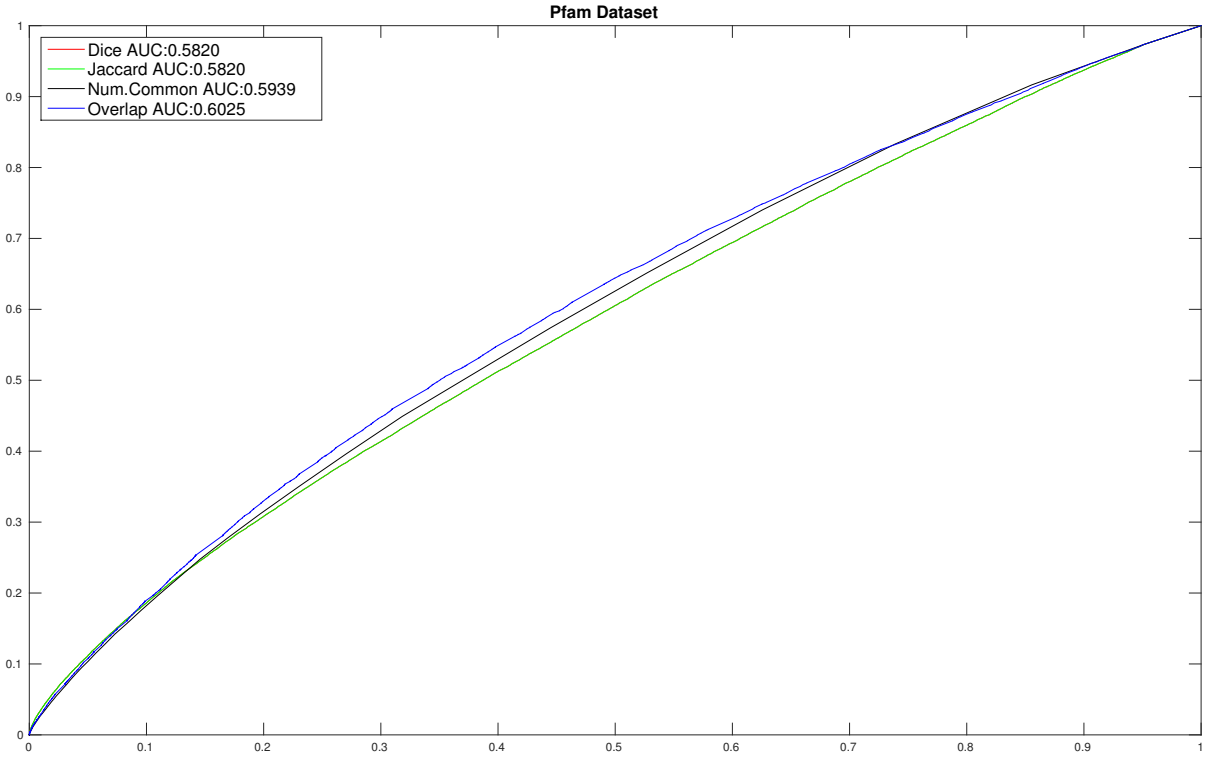

Figure 5.4: ROC curve of the simple similarity measures in the Pfam dataset.

are distant from one another in the ontology.

5.2 Correct use of the MeSH ontological structure is essential for accurate disease similarity calculations

To fully take advantage of the quality of the annotations, the ontology must be used appropriately. The studied semantic similarity measures however, do not perform equally well. Figure 5.7 shows a comparison of the semantic similarity measures evaluated, which shows Resnik's similarity measure outperforming all others. However, considering that Lin's [25] and Jiang and Conrath's [51] semantic similarity measures

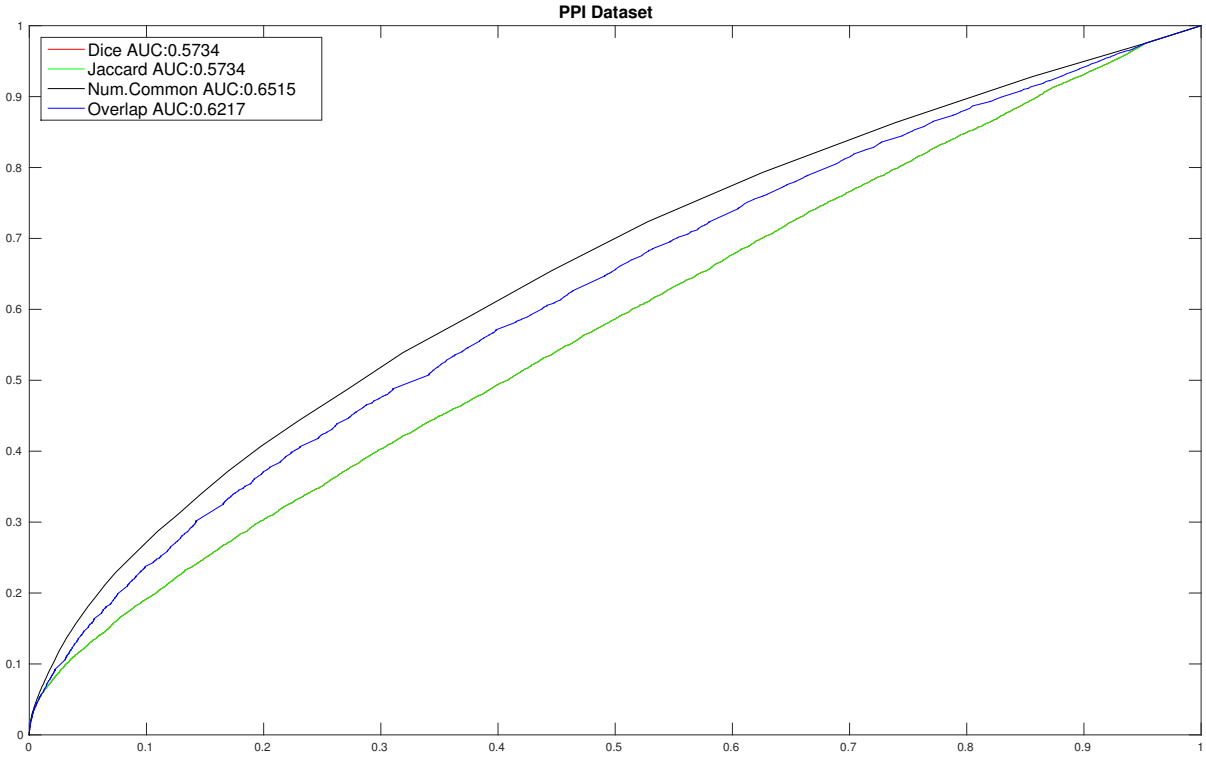

Figure 5.5: ROC curve of the simple similarity measures in the PPI dataset.

are similar to Resnik's [71], further analysis is required.

An important characteristic of Lin's and Jiang and Conrath's measure is that, any two diseases having at least one term in common would have a similarity of 1, that is, maximal irrespective of the specificity of this common term and the number and specificity of the non-overlapping annotations. To illustrate this scenario, consider two diseases, D_a and D_b, annotated as follows: $D_a = \{t_1, t_2, t_4, t_6\}$ and $D_b = \{t_1, t_7\}$. The similarity of these diseases is 1 given by the similarity of t_1 with itself, computed using either Jiang and Conrath or Lin. A similarity measure that assigns the maximum possible similarity whenever an annotation is shared will result in a large proportion of high-similarity pairs whenever high overlap in the annotations exists. Since these measures have a performance comparable to that of Resnik's in the GO [42] the question is: do the annotations in MeSH overlap substantially more than they do in

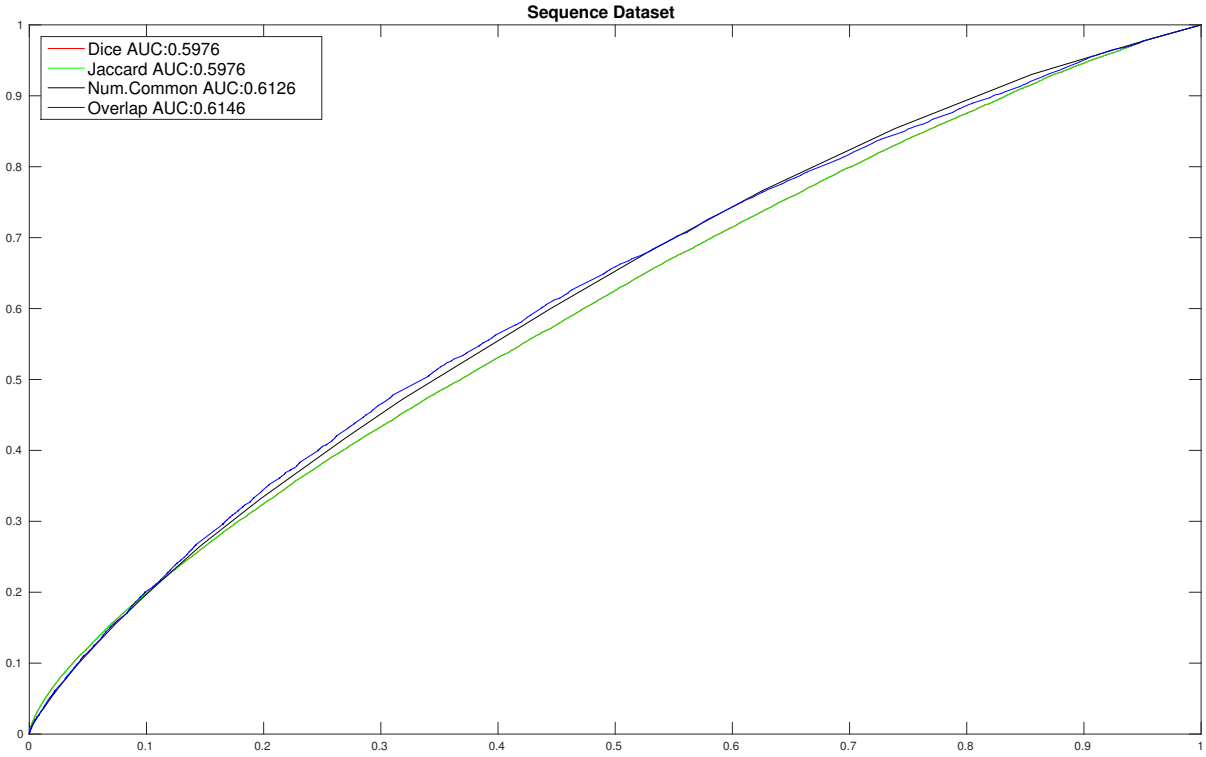

Figure 5.6: ROC curve of the simple similarity measures in the Sequence Similarity dataset.

GO?

To answer these question, I obtained GO annotations with experimental evidence codes (*i.e.* EXP, IDA, IPI, IMP, IGI and IEP [96]) for the model organisms *A. thaliana*, *H. sapiens*,*M. musculus*, *S. cerevisiae* and *C. elegans* from UniProt GOA [23]. I calculated the overlap of the annotations by counting the number of times the genes were annotated with the same GO term. I performed the same calculation for the MeSH annotations, and compiled the results in the box plot, shown in figure 5.8. The difference in the means between MeSH and every model organisms is significant. The p-values are: *A. thaliana*: $3.89 * 10^{-16}$, *H. sapiens*: $3.96 * 10^{-12}$, *M. musculus*: $1.37 * 10^{-11}$, *S. cerevisiae*: $2.84 * 10^{-16}$ and *C. elegans*: $3.87 * 10^{-18}$.

In general, the similarity scores according to Lin and Jiang and Conrath will be

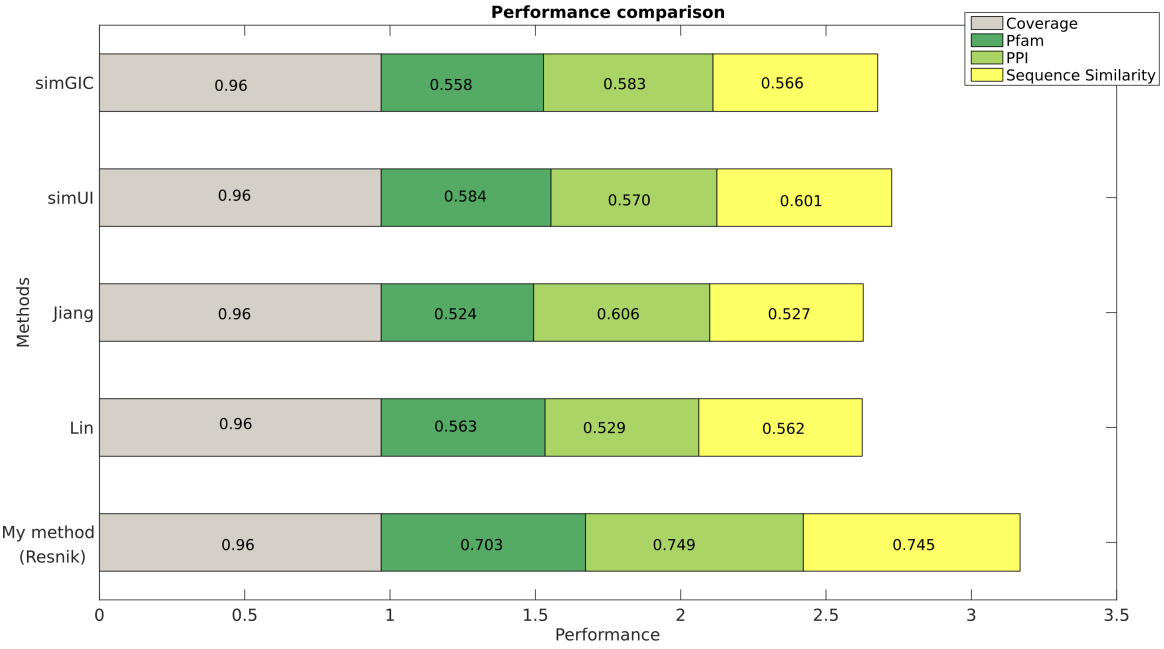

Figure 5.7: Performance of the evaluated semantic similarity measures on the combined MeSH ontologies

maximal whenever the similarity of the pair with the most informative LCA (*i.e.* with the highest information content) is chosen . This situation would not arise if other combinations were to be used (see chapter 1 §1.6). However, the maximum has proven to be successful in GO [42], and has the advantage of providing the actual most informative LCA. Knowing the most informative term that describes both diseases allows manual analysis of the pairwise similarity scores [16].

Considering that Resnik's similarity measure does not suffer from the shortcomings of Lin and Jiang and Conrath, and outperforms all other semantic similarity measures, I chose Resnik's to quantify the similarity between the sets of MeSH terms that describe the disease.

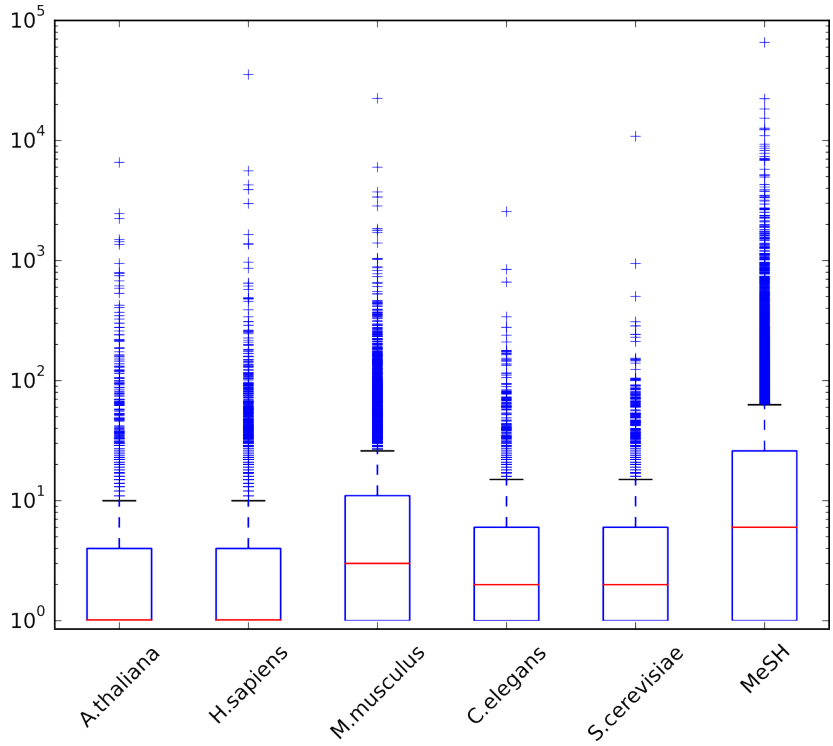

Figure 5.8: Comparison of the overlap of MeSH annotations in OMIM and the model organisms *A.thaliana*, *H. sapiens*, *M. musculus*, *C. elegans* and *S. cerevisiae*. The X-axis shows the different model organisms annotated with the Gene Ontology and OMIM annotated with MeSH. The Y-axis shows the distribution of overlapping annotations for each test case, in log scale. Notice the greater variability for the OMIM case. The difference between the means of MeSH and the model organisms is significant, as indicated by the p-values: *A. thaliana*: $3.89 * 10^{-16}$, *H. sapiens*: $3.96 * 10^{-12}$, *M. musculus*: $1.37 * 10^{-11}$, *S. cerevisiae*: $2.84 * 10^{-16}$ and *C. elegans*: $3.87 * 10^{-18}$.

5.3 The choice of MeSH subset

MeSH categorises the terms associated to a publication into Major Topic and non-Major Topic. A "MajorTopics" term designates a term extracted from the title or statement of purpose of the publication, and refers to its central focus point. The remaining MeSH terms are either qualifiers for the Major Topics or refer to topics substantially discussed in the publication.

The coverage of the proposed method was lower when considering only the MajorTopics set of annotations, with 7,094 (90.8% of OMIM) of the diseases having associated Major Topics against 7,575 (96.8% of OMIM) when considering all MeSH terms available. Performance between both sets was similar, shown in figures 5.9, 5.10 and 5.11. To ensure the widest possible coverage, I chose the entire set of MeSH terms.

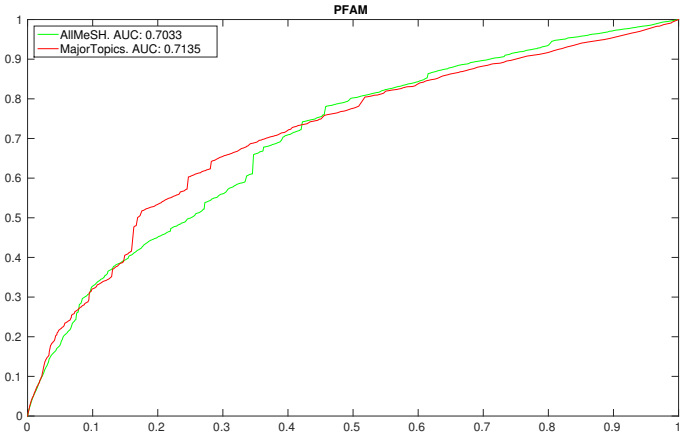

Figure 5.9: Performance comparison of the proposed method using all MeSH terms available and the major topics subset on the Pfam dataset

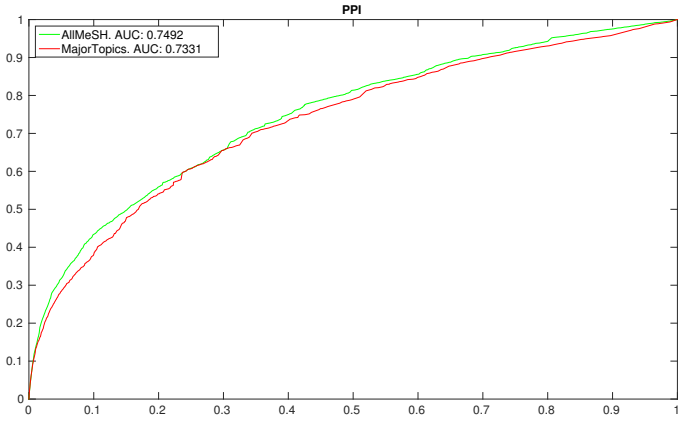

Figure 5.10: Performance comparison of the proposed method using all MeSH terms available and the major topics subset on the PPI dataset

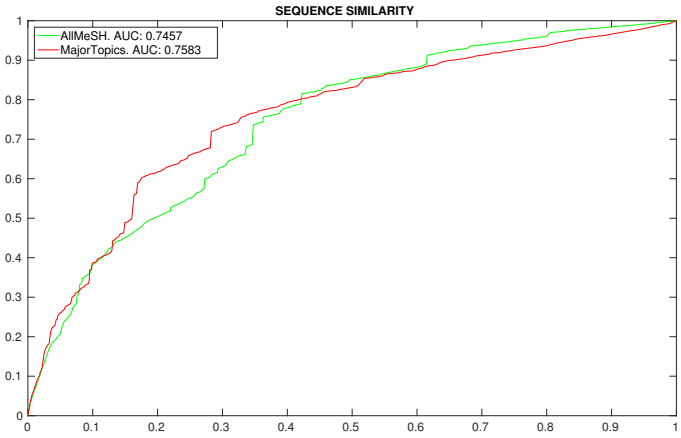

Figure 5.11: Performance comparison of the proposed method using all MeSH terms available and the major topics subset on the Sequence Similarity dataset

5.4 Decomposing the method: annotation and calculation

Conceptually, the proposed method can be thought of as a two-step process: an annotation step that results in MeSH terms being assigned to OMIM diseases (i.e. an OMIM-to-MeSH mapping) and a similarity calculation step. To gauge the impact of each step in the overall similarity of two diseases, the entire process of calculating disease similarity was decoupled for both the proposed method and van Driel's *et al.* [64] method. This was done as follows:

1. Replacing the proposed OMIM-to-MeSH mapping with van Driel's *et al.* OMIM-to-MeSH mapping (resulting from the text-mining analysis of the Clinical Synopsis (CS) and Text (TX) fields of OMIM). Note that, since the proposed method does not require weights for this initial mapping, the weights in van Driel's *et al.* OMIM-to-MeSH mapping, were removed.

2. Conversely, to verify the similarity calculation step, van Driel's *et al.* pipeline

was altered, replacing the implemented OMIM-to-MeSH mapping with the mapping provided by the proposed method. As van Driel's *et al.* method requires weights, a weight of 1 was assigned to each MeSH term. Since the annotation procedure defined for the proposed method does not consider repeated MeSH terms as relevant, a weight of 1 is appropriate according to van Driel's *et al.* annotation method.

The results are shown in figures 5.12, 5.13 and 5.14. As can be seen in these figures, using the MeSH terms associated to the publications to annotate the OMIM diseases provides rich descriptions that when used appropriately, can accurately quantify disease similarity. These comparisons highlight the fact that OMIM provides richer descriptions of the diseases that go beyond the symptoms and signs of a disease.

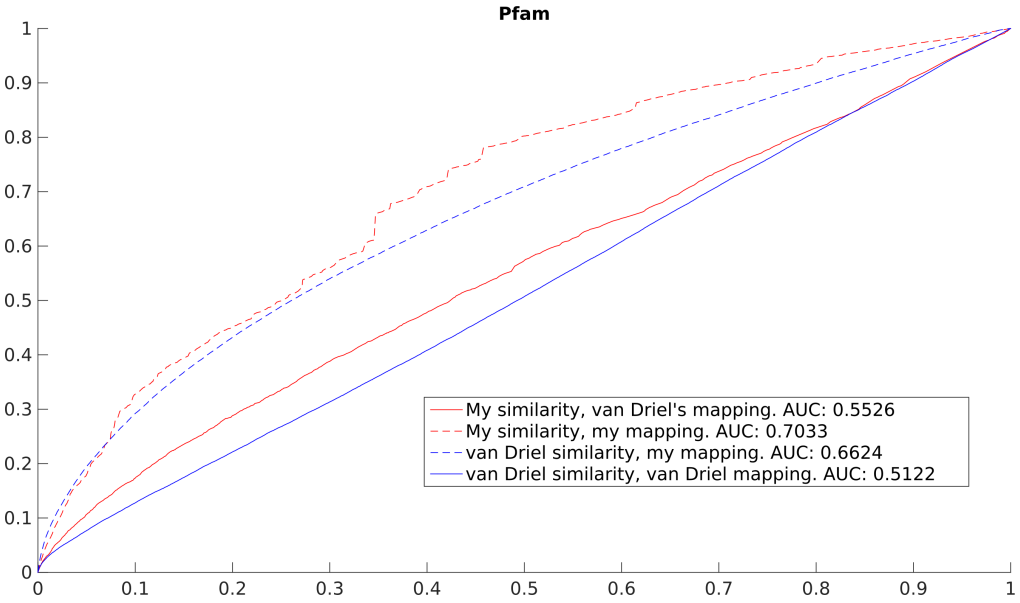

Figure 5.12: Performance comparison of the OMIM to MeSH mapping (Step 1) and the similarity calculation (Step 2) between the proposed method and van Driel's *et al.* method on the Pfam dataset.

Further analysis of the annotations produced by both the proposed method and van Driel's *et al.* method provides insight into the reasons for the poor performance of

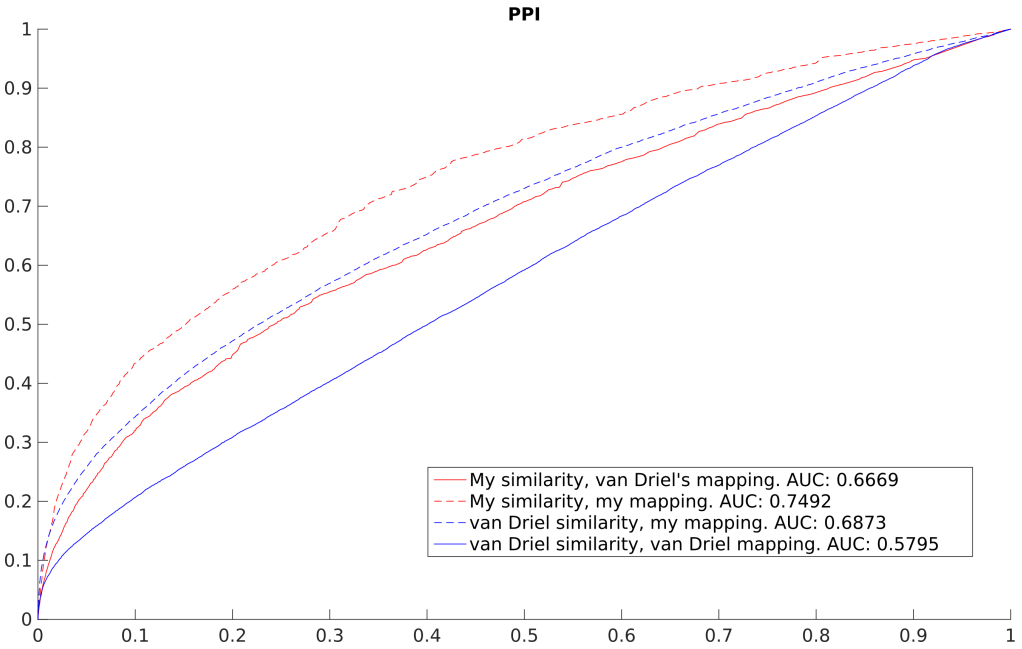

Figure 5.13: Performance comparison of the OMIM to MeSH mapping (Step 1) and the similarity calculation (Step 2) between the proposed method and van Driel's *et al.* method on the PPI dataset.

the text-mining method when analysing such complex entities as the OMIM entries. Consider the case of *Hyperpigmentation of Fuldauer and Kuijper* (`MIM:145200`), a disease with suspected Mendelian trait. This disease has no known genes associated to it and its short description provides a concise example.

The building blocks of the record analysed by van Driel *et al.* , namely the CS and TX fields, are reproduced from OMIM [4] in excerpt 5.4.

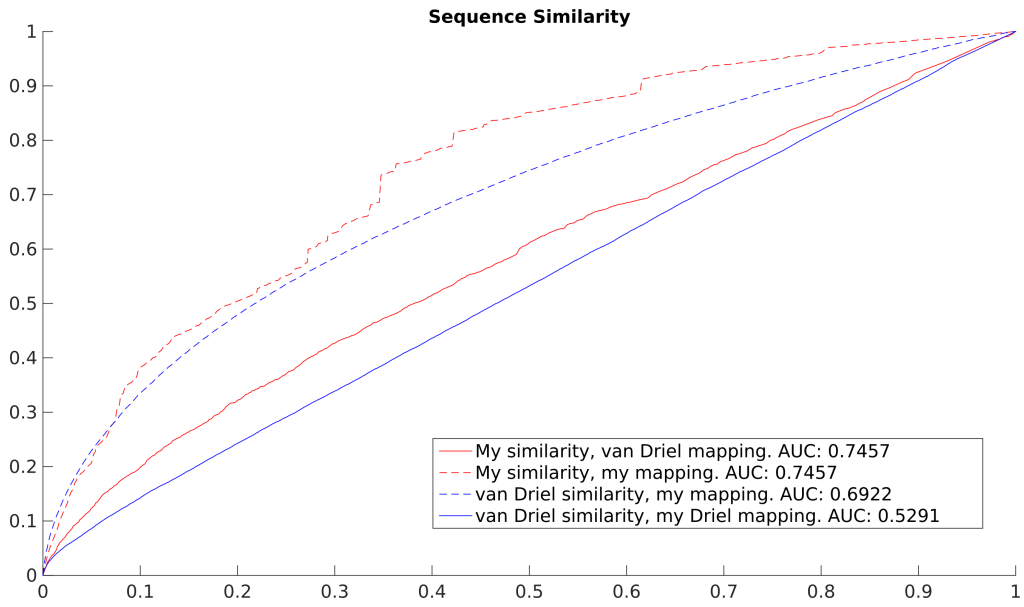

Figure 5.14: Performance comparison of the OMIM to MeSH mapping (Step 1) and the similarity calculation (Step 2) between the proposed method and van Driel's *et al.* method on the Sequence Similarity dataset.

*[**Text** TX]*

*Fuldauer and Kuijpers (1964) described a pigmentary anomaly in many members of a Dutch family. Although the paper was entitled **"Incontinentia Pigmenti,"** the distribution of the **hyperpigmentation** was quite different, being located on the **wrists**, **hands**, and **neck** and less consistently on the axillary folds, dorsa of the **feet**, and lines of the **hands**. Furthermore, incontinentia pigmenti is probably an X-linked dominant lethal in males. Many males were affected in this family.*

*[**Clinical Synopsis** CS]*

- *Inheritance: Autosomal dominant.*
- *Skin: **Hyperpigmentation** of **wrists**, **hands**, **neck** and less consistently on axillary folds, dorsa of the **feet** and lines of **hands***

Excerpt 5.4: OMIM record for Fuldauer and Kuijpers syndrome.

The boldface words represent the features selected by van Driel's *et al.* method. It is clear to see that while the term *Hand* (D006225) is chosen as descriptive of the disease, in reality, it provides little information on the disease itself, as in this case it was used to indicate the absence of a relationship.

Comparing the terms proposed by van Driel *et al.* with the terms obtained by my method, the relevance of the terms chosen becomes evident:

van Driel	My method
Foot (D005528), Hand (D006225), Incontinentia Pigmenti (D007184), Neck (D009333), Skin (D012867), Wrist (D014953), Hyperpigmentation (D017495)	Diagnosis, Differential (D003937), Ectodermal Dysplasia (D004476), Pigmentation Disorders (D010859)

While some annotations such as *Diagnosis, Differential* (D003937) might not be very informative, they are consistent with the description of the disease. My method is incapable of producing contradictory annotations such as the case of the term Foot (D005528) which van Driel's *et al.* method uses to annotate the disease. The MeSH annotations, by design, consist exclusively of terms which are relevant to the publication, therefore, cases such as the one mentioned, cannot happen.

5.5 Low variability of scores

After analysing the scores of highly similar diseases, I noticed that there is little variability in the scores. That is, in a relatively large set of disease pairs, very few different scores are present. As an example, a list of the similarities between *Breast Cancer* (MIM:114480) and the 10 diseases most similar to it is shown below:

1. Similarity 3.4:
    - Cervical Cancer (MIM:603956), LCA: Core binding factor beta (D050658)

- Mammographic density (MIM:607308), LCA: Mammography (D008327)
- Episodic Kinesigneic Dyskenisia (MIM:128200) LCA: Cerumen (D002571)

2. Similarity 3.58:

- Breast-ovarian cancer (MIM:604370), LCA: Mastectomy, Simple (D015413)
- Phosphoglycerate Dehydrogenase Deficiency (MIM:603956), LCA: Phosphoglycerate Dehydrogenase (D050543)
- Retinoblastoma (MIM:180200), LCA Neoplasm Seeding (D009366)
- Severe combined immunodeficiency (MIM:102700), LCA: Deamination (D003641)
- Estrogen receptor (MIM:133430), LCA: Nuclear Receptor Coactivator 3 (D056921)
- Epidermolysis bullosa (MIM:226730), LCA: Integrin alpha6 (D039503)
- Hypertrichosis (MIM:135400), LCA: Adenofibroma (D000232)

The little variability is due to the fact that the score depends on the number of diseases annotated by the lowest common ancestor, and it can happen that this number is the same for different pairs of diseases, even if the common ancestor is different. The LCA's for the diseases pairs with similarity score 3.58 annotate no other diseases than the ones in the example. The LCA's corresponding to disease pairs with score 3.40 are used to annotate three diseases each.

5.6 A brief analysis of the Goh *et al.* disease classes

When analysing Figure 4.26 in Chapter 4.3 it is important to consider that every disease is coloured according to a single class which based on the primary physiological system affected by the disease. My measure is not based only on the physiological

system affected by the diseases, but rather on their wider aetiology and also includes risk factors, related drugs and known associations to other diseases —hour measure is aimed at reflecting closeness on the interactome. This results in some diseases being placed among diseases of different classes, according to their "location" on the interactome. This effect is particularly noticeable for complex multifactorial diseases such as the *Cardiovascular* diseases.

In Figure 5.15 and have highlighted diseases classified as *Cardiovascular* that are embedded among other disease classes. *Ischemic Stroke* (`MIM 601367`) is located in a group of diseases classified as *Metabolic*. There are associations reported between Stroke and metabolic disorders such as *AOMS1* (`MIM 605552`) and my disease similarity measure scores the pair in the 99Th percentile. In the same group of *Metabolic* diseases is *Coronary artery disease* (`MIM 608320`), whose risk factors include obesity (disease similarity 95Th percentile), hypertension (disease similarity 99Th percentile), hypercholesterolemia (disease similarity 99Th percentile) and diabetes (disease similarity 96Th percentile) conditions related to the metabolic system. *Myxoma, Intracardiac* (`MIM 255960`) is located in the Cancer group and has high similarity to *Cancer* related disorders such as Carney Complex (99Th percentile) and Thyroid Carcinoma (99Th percentile). While myxomas are in general benign tumours, they share important hallmarks of cancer.

On the contrary, the tight group of Cardiovascular diseases at the centre of the plot (dashed line) contains diseases which are well described by a single class. In fact, these are intrinsically Cardiovascular and are related to mechanical failures of the hearth such as in the cases of *Ventricular Tachycardia* (`MIM 192605`), *Sick Sinus Syndrome* (`MIM 608567`) and *Hypoplastic Left Hearth Syndrome* (`MIM 241550`) to name a few. These diseases are highly similar with one another and dissimilar to most other diseases in OMIM.

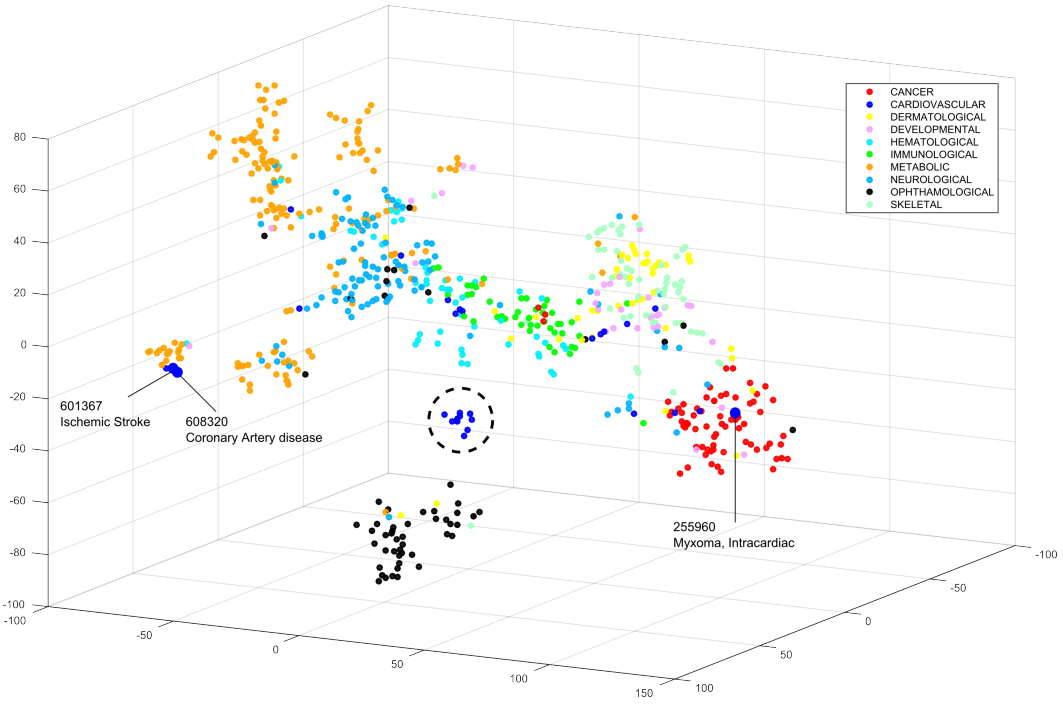

Figure 5.15: Embedding of OMIM diseases in 3D space. Each point in the plot represents an OMIM disease. The diseases are coloured according to the disease classes in Goh *et al.* . The highlighted diseases correspond to *Cardiovascular* diseases in the boundary with other classes. The dashed circle shows the tight group of cardiovascular diseases.

5.7 The effect of the number of genes on the similarity scores

While the method I developed does not rely on gene-disease associations the network medicine principles on which this work is based, imply that some relationships exist.

How does the number of genes in a disease affect its similarity to other diseases?

Diseases with many genes have, on average, slightly higher similarity scores. This is expected, as diseases with many genes will be more likely to be close to the other

diseases in the interactome —informally, one can think of their disease modules as being slightly "larger", and therefore closer.

To show this, I compared the mean similarity of two sets of diseases and all other diseases in OMIM. The first set consists of multigenic (strictly more than one gene) and the second set exclusively of monogenic (exactly one gene) diseases. The monogenic set consists of $3,743$ diseases and the multigenic set of 287 diseases.

The mean similarity between all diseases in the interactome and the diseases in the monogenic set is 1.19, compared to the 1.27 between all diseases and the multigenic diseases (p-value: 1^{-350}). This small difference is reflected on the smaller distance on the interactome: multigenic diseases are slightly closer to all other diseases (mean shortest path length 4.08) compared to the monogenic diseases (4.12); p-value: 1^{-350}.

Do disease-pairs share more common genes often have higher similarity scores than those pairs sharing one gene?

In Figure 4.25 I contrast two normalised histograms of disease similarity scores. The yellow histogram in the figure shows the distribution of scores of all pairs of diseases in OMIM, while the green histogram shows the distribution of scores for those diseases in OMIM which share at least one disease gene. The difference between both distributions shown in the figure is statistically significant, and interestingly, 90% of the pairs of diseases which share at least one gene, have similarity scores in the 99Th percentile or higher.

To verify the similarity scores of the diseases that do share genes, I calculated their similarities and represented them in a box plot shown in Figure 5.16. In this figure, the X-axis correspond to the number of shared genes, and the Y-axis to the distribution of similarity scores, shown as a box-and-whiskers. Each box-and-whiskers diagram represents the median similarity value (indicated by the red line) the upper and lower quartiles (indicated by the box segment below and above the median,

respectively) and the maximum and minimum values.

As can readily be observed, the similarity scores grow the more genes a pair shares. To verify the significance of the difference in the similarity distributions represented in Figure 5.16 I performed a pair-wise t-test between the diseases sharing no genes (labelled 0 in the X-axis) an all other diseases, and between the diseases sharing 1 gene and all other pairs. In Table 5.2 I show the p-values of these pairwise t-tests.

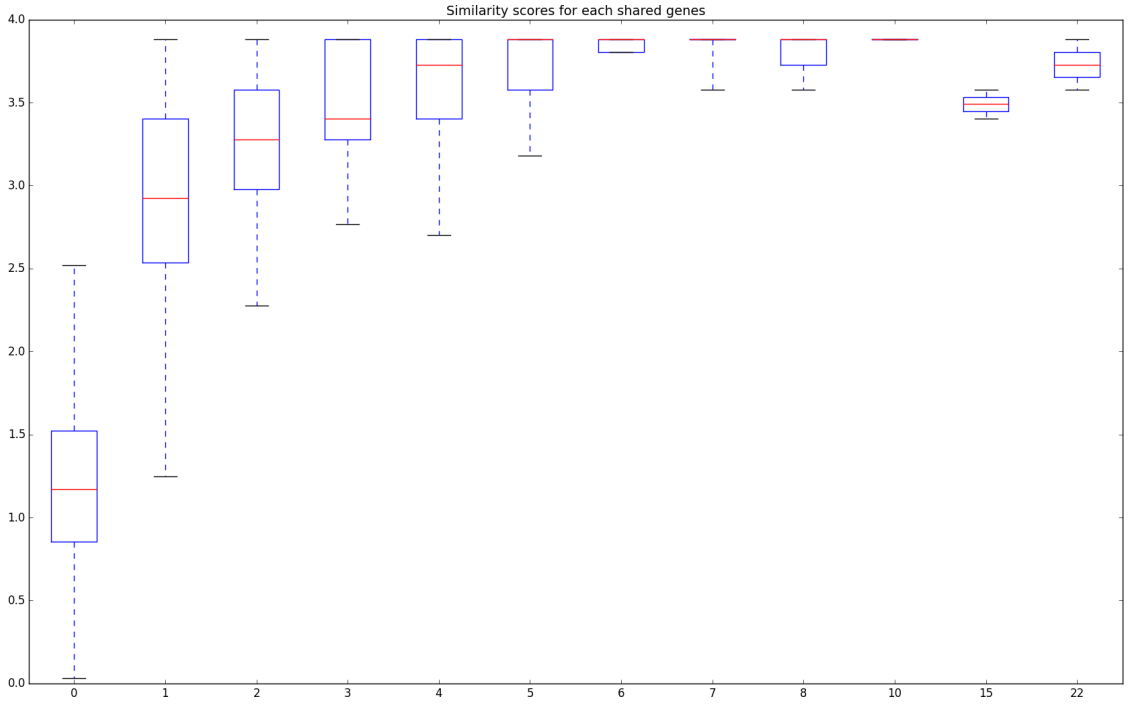

Figure 5.16: Distribution of similarity scores with respect to the number of shared genes. The plot shows the distribution of similarity scores for pairs of diseases with respect to the number of genes shared by them. The X-axis shows the number of genes shared by the pairs and each corresponding box represents the distribution of scores for those pairs of diseases. The red line in each box represents the median similarity value; the upper portion of each box represents the upper quartile of the distribution and the lower portion the lower quartile.

I have analysed this matter further by counting the number of shared genes in all pairs of diseases. The results are shown in Table 5.1, where we can see that the vast majority of diseases in OMIM do not share any genes.

Num. shared genes	Num. disease pairs
0	80,830,090
1	6,792
2	22
3	58
4	22
5	17
6	12
7	10
8	3
10	2
11	1
12	1
13	1
15	2
16	1
17	1
18	1
20	1
22	2
35	1

Table 5.1: Number of pairs of diseases with respect to the number of shared genes

	0	**1**
0	.	.
1	0.0	.
2	0.0	3.2^{-22}
3	6.7^{-261}	3.2^{-13}
4	4.8^{-111}	4.4^{-8}
5	5.2^{-94}	2.4^{-8}
6	3.7^{-72}	1.3^{-7}
7	4^{-1}	9.7^{-7}
8	5.8^{-19}	0.01
10	4.4^{-14}	0.01
15	1.1^{-10}	0.16
22	1.0^{-12}	0.04

Table 5.2: Pairwise t-test between diseases sharing no genes and all others and diseases sharing a single gene. While the p-value drops sharply above the 8 mark, for the 8, 10, 15 and 22 mark only 2 disease pairs exist.

While the significance at the 8 mark drops steeply, we must highlight that for the 8, 10, 15 and 22 mark only 2 disease pairs exist.

Chapter 6

Software

The very nature of the work I did during my PhD required the development of substantial amount of software. While exploratory data analysis constitutes the vast majority of the code I have written, two pieces of software stand out. I developed this software with a less technically oriented user in mind. I therefore provided easy to use interfaces and functionalities that allow the exploration of large amounts of data with simplicity. In this chapter I will present Gene Ontology Semantic Similarity Tool (GOssTo) and the Disease Similarity Explorer.

It is important to note that, although substantial, the pipeline developed for the calculation of the disease similarities is also fully available. However, since the construction of this pipeline is not aimed at wide usage, I have not included it in this chapter and is instead detailed in Appendix A.

6.1 The Gene Ontology Semantic Similarity Tool

Gene Ontology Semantic Similarity Tool (GOssTo) [39] is a user-friendly software system for calculating semantic similarities between gene products according to the Gene Ontology. GOssTo is bundled with six semantic similarity measures, including

both term- and graph-based measures, and has extension capabilities to allow the user to add new similarities. Few software tools have been proposed for calculating semantic similarities. ProteinOn [2] IT-GOM [37] and G-SESAME [107] stand out. These tools are provided as either stand-alone applications which are not readily extendible with new semantic similarity measures, or are available only as packages running within environments such as R or MATLAB. Other tools are exclusively available online and their use is impractical for high-throughput analysis on large bodies of data. Most tools do not allow for a straightforward calculation of semantic similarities for a whole genome, or an easy updating of the GO annotations.

GOssTo includes the Random Walk Contribution by Yang *et al.* [42] (see 1.7) , supports both term- and graph-based similarity measures and is available in downloadable binary form, with the entire source code released under GPLv3. GOssTo is easy to use and very fast Table 6.1 shows the time required for calculating the Resnik semantic similarity including the Random Walk Contribution for a few model organisms. GOssTo features a simple and concise command line interface and an application programming interface (API) for easy integration into high throughput data-processing pipelines.

GOssTo's design allows for user provided similarity measures to be independently developed, compiled and linked at runtime. A well-defined interface grants the user access to the data structures upon which new measures can be developed. After a new measure is independently compiled, it can be dynamically linked to GOssTo's application core, seamlessly integrating it to the main application and providing the same functionalities as the bundled measures. Thus, GOssTo can be used in three different ways: as a part of a larger data-processing pipeline; as a stand-alone application; as a static library for existing software. For easy processing of the results, all output is presented in structured plain text files.

GOssTo is also available online, through a clean web interface `www.paccanarolab.`

Organism	Number of GO terms	Number of annotated genes	Time term-wise	Time gene-wise
Arabidopsis	6,610	9,703	3m48s	43m35s
Rat	9,422	5,270	58m19s	29m54s
Mouse	12,961	15,020	24m35s	689m26s
Fly	7,304	8,235	4m56s	47m46s
Yeast	7,077	4,898	4m0s	23m55s
Worm	4,467	4,370	1m29s	5m1s

Table 6.1: For each organism: number of unique GO terms appearing in the GO annotation; number of annotated genes; time (in minutes and seconds) required for calculating the Resnik semantic similarity including the Random Walk Contribution term- and gene-wise. Calculations used GO experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS) and *is_a* and *part_of* GO relationships. Data downloaded in February 2014. Experiments run on a machine equipped with an AMD Opteron 6128 HE.

`org/gosstoweb`. GOssToWeb provides access to the same functionalities of the standalone application, allowing extensive configuration of the experiments through a user-friendly web form. The user can select GO evidence codes, GO relationships and a genome from the list of organisms available in UniProt-GOA. GOssToWeb automatically fetches the most recent version of the functional annotation from UniProt-GOA and of the GO from its official repository, thus ensuring that the most up-to-date data are used. Results are provided by redirecting the user to a page from which they can be downloaded. The system can notify the user with an email containing a link to the result download page.

The current version of GOssTo focuses on traditional semantic similarity measures which rely mostly on the GO structure. Future versions will include the possibility of handling Description Logic axioms which are being added to existing ontologies [44].

6.1.1 Technical details

GOssTo was developed, using the Java programming language, with the JAMA package providing the internal data types and the required mathematical routines. JAMA was modified slightly to rely on single-precision floating point numbers instead of double-precision floating point numbers. The decision to modify JAMA instead of the better known Apache Commons library has several reasons. Firstly, the Apache Commons library provides far more functionality than the required by GOssTo. Secondly, JAMA, being a more compact package, allowed simpler modification without compromising quality. The implemented changes resulted in a 50% reduction of GOssTo's memory footprint without compromising the quality of the results. The changes are extensively documented and all of GOssTo results, both final and intermediate, were thoroughly validated.

I developed GOssToWeb to widen the user base and allow less technically oriented users to have access to high-quality bioinformatics tools. GOssToWeb functions as a multiprocess queuing system to allow concurrent use of the resources in the shared server, and acts as an interface to the binary version of GOssTo. The system is illustrated in figure 6.1. After the user (green shaded area) submits the job to the server (blue shaded area) the User Interface (UI) will thoroughly validate all input parameters. This validation process will ensure that enough data was provided in order to calculate the results and that the provided parameters are valid. The validated job will be submitted to the Queuing System, who will check the load of the server and the status queue, and if there are enough resources available will spawn a worker process to handle the incoming job. The worker process will then run GOssTo (red shaded area) with the provided parameters and wait until it is complete. In this way, the queuing system is decoupled from the running of GOssTo, allowing more users to submit jobs that will be queued until resources become available . Once the job is complete, the worker process updates the queue status, and the UI shows

the result page or sends an email, according to the user preferences.

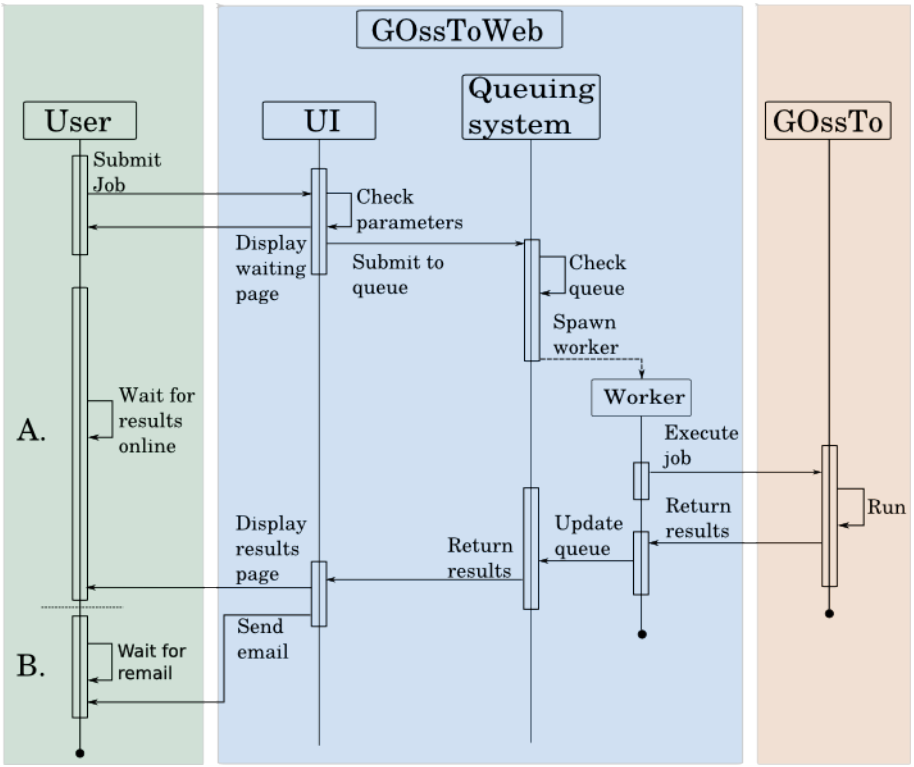

Figure 6.1: Simplified sequence diagram of GOssToWeb. The green shaded area corresponds to the user side. The blue shaded area to GOssToWeb, and is composed of the User Interface (UI) and the queuing mechanism. The standalone implementation of GOssTo is show in the red shaded area. Each time a new job gets submitted, the queuing system spawns a worker process to handle the request. Users can wait for the UI to display the result page (A) or provide an email address for GOssToWeb to notify them (B).

The User Interface components were developed using Php for server side programming, with JavaScript components on the client side. GOssToWeb keeps track of the jobs that were run, allowing a caching mechanism to be implemented that will check for identical parameter sets thus saving time and resources. The Queuing System was entirely written in Python, as were the Worker Processes.

The source code is freely available from GitHub at `https://github.com/pwac092/`

`gossto` released under the GPLv3 license. GOssTo runs on multiple platforms, and was extensively tested in on both GNU/Linux and Windows.

6.2 DisimWeb: A tool to explore disease similarities

I have developed DisimWeb in order to allow domain experts, medical doctors and the larger community to explore relatedness between the heritable diseases in OMIM. The browser, available at `http://www.paccanarolab.org/disimweb`, enables the users to obtain the similarity measure between over 28.5 million pairs of diseases, with dynamic links to OMIM, MeSH and UniProtKB databases.

The main page provides text fields, with preloaded options, where users can input a pair of diseases and obtain their similarity scores. Considering that there is a large number of diseases, disease names are auto completed to help the users quickly find the diseases. The pairwise similarity scores are presented in a single result page, a screenshot of which is shown in figure 6.2. The result page provides the similarity score as well as the MeSH terms that annotate each of the diseases being compared. Each MeSH term is linked to its own record page in the National Library of Medicine website.

Since the similarity scores are unbounded positive numbers, I include in the result page the percentile in which the score is located. This percentile is shown both as a number as well as graphically in a histogram that indicates the position of the similarity score of the disease pair in relation to the entire dataset. A red dot is located above the bar in the histogram corresponding to the similarity of the pair.

In addition to obtaining pairwise scores between the diseases, I developed a neighbourhood "explorer", a screenshot of which is shown in figure 6.3. The neighbourhood

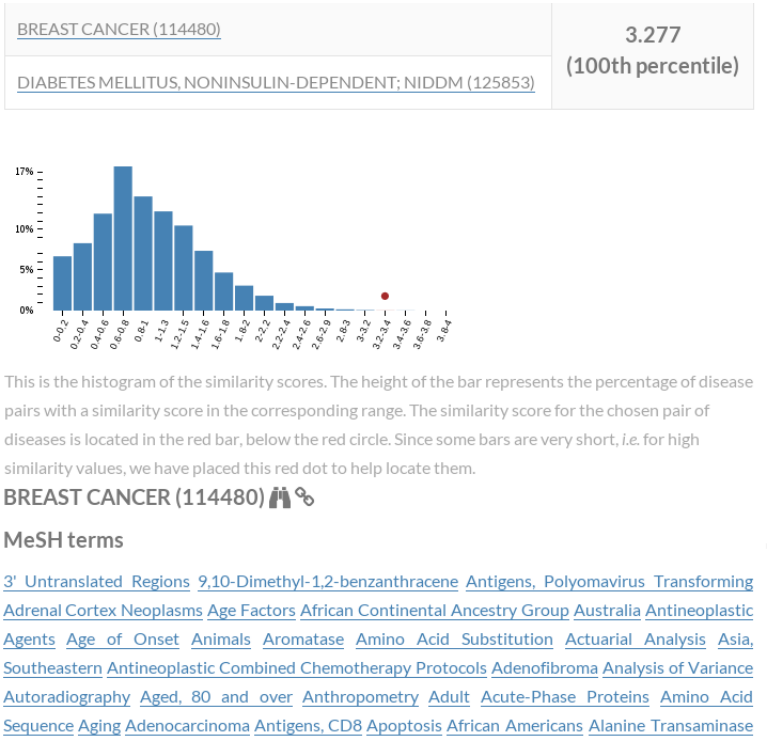

Figure 6.2: Screenshot of the "Search" feature. The pairwise similarity score and the percentile of the score is shown at the top right. To contextualise the score with all other scores, a histogram of similarity scores is shown. Clicking on the binoculars "Explores" the disease's neighbourhood. The link symbol redirects the user to the OMIM. The MeSH terms for each disease are listed as hyperlinks that point go the corresponding entry page in the National Library of Medicine website.

explorer allows user to graphically navigate through the disease associations by querying the similarity database and retrieving the diseases most similar to the ones he/she finds relevant. In the plot, every node is an OMIM disease and the links represent the similarity score between the two disease it connects, coloured according to its similarity value. The target disease appears in the centre, and it is connected to the 10 diseases that are most similar to it. In order to provide a wider picture, these 10 diseases are connected to their 5 most similar ones. The number of nodes displayed in the graph are not fixed, as the user can choose to include more neighbours at each level.

I have also added a "Fill Network" feature, that allows to fully connect the nodes in the graph. Considering that users can add a substantial number of nodes in each level (*i.e.* direct neighbours of the explored disease, as well as neighbours of the neighbours), a warning is displayed when filling using the "Fill Network" feature might excessively load the users computer.

The user can also choose whether to hide or display node labels as well as choose which labels to display. As a default MIM numbers are shown, but a single click displays the disease names. While the names are long and might, at first glance, confuse the graph, the nodes can be moved and rearranged by clicking and dragging them around the canvas. Additionally, all elements in the graph are clickable. Clicking on a disease node shows the disease name, while clicking on a link shows similarity between the connected diseases.

The default layout is a concentric layout, where the disease being explored is located in the centre of the plot, however, I have made several layouts available allowing users to obtain different perspectives on the same data. Users can select among a circular, breadth first, random, grid or a force-directed layout. The entire plot, with labels and colours, can be downloaded in a high-resolution PNG image.

6.2.1 Technical details

I developed DisimWeb using Django [30] a high-level web framework based on Python. The browser uses an SQLite in the backend to store the disease similarities, with indices specified to speed up the fetching of the data.

The front-end of the application relies on several JavaScript libraries both to produce a fluid user experience. jQuery [52] provides general functionality and improved user experience such as the autocompletion of disease names. The D3.js [29] library provides the capabilities to process the similarity scores and build the histogram shown in the results page. The neighbourhood explorer is built with the Cytoscape.js

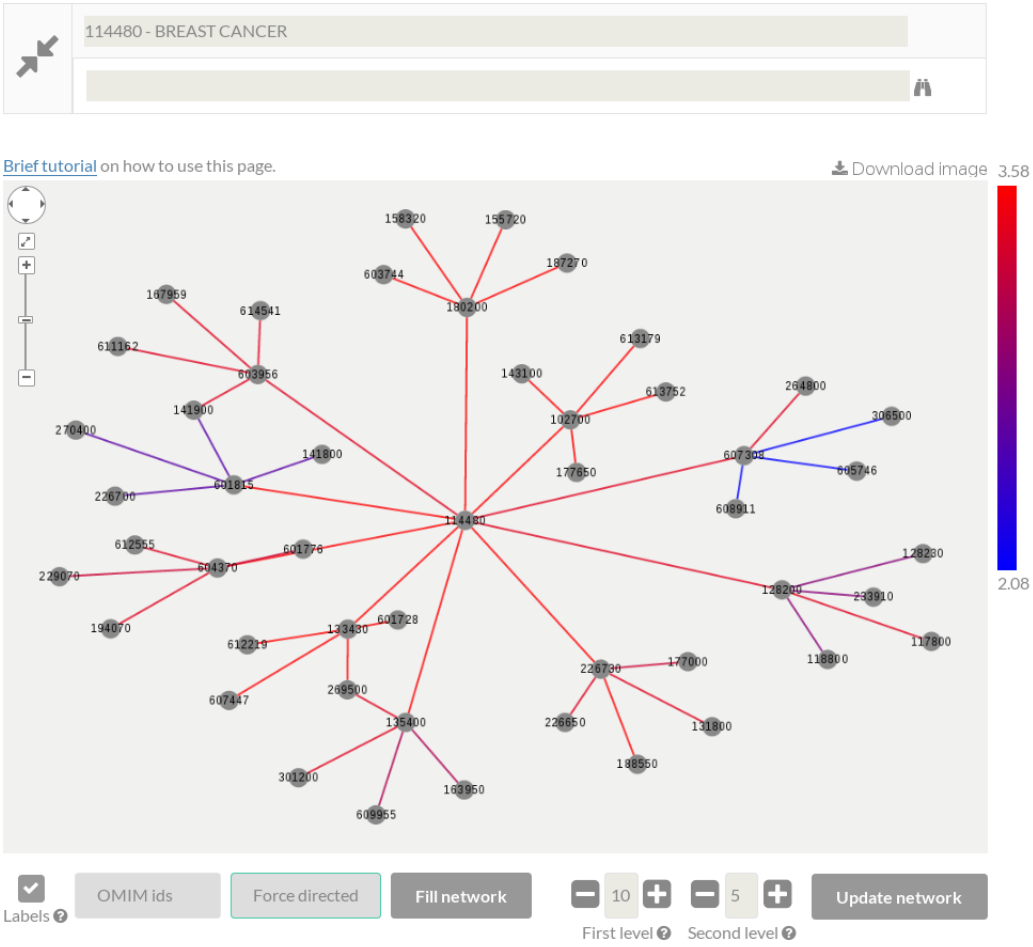

Figure 6.3: Force directed layout of a disease's neighbourhood.

[21] library.

The source code of the browser is freely available from GitHub at `https://github.com/pwac092/disimweb` and is released under a GPLv3 license.

Chapter 7

Future Work

In this chapter I present a brief analysis of future work related to my Thesis.

7.1 Disease gene prediction

The first, and perhaps more straightforward evolution of my work, is the prediction of disease genes. To aid in this task, I have produced the Disease Similarity Resource (DSR), a database of the disease pairs whose similarity scores was in the top 5% ($1,552,356$ pairs) and their associated disease genes. Each pair constitutes an entry, and contains 5 columns:

1. Disease A
2. Disease B
3. Similarity Score
4. UniProt identifiers of the proteins associated to disease A
5. UniProt identifiers of the proteins associated to disease B

Since similar diseases lie "close" on the interactome, similar these highly similar disease pairs are, suitable candidates for transferring knowledge between them. Thus the DSR provides a starting point for an in-depth analysis into the relationships and aetiology of the diseases, providing the basis for a statistical gene-discovery process.

In a recent collaboration with Valentini *et al.* [38] we show that weighted integration of networks improves the performance of kernel-based gene prioritisation techniques. The DSR provides an orthogonal source of information that correlates with closeness on the interactome, and could therefore be integrated into an existing gene prioritisation pipeline to improve performance.

7.2 Analysis of complex diseases

In this Thesis, I focused on diseases in Online Mendelian Inheritance in Man (OMIM), where very few disease-gene associations elucidated through "statistical" methods are included (see Appendix C). While OMIM includes complex diseases such as various Cancers, more complex inheritance patterns and larger sets of genes could be elucidated for the diseases through methods such as Genome-wide Association Studies (GWAS).

I explored the capability of my method to quantify disease similarity for diseases with larger set of genes, where a more complex genotype-phenotype relationship might exist. While there is not definitive way to classify diseases in OMIM as *Complex* or *Simple* I attempted three different ways of obtaining this classification. I report all of them below, even if only one gave results which I consider meaningful and was able to use afterwards.

1. Extracting GWAS traits from OMIM.

I classified the OMIM diseases whose disease-gene associations were obtained through GWAS as *Complex* and the remaining as *Simple*. To do this, I developed a method which classifies the OMIM diseases appearing in the EBI GWAS catalogue [31] as *Complex* while the remaining were classified as *Simple*. Considering that the traits in the GWAS catalogue and the OMIM diseases do not have identical names I used an approximate string matching algorithm that produces a similarity score for two given strings. This score (the Levenshtein distance) is based on the number of deletions, insertions and substitutions that are required to match the query strings. I calculated this score for every possible pair GWAS trait - OMIM disease. The dataset contained $21,529$ GWAS traits and $7,812$ OMIM diseases, so I calculated a total of $168,184,548$ scores. An OMIM disease was considered to be *Complex* if it was highly similar (similarity $\geq 90\%$) to a GWAS trait. Unfortunately, this process returned only 60 OMIM diseases being classified as *Complex* which is less than 1% of the total, as well as many diseases classified as *Simple* even if many disease genes have already been associated with them.

2. Filtering OMIM based on the Phenotype mapping key

I contacted the staff at OMIM, who recommended us to filter the OMIM database based on the Phenotype Mapping Key of the disease (see C). Following their recommendation I built a set of *Complex* diseases with those diseases that had the Phenotype Mapping Key 2, and a set of *Simple* diseases with the mapping key 3. Unfortunately, this process had similar issues as our previous method, as it returned only 63 OMIM diseases being classified as *Complex* which is less than 1% of the total, as well as many diseases (261) classified as *Simple* even if many disease genes have already been associated with them.

3. Extracting the multigenic disorders from OMIM

I classified all multigenic diseases (with more than one gene) in OMIM as *Complex* and all monogenic diseases as *Simple*. Here I assume that the multiple disease genes complicate the elucidation of the gene-disease relationship and therefore, multigenic diseases correspond to the set of inherently *Complex* diseases. In this way I obtained a set of 287 *Complex* diseases and a set of $3,743$ *Simple* diseases. There is a statistically significant difference (t-test p-value $\leq 10^{-350}$) between the mean number of disease genes associated to the *Complex* diseases (3.61) and to the *Simple* diseases (1). The results of the evaluation on the three datasets, Pfam, PPI and Sequence Similarity, are shown in Figure 7.1 .

The composite performance of my method is slightly inferior for the set of *Complex* diseases with respect to the *Simple* diseases —the overall composite score is 3.08 for *Complex* and 3.13 for *Simple*. Interestingly, the method by Park [86], which uses molecular level information, is the only method that shows the same behaviour; the methods of Köhler [81] and van Driel [64] obtain a better performance on each of the 3 datasets for *Complex* rather than *Simple* diseases. Overall, my method is the most stable as it varies the least in performance between the 2 sets of diseases.

Finally, It is important to note that in Figure 7.1 the coverage of the different methods is determined only for the diseases in the *Complex* and *Simple* sets and not for all of OMIM. Nevertheless, the Area Under the Curve (AUC) performance of the methods is comparable to those shown in Figure 4.21 in Chapter 4.3 where the methods are evaluated on all diseases in OMIM.

Interestingly, the EBI GWAS catalogue references the publications for the experiments. They might provide a starting point for a similarity measure of the disease traits based on my method. Conceptually, the problem is similar —the method would provide annotations for the disease traits in the catalogue.

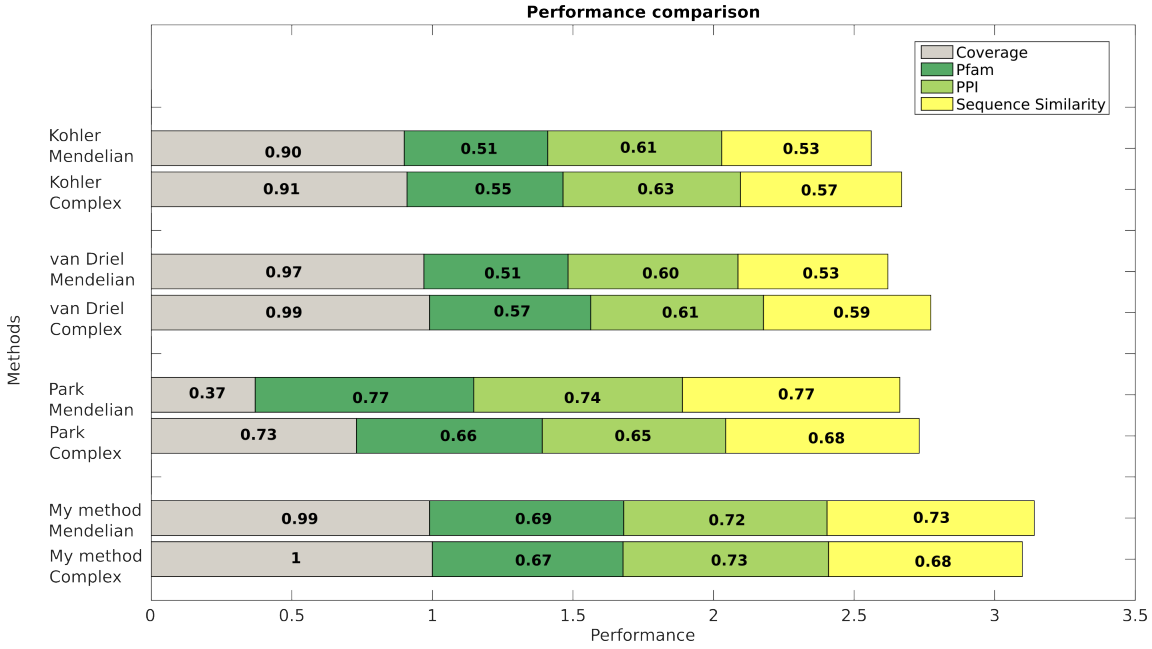

Figure 7.1: This figure compares the performance of my method and those by van Driel, Park and Robinson on the *Complex* and *Simple* sets of diseases. Coverage is defined as the fraction of diseases in the *Simple* and *Complex* sets for which a similarity can be calculated.

7.3 Computerised Medical Support Systems

Clinical Decision Support Systems (CDSS) are knowledge-based systems designed to aid the medical practitioner and patients in everyday decisions and proper care [14]. Several systems fall in the CDSS category such as automatic reminders, drug dosage verifiers, patient specific recommendations, automated differential diagnosis and others [14].

Medical diagnosis is a process that begins with a particular medical grievance, and concludes only with a specific, categorised identification of the cause of the grievance [7]. The diagnosis process is vague, and heavily dependent on the physicians abilities and preparation, as shown in studies such as the ones presented by Melo *et al.* [62] who explore Brain MRI's of Medical Doctors diagnosing a disease. To aid in this

process, several formulas in the form of Medical Algorithms have been developed [83]. A simple example is the calculation performed to obtain the Body Mass Index (BMI). This formula compiles various factors, height, sex and weight to obtain a single number to characterise a patients weight.

Many algorithms are available [83], and they attempt to summarise expertise and knowledge into a single, easily distributed procedure, that can help doctors. While they provide a clear path to follow once the disease or condition has been identified, the procedures for the elucidation of the disease is a constructive approach based on trial and error of conditions the medical practitioner identifies as relevant. The process of differential diagnosis, focuses on the correct discrimination of diseases from those that exhibit similar characteristics, but whose treatments are different.

This first step of providing alternative hypothesis is particularly complex in poorer, less developed areas, where few medical professionals are available and consultation with other physicians is not always a possibility. An automated system for the ranking of putative alternative diagnosis might provide a much needed sounding board in these poorer isolated regions. Systems to provide these automated Differential Diagnosis Generator (DDx) already exist, Bond *et al.* [102] evaluate 19 of such systems, ranging from general systems to condition specific systems.

As a first step, my measure could provide an up-to-date reference for physicians that could help discard diseases that are similar to the one experienced by the patient, but that require different patterns of care. It could allow physicians in remote areas to obtain more information to help contextualise the condition the patient presents.

I can also envision an interactive differential diagnosis system that would aid medical practitioners in identifying putative alternative diagnoses that are obscured by the complexity and multiplicity of the symptoms.

Appendices

Appendix A

Running the disease similarity pipeline

This Appendix presents a guideline to replicate the results of the disease similarity measure. Each step in the calculation process was implemented in an independent script, in order to confine any possible bugs. While this results in several programs various languages to be executed, the independence of the steps allows any and all steps to be replaced, as long as the file formats are respected.

All data is available from `www.paccanarolab.org/disease_similarity`. The code is released under GPLv3 and is available from `https://github.com/pwac092/disim_calculator`. The disease similarity browser is available at `www.paccanarolab.org/disimweb`, for details see chapter 6.

The pipeline was developed using Python 2.7.3 on a Debian 7 GNU/Linux system running the 3.16 Linux kernel. The guideline presented in this Appendix was developed and tested on a Debian 7 GNU/Linux environment, and should be applicable to most GNU/Linux systems.

A.1 Extracting the Online Mendelian Inheritance in Man (OMIM) data

The OMIM data has to be manually downloaded from `www.omim.org`. Registration is required, but the data is freely available. Since OMIM contains both genes and diseases, to extract the diseases from the catalogue the following prefixes have to be matched: $+, \#, \%$ and `null`, that is, a line without prefix.

While it is possible to extract the phenotypes from the morbidmap or genemap files, parsing the omim.txt file can be done exclusively using Linux GNU coreutils tools such as grep and awk. This simplifies the process and reduces the probability of making programmatic errors.

The following awk command can extract all OMIM identifiers from the omim.txt file, along with the name and prefix.

```
awk '/*FIELD* TI/{getline;print}' omim.txt > records
```

To extract only the phenotype prefixes:

- For the `null` prefix:

```
cut -f1 -d "␣" records|grep "^[0-9]"
```

- For the + prefix:

```
cut -f1 -d "␣" records|grep "^+[0-9]"|cut -c2-
```

- For the # prefix:

```
cut -f1 -d " " records|grep "^#[0-9]"|cut -c2-
```

- For the % prefix:

```
cut -f1 -d " " records|grep "^\%[0-9]"|cut -c2-
```

For convenience, the extracted OMIM should be sorted. This can be done using the `sort` command as follows:

```
sort diseases -o diseases
```

A.2 Extracting the referenced publications

To extract the publications referenced by each OMIM disease, the `OMIM_query.py` python script hast to be used. This script will query OMIM public API, extracting all the publications for a particular disease. An API key is required for this step, but this is freely available from OMIM at `http://omim.org/api`.

Alternatively, the omim.txt file could be manually parsed and the publications extracted in this fashion. Querying the API is safer and reduces the chances of programmatic errors, considering the free-text omim.txt file will require specific parsers and very detailed debugging to ensure their quality. The XML results provided by OMIM are easily parseable using python packages.

The required input for the script is:

1. A single-column file containing the list of OMIM diseases for which the publications will be obtained.

2. The name of the output file. For illustrative purposes, this file will be named omim2pubmed.

3. A configuration file which contains the API key provided by OMIM.

The output file is formatted in a way that makes it easy to parse in the Linux console. Each line consists of the OMIM disease followed by the PubMed identifiers obtained from OMIM separated by tabs. The configuration file is read for the API details using the python ConfigParser module. The format is as follows:

```
[APIconfig]
Server = api.europe.omim.org
Key = 5ED0AEDA215A37C589A9AF0E3EAF1F143033E50
```

The API key shown is for illustration purposes only and as such, is not valid. The server has to be chosen according the one's location. It is important to note that there are OMIM records for which no references can be fetched.

Once the PubMed identifiers are obtained form OMIM, they need to be extracted from the omim2pubmed file:

```
cut -f2- omim2pubmed | tr 't' 'n' | sort -n | uniq
```

A.3 Fetching the Medical Subject Headings (MeSH) terms

To extract the MeSH terms associated to each PubMed identifier, the `PubMed_query.py` python script hast to be used. This script will query PubMed's public API, extracting all the MeSH terms for a particular publication. This script will fetch the MeSH

terms for a given list of PubMed identifiers through API queries to Entrez E-utils. The input required for the script are:

1. The list of PubMed identifiers for which the MeSH terms will be fetched.
2. A string (Yes/No) indicating whether only to get the major topics MeSH terms or all MeSH terms.
3. Double column file, mapping MeSH term names (e.g. Adult) to their unique descriptor identifier (e.g. D000328).
4. The name of the output file. For illustrative purposes this file will be named pubmed2mesh.
5. The configuration file for the Entrez e-utils.

As with the mim2mesh file, the output file is formatted in a way that makes it easy to parse in the Linux console. Each line consists of the PubMed identifier followed by the MeSH identifiers obtained from PubMed separated by tabs.

A.4 Annotating OMIM with MeSH

To annotate the OMIM diseases with MeSH terms the mim2mesh.py Python script has to be used. This script will map each OMIM disease to the MeSH terms of the PubMed identifiers the disease references. It will produce the mapping file also, it will provide a file with all the OMIM diseases it could not map. The input required for the script are:

1. The mapping between OMIM records and PubMed identifiers.
2. The mapping between PubMed identifiers and MeSH terms.

3. The desired output file.

As with the mim2mesh and pubmed2mesh files, the output file is formatted in a way that makes it easy to parse in the Linux console. Each line consists of the OMIM disease followed by the MeSH identifiers associated to the disease's referenced publications.

A.5 Computing the pairwise disease similarities

The pipeline allows the calculation of similarity scores using a subset of the ontologies or combining the ontologies according to the method I propose in chapter 5. The `compute_matrices.py` script computes the scores for each individual ontology, while the `compute_combined_similarities.py` script computes the combined similarities.

In both cases, the similarity scores are presented in a triplet format, where the first two columns correspond to the diseases and the last column the similarity score between both diseases. The file has the format:

```
OMIM-1 OMIM-2 sim_score
OMIM-1 OMIM-3 sim_score
...
```

A.6 Producing the benchmarks

The benchmarks (Pfam dataset (Pfam), Protein-Protein interaction dataset (PPI) and Sequence Similarity dataset (SS)) Files are represented as triplets, where the first two columns contain the diseases, and last column contains 1 or 0, depending on the physical evidence supporting the similarity of the diseases. The file has the format:

```
OMIM–1 OMIM–2 1/0
OMIM–1 OMIM–3 1/0
...
```

The construction of the evaluation datasets requires three sources of data:

1. The Disease-protein mapping.
2. The Protein-protein interaction network.
3. The sequences of the diseases proteins.
4. The Pfam signature information for the disease proteins.

A.6.1 Getting the data required for the benchmarks

There are several Protein-Protein interaction datasets available. I have chosen the Human Protein Reference Database (HPRD) [93] available from `www.hprd.org`.

The disease protein to OMIM disease can be obtained from several sources. Primarily, it can be parsed from the `morbidmap` file provided by OMIM. However, UniProt provides a mapping file, named `mimtosp.txt`, which maps OMIM diseases to Unpaired identifiers and Gene Names. This file is simpler to parse, and the use of UniProt identifiers removes the need for further translation of gene names and protein identifiers.

For convenience, the script `convert_mimtosp.pl` transforms the mimtosp.txt file into a file that is simpler to process in the GNU/Linux console. This script has no input, it fetches the `mimtosp.txt` file from UniProt and converts it to the `mimtoprot.txt` file.

The resulting `mimtprot.txt` file has a simple two column format, where the first column contains the OMIM disease and the second column the UniProt ID of the

disease protein. Diseases with multiple proteins will appear several times in the file. For example:

```
101900  P16615
102200  O00170
102200  P63092
...
```

To obtain the sequences of the proteins associated to the diseases (required for the Sequence Similarity dataset), the script `get_sequences.pl` has to be used. This script automatically fetches the sequences for the proteins producing a sing file with the sequences in FASTA format. It requires a single input:

1. The `mimtoprot.txt` file.

Lastly, to obtain the Pfam-A signatures of the proteins, the `pfam_scan.pl` has to be used. This script is available from Pfam, and will produce a tabular file associating the Family, Domain, Motif and Repeats associated to the each protein.

A.6.2 The Pfam dataset

The `pfamBenchmark.py` script will build the Pfam benchmark following the criteria defined in chapter 4.3. The script takes three parameters:

1. The output of the `pfam_scan.pl` script.
2. The `mimtosp.txt` file.
3. A single-column file of valid OMIM diseases to consider.
4. A single-column file of Pfam identifiers to exclude. This parameter is optional, and can be left blank. See 4.3 for details on the exclusion of Pfam signatures.

The output is produced with the default name `omim_pfam` in the local directory where the script was run.

A.6.3 The PPI dataset

The `MIM2gene.py` script produces will build the PPI benchmark following the criteria defined in chapter 4.3.

The required inputs are:

1. UniProt ID to Gene Name mapping file.
2. `mimtosp.txt` file. This file is provided by UniProt and maps OMIM records to their known proteins using UniProt identifiers `http://www.uniprot.org/docs/mimtosp.txt`
3. Protein protein interaction dataset file.
4. A single column file of accepted OMIM numbers.

Different PPI networks can be chosen by defining a vector of columns in the class Interactions in the `MIM2gene.py` script. After this vector is defined, suffices with appropriately replacing the call to the constructor

```
hprd = Interactions(sys.argv[3], 'columnsHPRD')
```

by replacing the second argument with the appropriate parameter.

A.6.4 Sequence similarity dataset

Two Perl scripts are required to construct the Sequence Similarity dataset. In the following, they are detailed in the order they should be executed.

The `makeblast.pl` script calculates the Smith-Waterman alignment of the proteins provided. This script requires only the FASTA sequences of the proteins to compare.

The `produce_sequence_similarity.pl` produces the Sequence Similarity dataset. Two diseases are positively related when the sequence similarity e-value is lower than $10^{-}6$. The inputs are:

1. `mimtoprot.txt`, the file mapping OMIM diseases to UniProt proteins.

2. The alignment of the proteins.

A.7 Final comments

This guide is intended to allow results to be replicated with ease. Each script mentioned in this Appendix is commented explaining its behaviour.

The time required for fetching the data through the API interfaces varies depending on the load of the server and the network resources available. The bottleneck step is the actual calculation of the similarity scores. Dr. Alfonso E. Romero and I have implemented several mechanisms to speed up the process, nevertheless, the calculations could exceed 10 hours.

Further improvements in the implementation will be performed based on need.

Appendix B

Publications referenced by the old OMIM data

The following tables present the publications referenced by the OMIM diseases explored in chapter 4.3 section 4.7.

The following list corresponds to the publications referenced by the April 2013 version of OMIM.

- Tetralogy of Fallot (`MIM:187500`): 10587520, 5065286, 20807224, 4003436, 9132487, 11152664, 4834778, 11714651, 19597493, 8923932, 9188669, 4050848, 15937089, 20631719, 20581743, 1425789, 2260602, 21110066, 13943847, 14517948, 19948535, 21919901, 18055909, 18672102

- Right Atrial Isomerism (`MIM:208530`): 6638068, 7715640, 9152295, 874654, 3674113, 4003441, 3425603, 8834045, 6712272, 8873667, 7172476, 6622295, 1021593, 6050934, 6638069, 14929628, 2012140, 1191445, 14128648, 4774542, 9155619, 9443444, 7277426

- Noninsulin-dependent Diabetes (NDDIM) (`MIM:125853`): 17726085, 11443197, 10973253, 16885549, 2695375, 22286214, 9038347, 10720052, 12874106, 18323454,

17463246, 19657112, 15808156, 10199785, 10331426, 11575290, 10958757, 9745421, 15924147, 17273962, 8528247, 12915642, 8528248, 11032783, 21186350, 16775236, 17603485, 12783844, 9758619, 11130726, 18477659, 15472205, 20016592, 7971976, 9032096, 11158011, 19020324, 15940393, 11030756, 1357346, 17066296, 17906635, 19020323, 15070960, 16142453, 11916952, 1587533, 20085713, 20574426, 19933169, 12750520, 14960743, 10902787, 12045211, 9498630, 9541507, 18952314, 12851856, 17603484, 12118251, 15980866, 17463248, 9312173, 9482914, 11723072, 17179727, 17293876, 18008060, 8897863, 9062343, 11904371, 9892237, 18711366, 12727978, 18231124, 11067779, 11032784, 20360734, 17554300, 21118154, 22456733, 16034410, 18711367, 22456732, 11533494, 17463249, 22456734

- SHORT syndrome (`MIM:269880`): 6407320, 8574420, 8790109, 15481036, 12514365, 18384141, 21340693, 8279490, 8669449, 2729352, 4050863

- Dermatofibrosarcoma protuberans (`MIM:607907`): 9738795, 11291071, 11435686, 12209598, 17478383, 12202658, 12660034, 8988177, 12661001, 15221986

- Juvenile Myelomonocytic Leukemia (`MIM:607785`): 19420352,18182584, 19388938, 10086728, 19372255, 15723289, 21562564, 19571318, 11588050, 17332249, 20008299, 9160658, 20543203, 9616134, 16474405, 12717436

The following list corresponds the publications referenced in the August 2014 version of OMIM.

- Tetralogy of Fallot (`MIM:187500`): 10587520, 5065286, 20807224, 4003436, 9132487, 11152664, 4834778, 11714651, 19597493, 8923932, 9188669, 4050848, 15937089, 20631719, 20581743, 1425789, 2260602, 21110066, 13943847, 14517948, 19948535, 22939634, 18055909, 18672102

- Right Atrial Isomerism (`MIM:208530`): 6638068, 7715640, 9152295, 874654, 3674113, 9201627, 4003441, 3425603, 14648004, 8834045, 6712272, 8873667,

7172476, 6622295, 20413652, 6050934, 6638069, 14929628, 2012140, 1191445, 14128648, 4774542, 9155619, 9443444, 7277426

- Noninsulin-dependent Diabetes (NDDIM) (`MIM:125853`): 17726085, 11443197, 10973253, 16885549, 2695375, 22286214, 9038347, 10720052, 12874106, 18323454, 17463246, 19657112, 15808156, 10199785, 10331426, 11575290, 10958757, 9745421, 15924147, 17273962, 8528247, 12915642, 8528248, 11032783, 21186350, 16775236, 17603485, 12783844, 9758619, 11130726, 18477659, 15472205, 20016592, 7971976, 9032096, 11158011, 19020324, 15940393, 11030756, 1357346, 17066296, 17906635, 19020323, 15070960, 16142453, 11916952, 1587533, 20085713, 20574426, 19933169, 12750520, 14960743, 10902787, 12045211, 9498630, 9541507, 18952314, 12851856, 17603484, 12118251, 15980866, 17463248, 9312173, 9482914, 11723072, 17179727, 24390345, 17293876, 18008060, 8897863, 9062343, 11904371, 9892237, 18711366, 12727978, 18231124, 11067779, 11032784, 20360734, 17554300, 21118154, 22456733, 16034410, 18711367, 22456732, 11533494, 17463249, 22456734

- SHORT syndrome (`MIM:269880`): 6407320, 8574420, 11135494, 8790109, 23810379, 23810382, 15481036, 12514365, 18384141, 21340693, 8279490, 8669449, 2729352, 23810378, 4050863

- Dermatofibrosarcoma protuberans (`MIM:607907`): 9738795, 11291071, 11435686, 12209598, 17478383, 12202658, 12660034, 8988177, 12661001, 15221986

- Juvenile Myelomonocytic Leukemia (`MIM:607785`): 19420352, 18182584, 19388938, 10086728, 19372255, 15723289, 21562564, 19571318, 11588050, 17332249, 20008299, 9160658, 20543203, 9616134, 23832011, 16474405, 12717436

Finally, table B.1 shows the changes in the both the April 2013 and August 2014 releases of OMIM. All publications are shown as PubMed identifiers.

OMIM disease	Added	Removed
Tetralogy of Fallot (MIM:187500)	22939634	21919901
Right Atrial Isomerism (MIM:208530)	9201627, 14648004, 20413652	1021593
Noninsulin-dependent Diabetes (NDDIM) (MIM:125853)	24390345	-
SHORT syndrome (MIM:269880)	11135494, 23810378, 23810379, 23810382	-
Dermatofibrosarcoma protuberans (MIM:607907)	-	9920784, 10607907, 10607907
Juvenile Myelomonocytic Leukemia (MIM:607785)	23832011	-

Table B.1: Publications associated to the diseases in the August 15, 2014 version of OMIM.

Appendix C

Dividing the set of OMIM diseases

To divide the diseases in OMIM based on the mapping method, I contacted the OMIM staff. Their reply is included below:

> *Dear Horacio,*
>
> *If you are trying to exclude phenotypes that are placed on the map by GWAS, I suggest selecting by "Phenotype Mapping Key". Phenotypes that have a mapping code of 3 will have a known molecular basis. Phenotypes with a 2 will be placed on the map by linkage, GWAS, or other "statistical" methods. The mapping method codes available in the FTP download are not aggressively curated and generally the code "Fd" has been used for so-called statistical mappings of disease to the genome. In addition, we do not generally add GWAS information to OMIM unless the P value is astronomical. GWAS Catalog and GWAS Central are dedicated to GWAS data and are available from the "External Links" link at the top of every OMIM.org page.*
>
> *Sincerely,*
>
> *Joanna Amberger*

Bibliography

[1] History of the development of the ICD. http://www.who.int/classifications/icd/en/HistoryOfICD.pdf.

[2] Proteinon: A web tool for protein semantic similarity. Tech. rep.

[3] The Gene Ontology structure. http://geneontology.org/page/ontology-relations.

[4] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM®. http://www.omim.org.

[5] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). Online Mendelian Inheritance in Man, OMIM®. MIM Number: 101400. http://www.omim.org.

[6] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics 7*, Suppl 1 (2006), S7.

[7] A. Baerheim. The diagnostic process in general practice: has it a two-phase structure? *Family Practice 18*, 3 (jun 2001), 243–245.

[8] A. Brückner, C. Polge, N. Lentze, D. Auerbach and U. Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences 10*, 6 (jun 2009), 2763–2788.

[9] A. Chatr-aryamontri, A. Ceol, L. M. Palazzi, G. Nardelli, M. V. Schneider, L. Castagnoli and G. Cesareni. MINT: the molecular INTeraction database. *Nucleic Acids Research 35*, Database (jan 2007), D572–D574.

[10] A. G. Clark. Determinants of the success of whole-genome association testing. *Genome Research 15*, 11 (nov 2005), 1463–1467.

[11] A. H.C. Wong. Phenotypic differences in genetically identical organisms: the epigenetic perspective. *Human Molecular Genetics 14*, suppl_1 (apr 2005), R11–R18.

[12] A-L. Barabási, N. Gulbahce and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Review Genetics 12*, 1 (jan 2011), 56–68.

[13] A. Pinna, S. Heise, R. J. Flassig, A. Fuente and S. Klamt. Reconstruction of large-scale regulatory networks based on perturbation graphs and transitive reduction: improved methods and their evaluation. *BMC Systems Biology 7*, 1 (2013), 73.

[14] A. X. Garg, N. K. J. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam and R. B. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *Journal of the American Medical Association 293*, 10 (mar 2005), 1223.

[15] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J Goldberg, K. Eilbeck, A. Ireland, C. J Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S. Sansone, R. H Scheuermann, N. Shah, P. L Whetzel and S. Lewis. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology 25*, 11 (nov 2007), 1251–1255.

[16] C. Pesquita, D. Faria, H. Bastos, A. Falcão and F. Couto. Evaluating go-based semantic similarity measures. In *Proc. 10th Annual Bio-Ontologies Meeting* (2007), vol. 37, p. 38.

[17] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcão and F. M. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics 9*, Suppl 5 (2008), S4.

[18] C. Stark. BioGRID: a general repository for interaction datasets. *Nucleic Acids Research 34*, 90001 (jan 2006), D535–D539.

[19] C. Thauvin-Robinet, M. Auclair, L. Duplomb, M. Caron-Debarle, M. Avila, J. St-Onge, M. Le Merrer, B. Le Luyer, D. Héron, M. Mathieu-Dramard, P. Bitoun , J. Petit, S. Odent, J. Amiel , D. Picot, , V. Carmignac, , J. Thevenon, , P. Callier , M. Laville , Y. Reznik , C. Fagour , M. Nunes , J. Capeau , O. Lascols , F. Huet , L. Faivre , C. Vigouroux , J. Rivière . Pik3r1 mutations cause syndromic insulin resistance with lipoatrophy. *The American Journal of Human Genetics 93*, 1 (Jul 2013), 141–149.

[20] CDC. Sickle Cell Disease (SCD). Centers For Disease Control. `http://www.cdc.gov/ncbddd/sicklecell/data.html`. Online; accessed July 2015.

[21] Cytoscape JS. Cytoscape.js. `http://js.cytoscape.org/`. Online; accessed June 2015.

[22] D. B. Goldstein. Common genetic variation and human traits. *New England Journal of Medicine 360*, 17 (apr 2009), 1696–1698.

[23] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan and R. Apweiler. The GOA database in 2009–an integrated gene ontology annotation resource. *Nucleic Acids Research 37*, Database (jan 2009), D396–D403.

[24] D. Hanahan and R.A. Weinberg. Hallmarks of cancer: The next generation. *Cell 144*, 5 (mar 2011), 646–674.

[25] D. Lin. An information-theoretic definition of similarity. In *ICML* (1998), vol. 98, pp. 296–304.

[26] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. de Ferranti, J.-P. Despres, H. J. Fullerton, V. J. Howard, M. D. Huffman, S. E. Judd, B. M. Kissela, D. T. Lackland, J. H. Lichtman, L. D. Lisabeth, S. Liu, R. H. Mackey, D. B. Matchar, D. K. McGuire, E. R. Mohler, C. S. Moy, P. Muntner, M. E. Mussolino, K. Nasir, R. W. Neumar, G. Nichol, L. Palaniappan, D. K. Pandey, M. J. Reeves, C. J. Rodriguez, P. D. Sorlie, J. Stein, A. Towfighi, T. N. Turan, S. S. Virani, J. Z. Willey, D. Woo, R. W. Yeh and M. B. Turner. Heart disease and stroke statistics–2015 update: A report from the american heart association. *Circulation 131*, 4 (dec 2014), e29–e322.

[27] D. N. Cooper, M. Krawczak, C. Polychronakos, C. Tyler-Smith and H. Kehrer-Sawatzki. Where genotype is not predictive of phenotype:

towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Human Genetics 132*, 10 (jul 2013), 1077–1130.

[28] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen and C. von Mering. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research 43*, D1 (oct 2014), D447–D452.

[29] D3.js. D3.js. `http://d3js.org/`. Online; accessed June 2015.

[30] Django. Django web framework. `https://www.djangoproject.com/`. Online; accessed June 2015.

[31] EBI. GWAS Catalog: The NHGRI-EBI Catalog of published genome-wide association studies.

[32] EBI. UniProtKB/TrEMBL Protein database release 2015 08 statistics.

[33] F. Azuaje, H. Wang and O. Bodenreider. Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies* (2005), pp. 9–10.

[34] F. M. Couto and M. J. Silva. Disjunctive shared information between ontology concepts: application to gene ontology. *Journal of Biomedical Semantics 2*, 1 (2011), 5.

[35] F. Prinz, T. Schlange and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery 10*, 9 (aug 2011), 712–712.

[36] G. D. Bader. BIND: the biomolecular interaction network database. *Nucleic Acids Research 31*, 1 (jan 2003), 248–250.

[37] G. K. Mazandu and N. J. Mulder. Information content-based gene ontology semantic similarity approaches: Toward a unified framework theory. *BioMed Research International 2013* (2013), 1–11.

[38] G. Valentini, A. Paccanaro, H. Caniza and A. E. Romero and M. Re. An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods. *Artificial Intelligence in Medicine 61*, 2 (jun 2014), 63–78.

[39] H. Caniza, A. E. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, M. Mesiti, G. Valentini and A. Paccanaro. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology. *Bioinformatics 30*, 15 (mar 2014), 2235–2236.

[40] H. Janssen, J. Meinardi, F. Vleggaar, S. van Uum, E. Haagsma, F. van der Meer, J. van Hattum, R. Chamuleau, , R. Adang, Rob P, J. Vandenbroucke and R. FR. van Hoek. Factor v leiden mutation, prothrombin gene mutation, and deficiencies in coagulation inhibitors associated with budd-chiari syndrome and portal vein thrombosis: results of a case-control study. *Blood 96*, 7 (2000), 2364–2368.

[41] H. Jeong, S. P. Mason, A.-L. Barabási and Z. N. Oltvai. *Nature 411*, 6833 (May 2001), 41–42.

[42] H. Yang, T. Nepusz and A. Paccanaro. Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics 28*, 10 (apr 2012), 1383–1389.

[43] H. Yu, R. Jansen, G. Stolovitzky and M. Gerstein. Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics 23*, 16 (may 2007), 2163–2173.

[44] J. D. Ferreira, J. Hastings, F. M. Couto. Exploiting disjointness axioms to improve semantic similarity measures. *Bioinformatics 29*, 21 (sep 2013), 2781–2787.

[45] J. Das and H. Yu. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology 6*, 1 (2012), 92.

[46] J. Das, T. V. Vo, X. Wei, J. C. Mellor, V. Tong, A. G. Degatano, X. Wang, L. Wang, N. A. Cordero, N. Kruer-Zerhusen, A. Matsuyama, J. A. Pleiss, S. M. Lipkin, M. Yoshida, F. P. Roth, H. Yu. Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Science Signaling 6*, 276 (may 2013), ra38–ra38.

[47] J. E. Sadler. New concepts in von willebrand disease. *Annual Review of Medicine 56*, 1 (feb 2005), 173–191.

[48] J. Ott, J. Wang and S. M. Leal. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Review Genetics 16*, 5 (mar 2015), 275–284.

[49] J. Petschnigg, J. Snider and I. Stagljar. Interactive proteomics research technologies: recent applications and advances. *Current Opinion in Biotechnology 22*, 1 (feb 2011), 50–58.

[50] J. W. Whitaker, G. A. McConkey and D. R. Westhead. The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. *Genome Biology 10*, 4 (2009), R36.

[51] J.J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008* (1997).

[52] jQuery. jQuery.js. `https://jquery.com`. Online; accessed June 2015.

[53] K. B. Michels, C. G. Solomon, F. B. Hu, B. A. Rosner, S. E. Hankinson, G. A. Colditz and J. E. Manson. Type 2 diabetes and subsequent incidence of breast cancer in the nurses' health study. *Diabetes Care 26*, 6 (jun 2003), 1752–1758.

[54] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal and A.-L. Barabasi. The human disease network. *Proceedings of the National Academy of Sciences 104*, 21 (may 2007), 8685–8690.

[55] L. Cheng, J. Li, P. Ju, J. Peng and Y. Wang. SemFunSim: A new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS ONE 9*, 6 (jun 2014), e99415.

[56] L. Hartwell, J. Hopfield, John S. Leibler and A. Murray. From molecular to modular cell biology. *Nature 402* (1999), C47–C52.

[57] L. M. Schriml and E. Mitraka. The disease ontology: fostering interoperability between biological and clinical human disease-related data. *Mamm Genome* (jun 2015).

[58] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 2579-2605 (2008), 85.

[59] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J.M. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J.

E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics 25*, 1 (2000), 25–29.

[60] M. J. Holtzman, D. E. Byers, J. Alexander-Brett and X. Wang. The role of airway epithelial cells and innate immune cells in chronic respiratory disease. *Nature Reviews Immunology 14*, 10 (sep 2014), 686–698.

[61] M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research 28*, 1 (jan 2000), 27–30.

[62] M. Melo, D. J. Scarpin, E. Amaro, R. B. D. Passos, J. R. Sato, K. J. Friston and C. J. Price. How doctors generate diagnostic hypotheses: A study of radiological diagnosis with functional magnetic resonance imaging. *PLoS ONE 6*, 12 (dec 2011), e28752.

[63] M. Oti and H.G. Brunner. The modular nature of genetic diseases. *Clinical Genetics 71*, 1 (oct 2006), 1–11.

[64] M. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, J. A. M. Leunissen. A text-mining analysis of the human phenome. *European Journal of Human Genetics 14*, 5 (Feb 2006), 535–542.

[65] M. Vidal. A unifying view of 21st century systems biology. *FEBS Letters 583*, 24 (dec 2009), 3891–3894.

[66] M. Vidal, M.E. Cusick and A-L. Barabási. Interactome networks and human disease. *Cell 144*, 6 (mar 2011), 986–998.

[67] N. Guarino D. Oberle and S. Staab. What is an ontology? In *Handbook on Ontologies*. Springer Berlin Heidelberg, 2009, pp. 1–17.

[68] O. Hein, M. Schwind and W. König. Scale-free networks. *Wirtschaftsinformatik 48*, 4 (2006), 267–275.

[69] ONS. Office for National Statistics, United Kingdom.

[70] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H.-W. Mewes, A. Ruepp and D. Frishman. The MIPS mammalian protein-protein interaction database. *Bioinformatics 21*, 6 (nov 2004), 832–834.

[71] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007* (1995).

[72] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR abs/1105.5444* (2011).

[73] P. Wang, J. Qin, Y. Qin, Y. Zhu, L. Y. Wang, M. J. Li, M. Q. Zhang and J. Wang. ChIP-array 2: integrating multiple omics data to construct gene regulatory networks. *Nucleic Acids Research 43*, W1 (apr 2015), W264–W269.

[74] Princeton University. About WordNet.

[75] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. A. Fulcher, T. A. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research 42*, D1 (nov 2013), D459–D471.

[76] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate and M. Punta. Pfam: the protein families database. *Nucleic Acids Research 42*, D1 (nov 2013), D222–D230.

[77] R. Navigli and S. Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence 193* (dec 2012), 217–250.

[78] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin and C. ODonovan. The GOA database: Gene ontology annotation updates for 2015. *Nucleic Acids Research 43*, D1 (nov 2014), D1057–D1063.

[79] R. Van Noorden. Sluggish data sharing hampers reproducibility effort. *Nature* (jun 2015).

[80] S. E. Antonarakis and J. S. Beckmann. Mendelian disorders deserve more attention. *Nature Review Genetics 7*, 4 (mar 2006), 277–282.

[81] S. Köhler, S. C. Doelken, C. J. Mungall, S. Bauer, H. V. Firth, I. Bailleul-Forestier, G. C. M. Black, D. L. Brown, M. Brudno, J. Campbell, D. R. FitzPatrick, J. T. Eppig, A. P. Jackson, K. Freson, M. Girdea, I. Helbig, J. A. Hurst, J. Jahn, L. G. Jackson, A. M. Kelly, D. H. Ledbetter, S. Mansour, C. L. Martin, C. Moss, A. Mumford, W. H. Ouwehand, S.-M. Park, E. R. Riggs, R. H. Scott, S. Sisodiya, S. V. Vooren, R. J. Wapner, A. O. M. Wilkie, C. F. Wright, A. T. Vulto-van Silfhout, N. d. Leeuw, B. B. A. de Vries, N. L. Washingthon, C. L. Smith, M. Westerfield, P. Schofield, B. J. Ruef, G. V. Gkoutos, M. Haendel, D. Smedley, S. E. Lewis and

P. N. Robinson. The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research 42*, D1 (nov 2013), D966–D974.

[82] S. Mathur, D. Dinakarpandian. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics 45*, 2 (apr 2012), 363–371.

[83] S. Mushlin and H. Greene II. *Decision making in medicine: an algorithmic approach*. Elsevier Health Sciences, 2009.

[84] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and H. Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research 42*, D1 (nov 2013), D358–D363.

[85] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, N. H. Campbell, G. Chavali, C. Chen, N. del-Toro, M. Duesbury, M. Dumousseau, E. Galeota, U. Hinz, M. Iannuccelli, S. Jagannathan, R. Jimenez, J. Khadake, A. Lagreid, L. Licata, R. C. Lovering, B. Meldal, A. N. Melidoni, M. Milagros, D. Peluso, L. Perfetto, P. Porras, A. Raghunath, S. Ricard-Blum, B. Roechert, A. Stutz, M. Tognolli, K. van Roey, G. Cesareni and

H. Hermjakob. The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research 42*, D1 (nov 2013), D358–D363.

[86] S. Park, J.-S. Yang, Y.-E. Shin, J. Park, S. K. Jang and S. Kim. Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Molecular Systems Biology 7*, 1 (jan 2011), 494–494.

[87] S. Russell and G. Roth. Pseudo-von willebrand disease: a mutation in the platelet glycoprotein ib alpha gene associated with a hyperactive surface receptor. *Blood 81*, 7 (1993), 1787–1791.

[88] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie and A. J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology 6*, 2 (feb 2010), e1000662.

[89] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis and A. Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research 41*, D1 (nov 2012), D991–D995.

[90] T. Fawcett. Roc graphs: Notes and practical considerations for researchers. *Machine learning 31* (2004), 1–38.

[91] T. Nepusz, H. Yu and A. Paccanaro. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods 9*, 5 (mar 2012), 471–472.

[92] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition 5* (1993), 199–220.

[93] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey. Human protein reference database–2009 update. *Nucleic Acids Research 37*, Database (jan 2009), D767–D772.

[94] T. Schlitt and A. Brazma. Current approaches to gene regulatory network modelling. *BMC Bioinformatics 8*, Suppl 6 (2007), S9.

[95] Taylor, I. W., and Wrana, J. L. Protein interaction networks in medicine and disease. *Proteomics 12*, 10 (may 2012), 1706–1716.

[96] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research 43*, D1 (nov 2014), D1049–D1056.

[97] UMBEL. Upper Mapping and Binding Exchange Layer. `http://umbel.org`.

[98] United States National Library of Medicine. The Medical Subject Headings – MeSH. `http://www.nlm.nih.gov/mesh/mbinfo.html`.

[99] V. Marx. Biology: The big challenges of big data. *Nature 498*, 7453 (jun 2013), 255–260.

[100] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, H. Parkinson and

L. M. Schriml. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research 43*, D1 (oct 2014), D1071–D1078.

[101] W. Criekinge and R. Beyaert. Yeast two-hybrid: State of the art. *Biological Procedures Online 2*, 1 (oct 1999), 1–38.

[102] W. F. Bond, L. M. Schwartz, K. R. Weaver, D. Levick, M. Giuliano and M. L. Graber. Differential diagnosis generators: an evaluation of currently available computer programs. *Journal of General Internal Medicine 27*, 2 (jul 2011), 213–219.

[103] W. H. Dunham, M. Mullin and A. Gingras. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *PROTEOMICS 12*, 10 (may 2012), 1576–1590.

[104] WHO. World Health Organization. Causes of death in the developed world. `http://www.who.int/mediacentre/factsheets/fs310/en/`. Online; accessed July 2015.

[105] X. Wang and N. Gulbahce and H. Yu. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics 10*, 5 (jul 2011), 280–293.

[106] X. Zhou, J. Menche, A-L Barabási and A. Sharma. Human symptoms–disease network. *Nature Communications 5* (jun 2014).

[107] Z. Du, L. Li, C.-F. Chen, P. S. Yu and J. Z. Wang. G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Research 37*, Web Server (jun 2009), W345–W349.