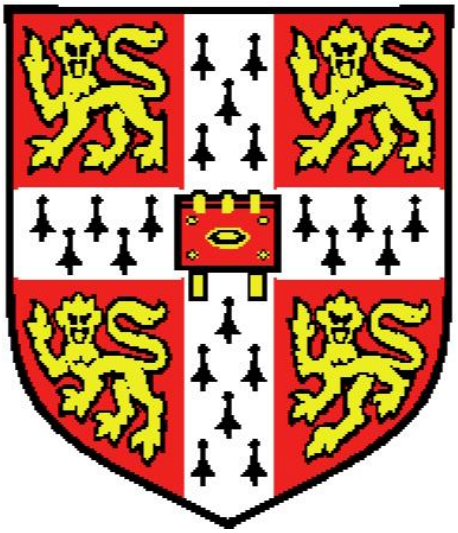
SUPPORTING DISEASE CANDIDATE GENE DISCOVERY BASED ON PHENOTYPE MINING

ANIKA OELLRICH

Wolfson College



A dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy

European Molecular Biology Laboratory
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD
United Kingdom

anika@ebi.ac.uk

29th August 2012

To my grandparents and parents.

DECLARATION

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university; and no part has already been, or is currently being submitted for any degree, diploma or other qualification. This dissertation does not exceed the specified length of 60000 words as defined by The Biology Degree Committee.

Cambridge, 29th August 2012

Anika Oellrich

SUPPORTING DISEASE CANDIDATE GENE DISCOVERY BASED ON PHENOTYPE MINING

— *by Anika Oellrich* —

Even though numerous biological and computational experiments have been devoted to the understanding of the molecular mechanisms underlying human genetic disorders, a large number of those disorders is still without identified genetic mechanisms. Genetic pleiotropy as well as the polygenic nature of some human genetic disorders pose challenges which still need to be overcome before an understanding of the disease underlying molecular mechanisms may be achieved. Mouse models are used to extensively study human disease through mutagenesis experiments and their findings are reported in publicly accessible databases as well as scientific publications. In my thesis, I focus on supporting the identification of disease gene candidates by mining phenotype information from three different resources: the Mammalian Genome Informatics (MGI) database, the Online Mendelian Inheritance in Man (OMIM) database and the scientific literature.

To enable the integration of the resources, I developed a pipeline which ranks mouse models for human genetic disorders and with that enables the identification of promising disease gene candidates. Mouse models are ranked according to their phenotype similarity and hence the ranking pipeline can be used as long as a phenotype description of the human disorder is at hand. No prior information about the genetic causes is required which makes this approach especially valuable in the case of orphan diseases, where it is hard to identify the molecular mechanisms due to their rare occurrence. Furthermore, I generated mouse-specific disease profiles and demonstrate their validity by applying them to the mouse model ranking pipeline and evaluating the obtained results against disease gene reporting databases. Those mouse-specific profiles may further broaden our knowledge about genetic diseases by using them for annotation enrichment. To illustrate their potential, I applied the mouse-specific profiles to a disease classification task. Manual investigation of the obtained classification results reveals phenotypes to enrich existing OMIM disease annotations.

Due to the incompleteness of the existing phenotype resources and the intense labour and time consumption of manual curation, my work also focussed on the extraction of phenotype information from scientific literature. Not only the abstract but also every other part of a scientific publication is analysed for its potential to provide phenotype information. Textual features as well as several ontologies are used to identify and extract phenotype mentions into a formal representa-

tion. The extracted phenotypes can then be used to provide database curators with a selection of phenotypes contained in a paper and by doing so speed up the curation process. Thus, literature extracted phenotypes enrich existing phenotype databases and consequently support the data mining efforts based upon phenotype information.

ACKNOWLEDGEMENTS

First of all, I would like to thank my PhD supervisor Dietrich Rebholz-Schuhmann for his continuous support, his guidance, his encouragement to follow my intuitions and consequently learn to guide my own research. Dietrich was always very supportive with last minute paper submissions and travel funds to several workshops and conferences.

Furthermore, I would like to thank Robert Hoehndorf – LLAP \/ – and George Gkoutos. I collaborated closely with Robert and George, especially in the last three years of my PhD. Both provided me with insights into biological and computational aspects of phenotype mining and spent numerous hours discussing with me (I know GeorgioUs, you are always right apart from 1986!).

In addition, I would like to thank Robert Busch, Peer Bork and Anton Enright who, as part of my Thesis Advisory Committee, have challenged my ideas and therefore let me gain valuable insights through discussion.

I am very grateful to everyone who has proof-read parts of this thesis, especially those reading substantial parts of it and having had to put up with lots of questions - refunds and compensations will be provided!

My special thanks also go to Christoph Grabmueller and Maria Liakata. Christoph supported me in technical questions while Maria provided insights into the theory of research work.

I further would like to thank Adam Bernard who has not only become a friend over the last couple of years but also provided a lot of feedback to my work.

My research group was also very supportive with various discussions, prep-talks and journal clubs. I would like to thank Samuel, Yumi, Senay, Chen, Irina, Silvestras, Jee-Hyub, Antonio, Ian, and Shyama for their continuous help!

The vivid predoc community at EBI was also a great support, not only with morning coffees and lunch time seminars. I would like to especially thank the predocs from my year – Tim, Inigo, Pablo, Joe, Nenad, the pink unicorn, and Charlie; those from previous years – Joern, Dominic and Dace; and those who were adopted – Dagmar, Claire and Mikhail.

Heidi and Elin, thank you very much for helping me when I needed it the most! Heidi – many thanks for all the meetings and sharing your knowledge with me.

I am also very thankful for having found a role model during those years at EBI – Mela, thank you very, very much for the way you are and the inspiration and guidance you gave me for almost six years now!!!

Last, but by no means least, I would like to thank the individuals closest to me. First, I would like to thank my family who provided me with the stepping-stones to become what I am and were supportive all the way long: my parents – Christina & Ulrich; my brother and his family – Daniel, Doreen, Lara & Lena; my grandparents – Brunhilde & Heinz. And finally many, many thanks to Steviiie & Bobo: *dankeschon, dankeschon, ich bin ganz comfortable, Kartoffelkopf*.

CONTRIBUTIONS

Contributions to the individual parts of this thesis are listed here according to chapter.

CHAPTER 2

I analysed and recorded the co-existing ways of describing phenotypes, derived the categorisation and drafted the manuscript for the workshop. Dietrich Rebholz-Schuhmann supervised the work and contributed to the workshop manuscript.

CHAPTER 3

I designed the study with the help of Robert Hoehndorf. I implemented the required R and Groovy scripts. Robert Hoehndorf contributed two Groovy Scripts, and generated the combined mappings. Robert Hoehndorf and Dietrich Rebholz-Schuhmann supervised the research. Georgios Gkoutos and Robert Hoehndorf helped with the validation of the initial pairs and the biological validity of the results. All of us contributed to the submitted manuscripts.

CHAPTER 4

I designed the study and implemented the required groovy scripts as well as the web interface. Robert Hoehndorf and Dietrich Rebholz-Schuhmann supervised the work. Robert Hoehndorf provided the classification of the diseases based on the mouse profiles. George Gkoutos helped with the manual evaluation of the classification results. All contributed to a manuscript which is to be submitted soon.

CHAPTER 5

Christoph Grabmueller implemented a general set-up of the annotation servers, while I derived the domain specific dictionaries underlying those servers. Irina Colgiu provided a speed-improved and corrected version of the Gene Ontology (GO) server implementation described in (Gaudan et al., 2008). I designed the study and implemented all the other required software using Groovy. I also manually evaluated a subset of the error cases in the entity–quality (EQ) statements and provided feedback to Georgios Gkoutos. Georgios Gkoutos updated the logical definitions accordingly and provided additional

information in unclear cases. Dietrich Rebholz-Schuhmann supervised the study.

CHAPTER 6

I designed the study and implemented all required testing software as Groovy scripts. I also carried out the manual curation of the generated EQ statements to find flaws and allow for the definition of generalised patterns in the decomposition process. Doubtful cases were clarified with Georgios Gkoutos as were cases in which an error in the EQ statements was suspected. Georgios Gkoutos corrected the EQ statements accordingly and provided support. Dietrich Rebholz-Schuhmann supervised the studies. The server set-up was the same as in chapter 5, partially contributed from Christoph Grabmueller and Irina Colgiu. Christoph Grabmueller, Dietrich Rebholz-Schuhmann and me, we all contributed to the submitted manuscript.

GLOSSARY

The following terms are used throughout my dissertation.

Genotype	A genotype is the entire genetic information of an organism.
Genome	A species genome comprises all genes appearing in this species.
Phenotype	A measurable or observable characteristic of an organism.
Phenome	A species phenome is the collection of all phenotypes occurring in this species.
Disease	A disease affects the body of an organism (mostly in an impairing or dysfunctional sense) and causes the phenotypes of this organism to differ from an as *normal* defined state.
Ontology	An ontology is the conceptualisation of knowledge in a domain.
Concept	A concept in an ontology is one of the building blocks to construct the conceptualisation.
Text Mining	Text mining is a research domain focusing on the retrieval, extraction and discovery of knowledge from text.
Special character	A special character is a character which is neither alphabetic nor a number.
Word	A word is a consecutive string of characters which have a meaning attached to them and can stand on their own. Words as opposed to terms may have several meanings at the same time and the context of the word determines its meaning.
Token	A token is a consecutive string of characters and numbers, potentially also including special characters but no white spaces. In text, tokens are joined together with white spaces to form sentences. As opposed to a word, a token does not necessarily possess a meaning.
Term	A term is a word that has been given a specific meaning in a specific context.
Stop word	A stop word is a word considered not to carry a lot of information and which is therefore removed when applying text mining.
Information content	The information content of a word is a measure to quantify the information content carried by the word.

Annotation	An annotation is an assigned ontology concept to specifiy the meaning of the entity it is attached to, e.g. the content of a database entry.
Annotation server	An annotations server processes text and assigns annotations to text spans, e.g. according to a defined dictionary.
Curator	A curator is a person who manually assigns annotations.

CONTENTS

1

INTRODUCTION

Numerous biological as well as computational efforts are ongoing to identify the underlying molecular mechanisms of human genetic diseases. Despite those efforts, a large number of the diseases is still without an identified genetic basis (Amberger et al., 2011). Genes impacting multiple characteristics of an organism (genetic pleiotropy) as well as the polygenic nature of a subset of human genetic diseases pose additional challenges to the quest of identifying the key players underlying disease (Raychaudhuri, 2011). Only if we are able to fully understand the genetic causes of a disease will we be able to find reliable treatments or prevention mechanisms.

Since the discovery of the interplay of genotype and phenotypes, which are observable/measurable characteristics of an organism, genotype–phenotype associations are recorded to improve our biological understanding. With the increasing availability of technologies to determine the genotype of an organism, large-scale efforts are ongoing to systematically record phenotypes corresponding to specific genotypes. For instance in mouse, an almost completed stem cell library exist, covering all genes in the mouse genome (Dolgin, 2011). In order to identify the resulting phenotype, mutant mice have to be bred from the stem cell lines. Projects, such as EuroPhenome (Morgan et al., 2009), record mouse phenotypes and make them available to the research community via databases. The resulting wealth of data cannot be handled manually and requires computational efforts to thoroughly analyse the data.

An independent research domain, bioinformatics, developed which helps with the integration and analyses of biological data. Bioinformatics originates from the developments in the 1960s to apply computers to solve complex biological problems (Hagen, 2000; Ouzounis and Valencia, 2003). For instance, improving the understanding of biological macromolecules with computers led to significant progress with the understanding of the evolution of genes and proteins (INGRAM, 1961; Zuckerkandl and Pauling, 1965). In the early 1970s, Paulien Hogeweg and Ben Hesper proposed the term Bioinformatics to refer to "the study of informatic processes in biotic systems" (Hogeweg,

2011). During the 1980s, bioinformatics evolved to be an independent research domain. Wide-spread applications have been developed to analyse not only genome data but also to model complex biological systems (Hogeweg, 2011).

Alongside genome wide association studies (GWAS) (see 1.3), phenome wide association studies (PheWAS) evolved, aiming to computationally identify connections between a genotype and a phenotype. PheWAS have gained significant importance in the quest of identifying disease genes by making use of the available phenotype resources, such as the EuroPhenome database. Some of the approaches focus on the development of single species phenotype networks, e.g. for human (Lage et al., 2007) or mouse (Espinosa and Hancock, 2011). Others facilitate the integration of phenotypes across different species (Hoehndorf et al., 2011c; Washington et al., 2009; Groth et al., 2007) leading to the discovery of new genotype-phenotype associations (McGary et al., 2010; Leach et al., 2009). Despite the achievements of PheWAS, Lussier and Liu judge:

> *"There is a scarcity of phenotypic discovery methods, theories, and predictions to exploit the rich and untapped phenotypic data repositories in current genetic model organism databases [...] The lack of high-throughput technologies to access well-networked and integrated phenotypes from heterogeneous sources and across multiple scales of biology under homoeostasis or disease conditions has prevented the effective use of phenotypic information." (Lussier and Liu, 2007)*

Other reviews on the availability and application of phenotype data present similar conclusions (Groth and Weiss, 2006; Gkoutos et al., 2012).

Therefore, the work described in this thesis focuses on supporting the identification of disease gene candidates by mining phenotype data. Mining phenotype data does not only include the integration and analysis of the data, but also includes the enrichment of existing phenotype resources. The work presented here was geared towards the development of a solution (called *InterPhen*) that would enable its users to integrate a variety of phenotype data repositories. Ultimately, *InterPhen* would allow for the prioritisation of genes for human diseases based on the integrated phenotype data. Moreover, not only human data would be used to derive gene candidates but also data from model organisms, e.g. obtained through biological investigations in mice.

With the increasing amount of available genotype data, most costs nowadays lie in phenotyping (Bilder et al., 2009). Automated methods can help reduce the costs by computationally determining relevant phenotype information. For example, Oti et al. (2009) showed in an initial study that phenotype data can be enriched by integrating human phenome databases. Following on from this study, I investigated the potential of using phenome databases covering several species to enrich existing phenotype descriptions of human diseases. The data used in this study was obtained from both the Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2011) (see 2.4.1) and the Mouse Genome Informatics database (MGD) (Blake et al., 2011) (see 2.4.2).

In addition to species-specific phenotype databases, I used the scientific literature to extract phenotype information for enrichment purposes. Due to most research work being published in scientific journals, published literature constitutes a rich resource of biological knowledge (Jensen et al., 2006). Initial studies have shown that the systematic analysis of phenotypes contained in scientific literature enables biological discoveries (Korbel et al., 2005). Furthermore, it has been shown that phenotypes extracted by text mining (TM) show biologically valid connections. Therefore, extracted phenotypes can facilitate biological knowledge discovery (van Driel et al., 2006). Even though generalised extraction systems exist, e.g. the Open Biomedical Annotator (Jonquet et al., 2009) (see 1.4.3), text mining has not yet successfully targeted the systematic extraction of phenotype data from literature and a demand for extracted phenotype information still exists (Hirschman et al., 2012). A more detailed description of phenotype extraction methods in the field of TM is provided in section 1.4.3.

This chapter aims to give an introduction to human genetic diseases, their heritability and the application of model organisms, in particular mouse models, to study human diseases. Furthermore, this chapter includes an overview of existing computational methods to identify disease gene candidates. In addition, it covers methods facilitating the enrichment of existing data through text mining. An insight into ongoing standardisation efforts in the biological and biomedical domain is also provided. Ontologies used for standardisation as well as their alignment for data integration are covered. In summary, the work presented in this thesis comprises a set of methods and insights

that can facilitate biological discovery in the field of human genetic diseases.

1.1 HUMAN GENETIC DISEASES

According to the online version of the Oxford English Dictionary[1], the second definition of disease is as follows:

> *"A condition of the body, or of some part or organ of the body, in which its functions are disturbed or deranged; a morbid physical condition; 'a departure from the state of health, especially when caused by structural change' (New Sydenham Soc. Lexicon). Also applied to a disordered condition in plants. "*

However, this definition focuses only on the consequences a disease has on an organism. It does not cover the potential origins of a disease. Diseases may be either caused by external organisms such as bacteria or viruses, or inflicted by the genotype of an organism placed into a certain environment. Consequences on an organism are usually described as signs and symptoms (phenotypes).

In the case of a disease possessing heritable origins, knowing the underlying molecular mechanisms will help to determine individuals at risk through genetic screening. Furthermore, with knowing the genetic causes, effective treatments can be developed as well as possible prevention mechanisms identified (Pritchard and Cox, 2002) .

Even though numerous efforts aim to elucidate the molecular mechanisms underlying human heritable diseases, a large number of those diseases is still without identified genetic background (Amberger et al., 2011). To support the quest for disease genes and facilitate communication about the diseases, databases have been developed to catalogue the diseases. Examples of these databases include the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2011) (see 2.4.1) and the OrphaNet database (Aymé, 2003; Aymé and Schmidtke, 2007; Rath et al., 2012). The databases also contain information about the molecular mechanisms where identified: e.g. genes, allelic variants or a chromosomal region to which the disease has been linked.

1 June 2012; Oxford University Press; `http://oed.com/view/Entry/54151?rskey=I70C5B`

1.1.1 *On the importance of phenotype information*

Phenotype information plays a crucial role in both the clinical as well as the biological domain. Even though both domains focus on recording phenotype information, definitions as to what constitutes a phenotype differ. In the biological domain, phenotypes are an observable, and in most cases also measurable characteristic of an organism. The entirety of all phenotypes existing within a species are referred to as the phenome of this species. Despite the suggestion to unify the collection of human phenotype data (Freimer and Sabatti, 2003), the human phenome is not defined as yet. In other species, however, the definition of the phenome is almost accomplished, e.g. in mice (Schofield et al., 2011).

In the clinical domain, human diseases are described by specific phenotypes, which are also referred to as signs and symptoms. An individual suffering from a disease exhibits one or more signs and symptoms used for the definition of the disease. However, the definition of a disease based on the exhibited phenotypes is an iterative process (Brunner and van Driel, 2004). One aspect requiring iterations is that, and especially occurring in case of genetic diseases, phenotypes are likely to change in severity in subsequent generations while the age of onset may decrease (Warren and Nelson, 1993). Another aspect requiring iterations is that some diseases possess low occurrences in a population. In this case, it is hard to define a disease by a *common* pattern and the pattern may change with each new occurrence of the disease. Due to these iterations, diseases are redefined over time and may even disappear, e.g. if they are merged, or split into more than one syndrome (Amberger et al., 2011; Brunner and van Driel, 2004).

Despite the best efforts in defining a disease, it may happen that important phenotypes are missed (Schofield et al., 2011). This is potentially caused by limitations in the targets of applied screening methods in human, or because effects are only visible on a molecular level. In experimental set-ups, such as those to assess the phenome of a species (see introduction of this chapter), a protocol is followed, leading to a standardised way of assessing the phenotype corresponding to a genotype. Consequently, thorough analyses of mouse models can support the identification of relevant phenotypes in human and its success has been shown in the past (Lisse et al., 2008).

In my work, I will focus on the definition used in the biological domain that phenotypes are observable and mostly measurable charac-

teristics of an organisms. I chose to work with the *biological* phenotype definition because the amount of available phenotype data coming from *biological* investigation is simply larger than accessible *clinical* data. While *biological* phenotypes are reported in different species, *clinical* data reports about humans. *Clinical* phenotypes can be transferred to *biological* phenotypes to facilitate translational research, with topics spanning further than identifying the causes for heritable diseases in human (see 7.3).

1.1.1.1 *Types of phenotype information*

Phenotypes reported for diseases and gathered in biological experiments span different areas: morphology, biochemistry, physiology, and behaviour (Freimer and Sabatti, 2003). Phenotypes also cover different levels, reaching from a molecular level to the environment an organism lives in (Oellrich and Rebholz-Schuhman, 2010). This subsequently means that phenotypes do not only address quantitative and qualitative attributions, such as *five* or *high/low*, but they also include:

- locations (either within the organism or the location of the organism itself), e.g. *liver* or *cage*
- processes, e.g. *insulin secretion*
- chemicals, e.g. *cholesterol*
- absences and presences, e.g. *absent carpal bone*

To ensure comparability of experimental results and facilitate large-scale analysis of the data, standardisation mechanisms are required. Research communities have responded with introducing species-specific phenotype projects, e.g. the mouse phenome project (Paigen and Eppig, 2000; Bogue and Grubb, 2004; Bogue, 2003), or consortia, e.g. the International Mouse Phenotyping Consortium[2] (Collins et al., 2007) to harmonise phenotype screens and environments to facilitate comparative analyses. Despite the efforts in other species and the suggestion of a human phenome project (Freimer and Sabatti, 2003), the initiation of such a project is still to come.

Chapter 2 provides an overview of co-existing ways to describe phenotype information.

2 http://www.mousephenotype.org/

1.1.2 *Heritability of diseases*

Heritability of phenotypes and consequently diseases is determined by DNA (Avery, 1915). DNA is coiled together into chromosomes, and each chromosome is further segmented into heritable units called loci or genes. Human cells are diploid cells and usually possess 23 pairs of chromosomes: 22 pairs of autosomes (gender-neutral chromosomes) and one pair of allosomes (gender-specific chromosomes). Some of the human genetic disorders affect the number of chromosomes in a cell, e.g. trisomy 21 (MIM:#190685) in which case a third copy of chromosome 21 exists.

Each of the genes contained in a cell can vary with respect to its sequence. Each individual sequence variant is an allele of the corresponding gene. Because of human cells being diploid, each cell possesses two copies of the genes located on the autosomes (for females also the genes located on the allosomes). Those two copies of the same gene may be either the same or different alleles. If a cell holds the same allele for a gene on either chromosome, it is said to be *homozygous* for this gene. If different alleles for the same gene are contained within a cell, this cell is referred to as being *heterozygous* for this particular gene.

In the process of gene mapping, genes are mapped to their function and corresponding phenotype. However, there is no one-to-one correspondence between a gene and a phenotype. One gene can cause multiple phenotypes (genetic pleiotropy); conversely, one phenotype can also be caused by multiple genes (polygenic phenotype). An example of genetic pleiotropy is phenylketonuria (PKU) (MIM:#261600). PKU is caused by deficiency of phenylalanine hydroxylase (UniProt:P00439) but leads to phenotypic effects, e.g. mental retardation and eczema (Lobo, 2008). In Mendelian diseases, a phenotype is frequently caused by different genes (Brunner and van Driel, 2004; Roberts et al., 2002).

Even though the causing gene may differ, Mendelian diseases are caused by the impact of single genes. But not all heritable diseases are Mendelian diseases. Some of the heritable diseases are caused by a certain set of genes, each contributing their share to the disease phenotype(s). Those diseases are called polygenic diseases (due to multiple genes contributing) and most human genetic diseases are polygenic in nature (Beckers et al., 2009).

Further to the influence of the genetic background, modifier genes may also play a role in human diseases. Modifier genes may change

their behaviour over the lifespan of an organism. Consequently, phenotypes and potential diseases impacted by modifier genes will change over time. Reasons for modifier genes to change over time could be environmental influences (Nadeau, 2001; Dipple and McCabe, 2000).

1.1.3 *Investigating human genetic diseases with animal models*

Even though projects are underway to assess the variation of the human genome (Consortium, 2010a), it is impossible to specifically modify genes and chromosomal regions in humans and assess the outcome. Due to the human genome possessing approximately 30,000 genes, chances are small that the influence of individual genes as well as their combination, will become visible in variation studies. With only studying small groups of individuals, it is hard to reliably establish patterns defining causative genetic components for phenotypes. Therefore, model organisms, such as mouse, zebrafish and fruit-fly, are used to investigate the influence of the genes and their combination on the phenotypes. The results of those experiments are then transferred to human by comparing phenotypes or genetic sequences.

Underlying comparative functional and phenotype studies is the assumption that orthologous genes also possess a conserved function and subsequently a conserved phenotype. Orthologous genes are genes with a conserved sequence in different species. Even though this assumption holds largely true, examples exist contradicting this assumption. Therefore careful consideration of results is required when applying comparative gene function and phenotype studies.

Another aspect which has to be taken into consideration when applying animal models to the study of human diseases, is the way of how phenotypes are reported in animal studies. Firstly, only phenotypes are reported which are required by the protocol of the experiment. Secondly, a phenotype is only reported when it constitutes an *abnormal* outcome for this particular experiment. For example, if studies are carried out in *obese* mice, this information will not be reported as a result unless the experiment also has an enormously increased effect on the body weight.

Even though those aspects have to be taken into consideration, mouse models are considered to be a good choice when investigating human diseases as both share 99% of their genes (Rosenthal and Brown, 2007). In mouse, an almost completed stem cell library exist, providing a knockout for each gene in the mouse genome (Skarnes

et al., 2011). An additional project was launched to comprehensively record the consequences on the phenotypes caused by gene knockouts (Abbott, 2010).

Outcomes of mutagenesis projects are stored in community supported model organism database (MOD)s, e.g. the Mouse Genome Informatics database (MGD) (Blake et al., 2011) or FlyBase (Drysdale and Consortium, 2008). Leonelli and Ankeny (2012) assessed four MODs in terms of content and biological consequences. They found that the biological content in those databases shapes the work of communities. Leonelli and Ankeny (2012) also argued that those databases facilitate a structured representation of knowledge. A structured representation of data improves cross-species comparability of data. However, with the variety of existing phenotype descriptions (see 2), automated analysis of those databases is still challenging.

1.2 TOWARDS THE STANDARDISATION OF BIOLOGICAL KNOWLEDGE

Not only is the amount of published research papers increasing, but so is all sorts of other data. For instance, with the recent improvements in sequencing technologies, an abundance of genome data are available for analysis. In order to facilitate large-scale automated analysis of these data, each individual data set has to be recorded in compliance with a standard. The application of a standard will also provide a fixed meaning for the data content.

Due to the need of standardised knowledge, ontologies have more and more found their way into biological and biomedical research. Ontologies provide the specification of the conceptualisation of a domain (Gruber, 1995) and therefore may provide the means to retrieve, integrate, and compare several different data sets (Rubin et al., 2008).

A requirement for the retrieval, integration and comparison of data, is the assignment of annotations to individual data sets. An annotation in this context is a specifically selected ontological concepts to represent the content and meaning of the data set. An annotation may even cover the aspect of how the data set was collected.

To illustrate the role of ontologies in biology and biomedicine, the next sections provide an insight to the existing ontologies (1.2.1), what possibilities exist to assign annotations (1.2.2) and how data sets can be integrated using ontology annotations (1.2.3).

Part of the standardisation efforts target the representation of phenotypes and their application to annotate and enable integration across

phenotype repositories. This section aims to provide a general introduction to ontologies and their use in the biological and biomedical domain and is not targeted to phenotype ontologies in particular. Chapter 2 provides a detailed view on the availability of phenotype descriptions, phenotype ontologies and their usage.

1.2.1 *Ontologies in the biological and biomedical domain*

Due to the enormous increase in available data sets in the biological and biomedical domain, the ontologies used for the description of biological data have also increased in size and numbers. An ontology does not only provide concepts, but also holds relations, functions, and other objects to describe a domain of interest (Gruber, 1993).

Ontological objects are described using a formal language. Additionally, textual data are used to assign a concept label, synonyms, and a textual definition. Links are provided to cross reference the object to other data collections, e.g. databases. The development of each ontology is an iterative process. User communities and discussion groups exist to evaluate the existing concepts and their relations, and decide upon the insertion, correction or deletion of concepts or relations.

An example of a highly axiomatised ontology in the biomedical domain is SNOMED Clinical Terms (SNOMED CT)[3]. SNOMED CT is extensively applied in the biomedical domain, mainly for the purpose of annotating Electronic Health Records. Furthermore, SNOMED CT is integrated into the Unified Medical Language System (UMLS) (Bodenreider, 2004) (see 1.2.3). Thus, SNOMED CT is aligned to other clinical ontologies and terminologies. Due to the integration into UMLS, data that are annotated with SNOMED CT can be broadly integrated with other data, even those not necessarily represented with the same terminology.

To define and develop ontologies, different formats are available, e.g. the OBO Flatfile format[4] (Horrocks, 2007) or the Web Ontology Language (OWL)[5] (Grau et al., 2008). There are several tools to support ontology development in the different formats, e.g. the OBO–Edit editor (Day-Richter et al., 2007) or Protégé [6]. To ensure a seamless integration between both ontology formats, several projects have ad-

3 http://www.ihtsdo.org/snomed-ct/, historic and content overview available from http://ontogenesis.knowledgeblog.org/834

4 http://www.geneontology.org/GO.format.obo-1_2.shtml

5 http://www.w3.org/TR/owl2-overview/

6 http://protege.stanford.edu/

dressed the conversion from one to the other (Tirmizi et al., 2011; Moreira and Musen, 2007; Hoehndorf et al., 2010a).

Repositories such as the OBO Foundry[7] (Smith et al., 2007) or BioPortal[8] (Whetzel et al., 2011; Noy et al., 2009) evolved to provide a central storage space for ontologies. For an ontology to be included into the group of OBO Foundry ontologies, it has to fulfil a list of criteria[9]. Those criteria aim to ensure that an ontology is orthogonal to the existing ontologies and it possesses sufficient quality to be provided to the broader public. Ontologies not fulfilling those criteria are handled as *OBO Foundry candidate ontologies* until the criteria are met.

1.2.1.1 *The Gene Ontology*

One ontology that gained a lot of attention is the Gene Ontology (GO) (Ashburner et al., 2000; Pesquita et al., 2008). GO is used by numerous databases to annotate their content, e.g. MGD and the Universal Protein resource (UniProt) (Consortium, 2010b). More and more algorithms emerge using GO annotations to automatically analyse the content of those databases, e.g. MedSim (Schlicker et al., 2010) or GOToolBox (Martin et al., 2004).

GO is developed in the OBO Flatfile format and has been separated into three independent parts: *biological processes*, *cellular components*, and *molecular function*. GO aims at the unification of biological knowledge across different species, and it also facilitates the comparison of data by comparing assigned GO concepts. Figure 1 provides a small selection of concepts from the *molecular function* part of GO to illustrate the structure of GO. To illustrate concepts definitions in GO, the definition of the concept *nucleus* (GO:0005634) is provided here:

```
[Term]
id: GO:0005634
name: nucleus
namespace: cellular_component
def: "A membrane-bounded organelle of eukaryotic cells in which
      chromosomes are housed and replicated. In most cells, the
      nucleus contains all of the cell's chromosomes except the
      organellar chromosomes, and is the site of RNA synthesis
      and processing. In some species, or in specialized cell
      types, RNA metabolism or DNA replication may be absent."
```

7 http://obofoundry.org/
8 http://bioportal.bioontology.org/
9 http://www.obofoundry.org/wiki/index.php/Category:Principles

```
      [GOC:go_curators]
subset: goslim_aspergillus
subset: goslim_candida
subset: goslim_generic
subset: goslim_metagenomics
subset: goslim_pir
subset: goslim_plant
subset: goslim_yeast
synonym: "cell nucleus" EXACT []
xref: NIF_Subcellular:sao1702920020
xref: Wikipedia:Cell_nucleus
is_a: GO:0043231 ! intracellular membrane-bounded organelle
}
```

An ontology file described in the OBO Flatfile format, is a collection of concept definitions as shown in the aforementioned example. For each concept, it is specified how it relates to the other concepts of the ontology. With the definition of concepts and relations between those concepts, an ontology forms a graph-like structure as illustrated in 1.

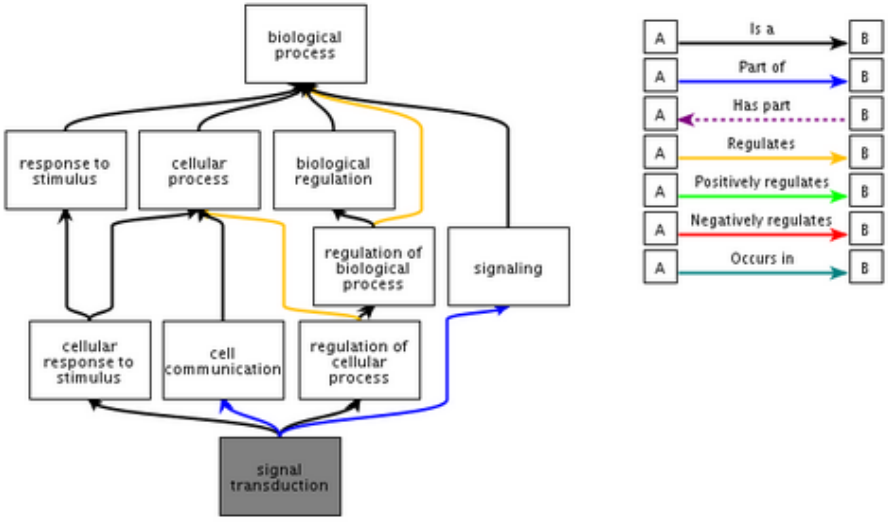


Figure 1: **Sub-tree taken from GO to illustrate its structure.** View generated by using QuickGO (`http://www.ebi.ac.uk/QuickGO/`), searched for concept *signal transduction* (GO:0007165).

1.2.2 *Annotating biologically relevant content*

A requirement for data integration and knowledge discovery in translational research are annotations (Bodenreider and Stevens, 2006). An annotation is an assigned ontology concept to an entity, e.g. a

Concept ID	Concept name
GO:0003677	DNA binding
GO:0003682	chromatin binding
GO:0003700	sequence-specific DNA binding transcription factor activity
GO:0008022	protein C-terminus binding
GO:0043565	sequence-specific DNA binding
GO:0047485	protein N-terminus binding

Table 1: GO *molecular function* **annotation for the human gene HESX1 (UniProt:Q9UBX0).** The annotations are only a subset of all available gene annotations. Evidence and origin of each of the annotations is also suppressed. Full information is available from `http://www.ebi.ac.uk/GOA/` and `http://www.ebi.ac.uk/QuickGO`.

database entry, to clarify its meaning. Large biological databases such as Universal Protein resource (UniProt) (Consortium, 2010b) and ArrayExpress (Parkinson et al., 2011), also adhere to the annotation of data to describe the biological content of those databases.

As an example, table 1 shows the GO annotations of the human gene HESX1 (UniProt:Q9UBX0). Those GO annotations represent its function and are extracted from the Gene Ontology annotation (GOA) database[10] (Barrell et al., 2009). Only a subset of the data provided in the original source is shown here to reduce complexity.

By now, the GOA database covers more than 32 million annotations for over 4.3 million proteins in over 160,000 taxa. As of September 2012, only 467,134 annotations out of over 32 millions were gathered manually (Barrell et al., 2009). Each annotation contained in GOA possesses an evidence code indicating on how this annotation was retrieved. For example, the evidence code *IEA ('Inferred from Electronic Annotation')* symbolises that annotations with this evidence code were retrieved by automated means. The diverse methods of gathering annotations for biological content and the population of databases with annotations is illustrated in the following section.

1.2.2.1 *Methods of automated annotation assignment*

Manually assigning ontology annotations is tiresome, costly, and time-intense (Jaeger et al., 2008). However, annotations are most reliable if done by expert curators applying their knowledge to the annotation assignment process (Shah et al., 2009; Barrell et al., 2009). With the ever

10 `http://www.ebi.ac.uk/GOA/`, `http://www.ebi.ac.uk/QuickGO`

growing data in literature and databases, it is impossible to handle all annotation work manually (Krallinger et al., 2008; Bairoch, 2009). To improve the long curation times, reliable methods are sought which allow for automated annotation of biological data. Some showcases are provided in the following paragraphs to illustrate the landscape of automated ontology annotation assignment.

For example, Shatkay et al. (2007) developed a system called SherLoc to predict the cellular location of proteins. Further to protein sequence similarity, SherLoc uses text mining to identify the location of a protein in a cell. Overall, SherLoc predicts protein locations with a reported accuracy of 71% and constituted a significant improvement over other methods. Experimental verification of the results was not available at the time of publication.

Jaeger et al. (2008) used conserved protein interaction graphs to predict protein function. Annotations are transferred from orthologous proteins contained in the conserved graphs. The obtained protein function predictions are then further validated with results reported in the scientific literature to increase the precision of the method. This validation step leads to a final set of protein function annotations. A manual evaluation by domain experts showed that all final protein function annotations obtained by this method were correct.

Magi et al. (2012) introduced the Weighted Network Predictor (WNP) to identify functional annotations of gene products. The WNP prediction algorithm relies upon species-specific Bayesian gene product interaction networks. Annotations are propagated from characterised to previously uncharacterised gene products. Applying this method, the authors demonstrated to outperform five state-of-the-art approaches and to automatically annotate 500 and 10,000 previously uncharacterised gene products in Saccharomyces cerevisiae and Arabidopsis thaliana respectively.

Automated methods contribute a huge number of annotations to public databases (see 1.2.2, example GOA database). However, an influence of automated annotations on data analysis methods cannot be neglected. The incorporation of automated annotations may lead to data circularity (Pesquita et al., 2009). For example, if gene sequence similarity is used to derive annotations and those annotations are then applied to sequence studies applying clustering, genes used for annotations will consequently cluster together. Thus, each study based upon annotations, should carefully investigate where annotations orig-

inate from and whether this may impact the analysis results (Pesquita et al., 2009).

1.2.3 *Integration of knowledge via ontologies*

With the advent of diverse ontologies spreading to annotate biological content, mechanisms are required that facilitate the seamless integration of data annotated with different ontologies (Agrawal et al., 2008). The process of finding corresponding concepts from different ontologies is called *ontology alignment*, or in terms of vocabularies or terminologies *term alignment*. One example of a resource resulting from a large-scale alignment process is the Unified Medical Language System (UMLS) (Bodenreider, 2004). As of 15 August 2012, it incorporates over 150 terminologies, vocabularies and ontologies concerned with the description of clinically and disease relevant information in human. SNOMED CT (see 1.2.1) is one of the included clinical ontologies.

A variety of alignment tools have been developed and the Ontology Alignment Evaluation Initiative (OAEI) campaigns[11] aim to provide a fair assessment of the existing tools. A unified benchmark set is provided to determine the performance of each system, e.g. by determining precision and recall scores. The benchmark set, together with the performance results and suggestions for possible improvements, is published as a report file (Euzenat et al., 2011).

More recent than the aforementioned report is a review from Shvaiko and Euzenat (2011), also covering some of the tools repeatedly presented and evaluated in OAEI campaigns. A short summary about the different types of tools is provided here focusing on two different criteria: features used for alignment and characteristics of the resulting ontology alignment produced by the tools.

Alignment of ontologies has been achieved considering different characteristics of the ontologies. One group of algorithms focuses solely on lexical features, e.g. Lexical OWL Ontology Matcher (LOOM) (Ghazvinian et al., 2009) or OntologyMapper[12]. LOOM aligns concepts based on a pairwise comparison of the strings of the labels and synonyms used to describe a concept (see 1.2.1.1 and 3.2.3.1). Concepts either sharing their label or synonym are recorded and the alignment is provided as a list of concept pairs. OntologyMapper aligns concepts

11 http://oaei.ontologymatching.org/
12 `http://www.ebi.ac.uk/efo/tools/`

based on their phonetics and supports the Metaphone and Double Metaphone algorithms (Philips, 2000). It also provides list of aligned concept pairs.

A second group of alignment algorithms does not only include the lexical information provided by the ontologies but also considers ontology features. Ontology features are e.g. particular relationships holding between the concepts. For instance, Zhang and Bodenreider (2003) did not only use the textual content but also used the ontologies' taxonomies to align anatomy ontologies. AgreementMaker (Cruz et al., 2009) also employs an algorithm focusing on labels and synonyms as well as ontological features such as information about ancestor and sibling concepts. The ancestor and sibling information can be used to correct semantic similarity scores of concepts.

Alignment algorithms can be further classified according to the output they produce. While some alignment tools produce a one-to-one alignment (Hu and Qu, 2008; Li et al., 2009) others produce a many-to-many alignment (Cruz et al., 2009). A one-to-one alignment means that one ontology concept is aligned to only one other concepts and mappings are symmetrical. In the case of a many-to-many alignment, one ontological concept can be aligned to multiple other concepts and the resulting mappings are not necessarily symmetrical.

Aside from classification aspects of alignment tools, another example of alignment mechanism should be only mentioned here (further explained in 2.2.2.2): entity–quality (EQ) statements. EQ statements are particular to phenotype representations and are used to bridge species-specific phenotype ontologies.

1.3 AUTOMATED DISEASE GENE DISCOVERY

Traditionally, two complementary approaches have been applied to the identification of genetic components underlying human disease: linkage analyses (Taylor et al., 1997) and genome wide association studies (GWAS) (Lunetta, 2008). Linkage analyses are carried out in families and are most suitable for finding rare variants with considerable penetrance causing severe changes in phenotypes. GWAS work on population samples of unrelated individuals and are able to identify common variants with moderate phenotype changes. Despite both approaches having delivered substantial contributions, their success is largely dependent on the disease causing alleles aggregating in families or sample population being studied. Furthermore, those studies

are mostly targeted towards one disease or a small number of similar diseases and cannot be applied to the large number of diseases which are still without identified cause.

In recent years, numerous tools to automatically identify the genetic causes of diseases have been developed falling into one of the two categories: gene prediction and gene prioritisation. A gene prediction tool determines one or more candidate genes out of a pool of genes while a gene prioritisation tool ranks the genes contained in the pool. Genes are ranked according to their relevance based on one or more criteria. Gene prediction can be compared with a labelling mechanism simply labelling the involvement (or not) of genes in a disease. In contrast, gene prioritisation algorithms assign a measure or a rank to each gene. Based on this measure/rank, the involvement of a gene in a disease is determined.

Tranchevent et al. (2010) recently reviewed 19 available online tools for gene prioritisation. The results of their assessment were published in the Gene Prioritization Portal [13]. The Gene Prioritization Portal has been updated since the publication and contains 33 tools as of 2 August 2012. One aspect of the study was to catalogue the different types of data underlying the prioritisation tools. The underlying data range from gene expression, over chemical and functional annotations to phenotypes, also covering other kinds of data.

Despite the growing number of algorithms being developed, no generalised procedures have been established to compare all available tools. However, Tranchevent et al. (2010) work on a standardised *critical assessment* which is, according to the website, available soon. Two aspects relevant to automated disease gene discovery and also to the work presented in chapter 3 and 4, are illustrated in the following subsections.

1.3.1 *Guilt-by association approaches*

Guilt-by association approaches rely on established links between a human disease and genes, which are causative for this disease. Those known gene–disease associations are used to *profile* the disease based on either annotations or other characteristics of the associated genes. Examples include determining what molecular functions and biological processes are impaired (Schlicker and Albrecht, 2010; Schlicker et al., 2010) or which anatomical components are affected (Tiffin et al.,

13 http://homes.esat.kuleuven.be/~bioiuser/gpp/stats.php

2005). Even though those approaches could help with the identification of novel disease gene candidates, *guilt-by association* approaches possess drawbacks:

- not applicable to diseases where no genetic cause is known yet
- potential biases in the annotations of genes associated to the disease

From over 7,000 diseases characterised in OMIM, only about half the diseases possess one or more identified causative genes (Schofield et al., 2012). Consequently, *guilt-by association* approaches can only be applied to about half of the to date known genetic diseases.

By using known gene–disease associations, the profiles are generated from characteristics of the associated genes. Therefore, the associated genes are assumed to be representative for this disease. This assumption poses limitations when genes are either not sufficiently annotated or extensively studied. Extensively studied genes possess a high number of annotations which are not necessarily relevant to all the diseases the gene is associated with. Genes may be also multifunctional and change function depending on the context (Gillis and Pavlidis, 2011). Thus, not all the annotations of a multifunctional gene may be relevant for each of the diseases it is associated with.

1.3.2 *Semantic similarity*

Semantic similarity is a measure of likeness between two entities according to the semantics of those entities. The semantics of entities can be described through e.g. ontology annotations or terms. For instance, in a biological context, semantic similarity measures have been applied to measure the similarity of gene function based on their representation with GO.

To date, different measures have been proposed to calculate semantic similarity. Those measures can be separated (among other criteria) into scores calculated between two ontology concepts (pairwise) and scores calculated between sets of ontological concepts (group-wise) (Xu et al., 2008). Group-wise similarity scores can be calculated by applying pairwise similarity measures as an intermediate. This means that to determine the semantic similarity of two sets of concepts, all possible combinations of concepts in those sets are built and their pairwise similarity is calculated. A final score is obtained by e.g. either

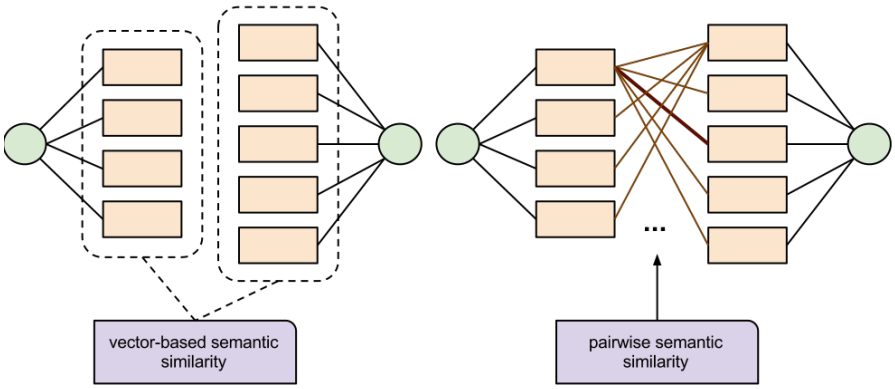


Figure 2: **Illustrates the difference between a vector-based and a pairwise similarity score to determine the semantic similarity of a set.** Circles represent objects described with ontological concepts (squares). Lines (not all possibilities included here) between squares constitute a pairwise similarity score while dashed rectangles correspond to vectors of ontological concepts.

maximising (Brameier and Wiuf, 2007) or averaging (Lord et al., 2003; Wang et al., 2005) the similarity scores for each pair.

The semantic similarity of two sets of ontology concepts can also be calculated based on vector-based algorithms, where each vector possesses one of the set's ontology concepts as elements. Figure 2 illustrates the calculation of a semantic similarity of two sets of concepts: vector-based on the left and by employing a pair-wise scoring as intermediate on the right.

Further to the distinction of whether the semantic measure is determined pairwise or group-wise, semantic similarity measures can be categorised as to whether or not using an information content (Pesquita et al., 2009). In case of algorithms applying the information content of concepts, the information content for each concept has to be calculated based on a representative reference resource, before semantic similarity measures can be determined. A high information content score indicates a low occurrences of the ontology concept in the reference resources. A low information content indicates a high occurrence of the ontology concept in the reference resource.

One example of such an approach is the similarity measure proposed by Resnik (1995), which has been evaluated in different studies. Even though this approach performed well or even best in those studies, it was also noted that information content scores may not necessarily be representative. Differences may occur depending on

the resource used for calculating the information content (Mistry and Pavlidis, 2008; Pesquita et al., 2009).

Semantic similarity scores in the context of GO have been applied in different ways, either solely based on GO (Wu et al., 2006) or jointly with other data, e.g. to construct a genotype network (Guo et al., 2006). The performance of the similarity score may differ according to the task it is applied to. Therefore, Pesquita et al. (2009) proposed guidelines as to which similarity measure is best to choose in the context of GO. The authors also suggested that if the similarity measure is not necessarily the best suitable, the results have to be interpreted carefully.

1.3.2.1 *Phenotype semantic similarity measures*

Phenotype semantic similarity measures are used to calculate the relatedness of entities based on their phenotype descriptions. For instance, phenotype similarity of genes is measured to identify members in pathways or disease causing genes (Washington et al., 2009). An obstacle to the application of phenotype similarity measures is that phenotypes are not represented consistently (see 2), neither in a single species nor across species (Gkoutos et al., 2012). Even though a small number of studies exists applying semantic similarity measures of phenotype information, the ontologies, terminologies and applied semantic similarity measures differ. Some examples are given here for illustration purposes.

Cantor and Lussier (2004) used OMIM's phenotype descriptions in the *Clinical Synopsis* part of an OMIM entry. Those phenotype descriptions were used to build binary vectors for each disease. The generated binary vectors were then used to hierarchically cluster 4,491 diseases. By evaluating their results, the authors found that for the investigated subset, diseases can be clustered using OMIM's phenotype descriptions and self-organising maps.

van Driel et al. (2006) also investigated the textual phenotype descriptions of OMIM diseases. The authors used TM to extract phenotype descriptions for each disease. Phenotype descriptions were based on the anatomy and disease branch of the Medical Subject Heading (MeSH) vocabulary. The terms inside the vectors were weighted according to their occurrence in OMIM entries and corrected for record length. A vector-based semantic similarity measure, cosine similarity, was applied to determine the phenotype similarity of the diseases. The obtained similarity measures were then used to build disease clusters and clusters were further investigated. By doing so, van Driel et al.

(2006) were able to connect more than 5,000 OMIM diseases and could demonstrate a positive correlation with gene sequence, protein motifs, functional annotations and known protein interactions inside their clusters.

Lage et al. (2007) used phenotype similarity to prioritise proteins whose producing genes are within a given linkage interval. Phenotype similarity was determined using the Unified Medical Language System (UMLS) (Bodenreider, 2004) and calculating the cosine similarity. Protein phenotypes were obtained from Ensembl (Spudich and Fernández-Suárez, 2011; Flicek et al., 2012). Disease phenotype annotations were obtained by applying MetaMap (Aronson and Lang, 2010) to the textual descriptions of the diseases in OMIM. Furthermore, a Bayesian network was trained to assign posterior probabilities to a protein involved in a disease based on the phenotypes. Evaluation of the system showed that using the system for protein prioritisation outperformed other methods at the time of publication.

Washington et al. (2009) used phenotype similarity to identify allelic variants of a gene, potential pathway members, orthologs in zebrafish and mouse, and orthologous pathway members. Phenotypes are represented using the entity–quality (EQ) formalism (Mungall et al., 2010) (see 2.2.2.2) and four different similarity scores are used: three involving an information content score and a Jaccard index. The information content scores for EQ statements were calculated based on the annotations of 11 human disease genes. The authors concluded from their results that EQ statements and the applied similarity measures can be used to link human disease and animal models.

Zhang et al. (2012) assessed the performance of a subset of GO similarity scores (Mistry and Pavlidis, 2008; Pesquita et al., 2009) applied to genes described with Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010) annotations. The authors showed that the chosen measures could be used to identify genes with similar phenotypes. Furthermore, the five applied similarity measures outperform the van Driel et al. (2006) method of determining phenotype similarity. Interestingly, Resnik's similarity score employing an information content based on the assigned annotations (Resnik, 1995) was outperformed on the OMIM data set. Zhang et al. (2012) concluded that better phenotype similarity measures are still be identified.

The list of examples provided here is not comprehensive and only serves the purpose of illustrating current attempts to use phenotype semantic similarity scores. Furthermore, the examples show that no

common approach exists to determine the semantic similarity of phenotype descriptions.

1.4 ENRICHMENT FROM SCIENTIFIC LITERATURE

Tools such as PhenomeNET and MouseFinder (Hoehndorf et al., 2011c; Chen et al., 2012) employ phenotype descriptions of human genetic diseases described in OMIM. Investigating the OMIM disease phenotypes, among other human phenotype data, Oti et al. (2009) showed that a broader coverage of phenotype data is required. Consequently, an enrichment of the disease phenotypes in OMIM could lead to a better performance of PhenomeNET and MouseFinder.

With the ever growing amount of publications, the scientific literature constitutes a good repository for biological discovery (Jensen et al., 2006). Korbel et al. (2005) showed that phenotypes extracted from literature can be successfully applied to annotate genomes. Those results suggest that systematically extracting phenotypes from scientific literature may be used to enrich existing resources.

This section aims to first give an overview about the different areas of TM (see 1.4.1), then provide an insight to evaluation measures (see 1.4.2) and finally summarise the efforts to text-mine phenotype data (see 1.4.3).

1.4.1 *Research areas in text mining*

The research domain of TM can largely be divided into three different areas:

1. information retrieval
2. information extraction and
3. knowledge discovery

Those three areas are illustrated in figure 3 and are briefly explained in the following.

1.4.1.1 *Information Retrieval*

Information retrieval identifies relevant documents or text spans from a defined text corpus according to a user defined query, e.g. all papers mentioning Septo-Optic Dysplasia (SOD) in their title. Diverse

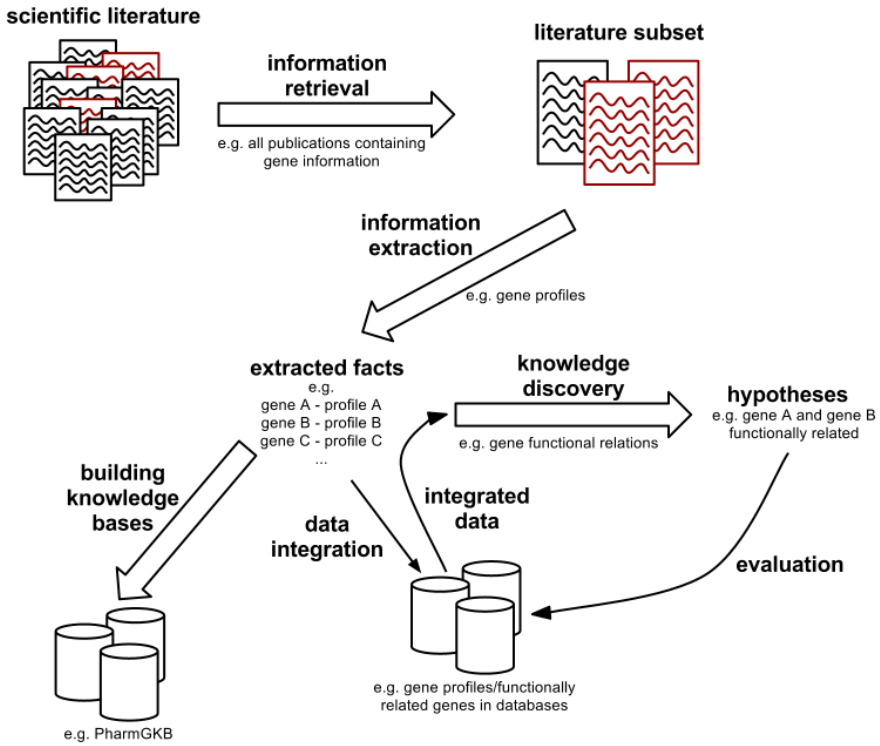


Figure 3: **Illustrates the three areas of text mining: information retrieval, information extraction and knowledge discovery.** After retrieving documents, information extraction algorithms can extract assertions which constitute, upon verification e.g. with existing resources, scientific facts. Either assertions or facts can be used to populate databases or to derive hypotheses.

algorithms are applied to identify the relevant documents, including simple keyword or complex query expansion searches. In the case of a simple keyword search, all documents mention one or more of the query-defined keywords are retrieved. In the case of query expansion, the user-defined query is expanded, e.g. via ontologies or an expert lexicon (see *BioLexicon* (Thompson et al., 2011)).

Systems performing information retrieval, so called retrieval engines, return results in different ways. Depending on the query, results may be numerous in which case a user would have to go through a large list to identify the relevant documents and text spans. To ease the user's search, algorithms are applied to make those list more manageable and "guide" the user through the results. For example, PubMed (Sayers et al., 2012) simply returns a list of the results but provides manually configurable filters to further narrow down the list. Another retrieval engine, GoPubMed (Doms and Schroeder, 2005), provides a classification of the results according to GO concepts. Users can then go through the results by browsing the ontology's hierarchy.

1.4.1.2 *Information extraction*

Information extraction focuses on the identification of scientific facts contained inside the literature. Facts in this context can either be entities, e.g. diseases, or relationships between those entities, e.g. gene–disease associations. With applying information extraction methods, the content of the papers is processed and only the relevant text spans are provided to the user. The applied abstractions provide faster access to the content of a paper.

To identify entities in a text, named entity recognition (NER) systems are applied. NER systems apply either dictionaries or complex machine learning algorithms to the identification of entities. For instance, BANNER (Leaman and Gonzalez, 2008) is a system which provides extracted genes and proteins contained in free text. In most cases, NER is a two-step process: the first step is to identify text spans that possibly contain entities, and the second step is to *normalise* the text spans to e.g. ontology concepts or database entries.

Relationships between the identified entities can be derived by either simple methods, such as co-occurrence of entities, or in complex ways, e.g. defining textual patterns which hold between the entities. Co-occurrence means that two entities appear "close" (based on either sentence, number of words, paragraph) to each other. For instance, Percha et al. (2012) applied patterns to identify drug–target relationships in scientific literature. Both entities and relations holding between those entities, extracted from literature are only assertions. Once they have been evaluated and verified, they constitute scientific facts.

1.4.1.3 *Knowledge discovery*

Further to retrieving documents and extracting facts from scientific literature, TM also includes aspects of knowledge discovery. Knowledge discovery derives testable hypotheses from the assertions and facts obtained through information extraction. Hypotheses can be derived by either accumulation of the extracted results or through integration with other resources (see figure 3). By applying TM and following the assumption that *if* $A \Rightarrow B$ *and* $B \Rightarrow C$ *then* $A \Rightarrow C$, Swanson famously showed the relevance of dietary fish-oil to Raynaud's syndrome and the impact of magnesium on migraine (Swanson, 1990). Hypotheses can be derived by applying textual, statistical and formal methods. Derived hypotheses can then be used to guide further experiments as implemented in the Robot Scientist "Adam" (King et al., 2009).

1.4.2 *Evaluation of Text mining solutions*

Due to the amount of data processed (e.g. approximately 20 million citations in MEDLINE), text mining solutions produce an enormous amount of results which are impossible to assess manually. Consequently, a potential evaluation of the results is either manually possible only on a subset or has to be automated to assess all results. To date, several evaluation procedures for TM tools have evolved, which can be roughly grouped into experimental, manual and (semi-)automated evaluation methods. While experimental evaluation is the most desired and most reliable option, it is not always applicable due to cost and time constrains. Manual evaluation requires experts to judge on the results while automated evaluation needs a test set to compare against. The latter two options are the most applied in the field of TM.

As part of my work, I aimed to develop a named entity recognition (NER) system for phenotypes in scientific literature (see 5) and one possibility to determine its performance is to calculate its precision, recall and the Harmonised F-measures. Thus, I will explain those measures in a little more detail.

Another option of evaluation, and which is not only applied in the context of TM tools, are receiver operating characteristic (ROC) curves. In fact, ROC curves are used in the studies presented in this thesis to determine the performance of a gene prioritisation pipeline. I will still introduce them here as one possible way of evaluation (see 1.4.2.2).

1.4.2.1 *Precision, Recall and F-measure*

Precision, recall and the Harmonised F-measure are broadly used measures to determine the performance of a TM solution. They are calculated based on a reference set or corpus (illustrated in figure 4).

An example of a text corpus is the manually annotated GENIA corpus (Kim et al., 2003; Tateisi et al., 2005; Kim et al., 2008). GENIA was derived from 2,000 Medline abstracts (>400,000 words, approximately 100,000 annotations), containing gene mentions, syntactical information, and events. GENIA contains abstracts relevant to transcription factors in human blood cells. 500 abstracts have syntactical annotations based on the Penn Treebank II scheme (Bies et al., 1995). Furthermore, 36,114 event annotations have been assigned to half of the corpus (1,000 abstracts with 9,372 sentences). All the relations contained in the corpus have also been formalised by Hoehndorf et al. (2011b).

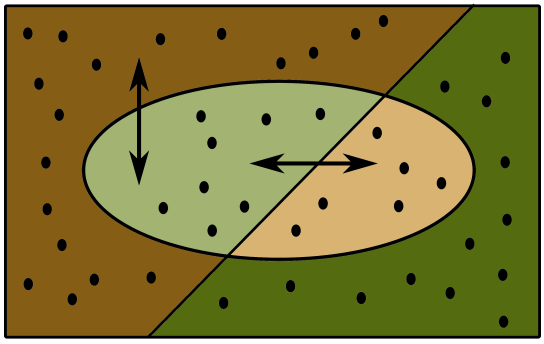


Figure 4: **Illustration of precision and recall.** Rectancle: entire result space (e.g. documents/annotations). Ellipse: results of application of a method (light green and brown). Diagonal: application of method divides result space into correct (left from diagonal) and incorrect (right from diagonal) results. Subset correctly identified: light green part of ellipse; subset incorrectly identified: light brown part of ellipse. Subset incorrectly not retrieved by method: dark brown part of rectancle; subset correctly not retrieved by method: dark green part of rectangle. Precision symbolised by horizontal arrow while recall by vertical arrow.

Precision and recall (and consequently F-measure) are calculated based on the annotations assigned to a corpus. Precision is the ratio of correctly assigned annotations over all assigned annotations of a method:

$$\begin{aligned} \text{precision} &= \frac{\text{correctly assigned annotations}}{\text{all assigned annotations}} \\ &= \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \end{aligned} \tag{1.1}$$

Recall is the ratio of annotations assigned by a method to all annotations contained in the corpus:

$$\begin{aligned} \text{recall} &= \frac{\text{correctly assigned annotations}}{\text{all text corpus annotations}} \\ &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \end{aligned} \tag{1.2}$$

In most cases, precision and recall deviate from each other and in some cases, the gap between them is large. Trying to improve one, mostly results in the decrease of the other. To provide a means of comparison for several solutions differing in precision/recall, commonly

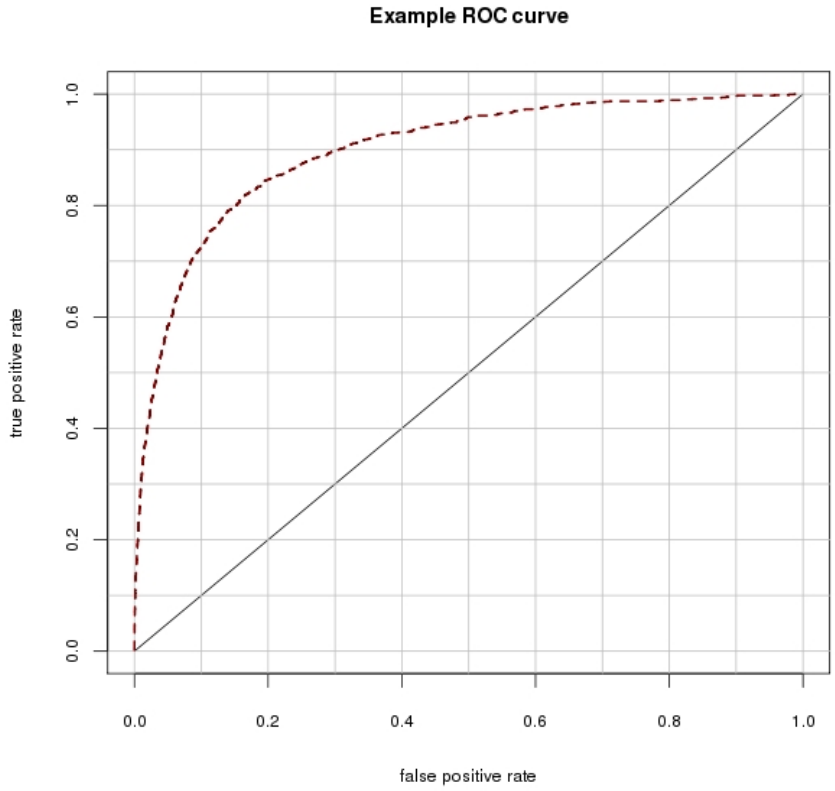


Figure 5: **Illustration of a ROC curve.**

the Harmonised F-Measure is calculated which is the harmonic mean of precision and recall:

$$F - harmonised = \frac{2 \times precision \times recall}{precision + recall} \tag{1.3}$$

1.4.2.2 *Receiver Operating Characteristic curves*

An alternative way of assessing the performance of a system, are receiver operating characteristic (ROC) curves. A ROC curve is a plot of the true positive rate as a function of the false positive rate. The area under curve (AUC) is a quantitative measure and is equivalent to the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative one. An illustration of a ROC curve is provided in figure 5.

To compare two or more different solutions with each other, the ROC curve for each solution has to be determined under the same

conditions. The differences of the resulting ROC curves can than be assessed with existing tools[14].

1.4.3 *Phenotype information extraction from scientific literature*

In the past years, genes and their products have been in the centre of attention of the TM community. Little effort has been spent to extract diseases or phenotypes from the scientific literature (Shah et al., 2009; Jimeno et al., 2008). Consequently also no normalisation of diseases and phenotypes to ontological concepts has been tackled. However, some studies used text-mined phenotype data and showed the potential lying therein.

Korbel et al. (2005) assumed that nouns with more than three characters co-occurring with 92 species names in MEDLINE abstracts are phenotype descriptors. For all the 92 species, orthologous gene groups were obtained and gene–phenotype associations have been identified based on an over-representation in literature. 2,700 gene–disease associations could be identified as plausible from a total of 1.9×10^9 possible relationships.

Cohen et al. (2011) proposed an advanced search functionality for OMIM by mining phenotype descriptions from the textual descriptions using MetaMap. The suggested algorithm also takes the context of the phenotype information and negation into consideration.

Despite systems focusing solely on the extraction of phenotype information, generalised concept recognition systems exist. Those generalised concept recogniser assign a normalised (to either ontologies or thesauruses) annotation to a text span. Examples include MetaMap (Aronson and Lang, 2010), MGrep (Dai et al., 2008) and Open Biomedical Annotator (OBA) (Jonquet et al., 2009), which are further explained in the following.

MetaMap is one of the first concept recogniser and is widely regarded as the gold standard for disease name recognition (Shah et al., 2009). MetaMap is tied to the UMLS and hence only recognises terms and ontology concepts included in UMLS. MetaMap provides a variety of features and is updated according to the updates of UMLS. Submitting free text to the system will result in a list of phrases together with the UMLS concept unique identifier and a confidence score. Each concept unique identifier is linked to all the vocabularies, terminologies or ontologies (all human-specific) possessing this term or concept. Even

14 e.g. `http://vassarstats.net/roc_comp.html` (Hanley and McNeil, 1982)

though MetaMap obtained in some studies precision measures in the range of 55% to 76% and recall measures in the range of 72% to 93% (Meystre and Haug, 2005; Pratt and Yetisgen-Yildiz, 2003; Chapman et al., 2004), Shah et al. (2009) measured a precision of only 9.1% when applying MetaMap to the recognition of disease names in MEDLINE abstracts.

Another generalised concept recognition system is MGrep. MGrep utilises a Radix-tree algorithm, incorporating labels/synonyms of ontological concepts as well as all word-wise permutations of the labels/synonyms. Labels/synonyms are restricted to a length of five words, but lexical variants of words contained in the labels/synonyms are supported. MGrep was assessed in initial studies determining its performance against MetaMap on four different reference sources (Bhatia et al., 2009; Shah et al., 2009). Those studies showed that MGrep outperforms MetaMap when applied to disease name recognition measuring precision. MGrep can be utilised in conjunction with various terminologies and ontologies.

The Open Biomedical Annotator (OBA) (Jonquet et al., 2009), e.g., has been developed to recognise all ontology concepts covered by the ontologies contained in BioPortal (Noy et al., 2009; Whetzel et al., 2011). BioPortal contains all OBO Foundry ontologies[15] (Smith et al., 2007) (see 1.2.1). Consequently, OBA also extracts Mammalian Phenotype Ontology (MP) (Smith et al., 2005) and Human Phenotype Ontology (HPO) (Robinson and Mundlos, 2010) annotations. These two ontologies are used to standardise phenotype descriptions (see 2). OBA uses MGrep for annotation extraction from text.

Another concept recogniser is LingPipe (Carpenter, 2007) which has been partially used in the studies presented here and is further explained in chapter 5. Note that the extraction of phenotypes and diseases are considered to be two independent though related tasks. Some studies use both terms interchangeably while they are considered to be two distinguishable and different entities in this thesis.

1.5 AIMS OF WORK AND OUTLINE OF THE REMAINING CHAPTERS

My research work focused on the identification of gene candidates for human heritable diseases through integration and analysis of a variety of resources containing genotype as well as phenotype data. In particular, the automated comparison of phenotype representations

15 http://obofoundry.org/

across species, especially mouse and human, was at the center of my work. My investigations were geared towards the development of a phenotype mining solution (called *InterPhen*) that would derive disease gene candidates through integration and analysis of phenotype data. However, the collection of resources would not only use databases derived from biological experiments but also enable the integration of scientific literature. The data content would not only span one particular species, instead, it would also take model organisms into consideration. The underlying assumptions are that orthologous genes exhibit the same phenotype as well as that the modifications of orthologous genes yield the same changes in phenotypes. In spite the fact that multiple aspects of diseases and genes could be used for data integration, e.g. anatomical structures or pathway involvement, my goal was to solely use phenotype data.

In addition to the prioritisation of disease genes, *InterPhen* could be used as a curation system at the same time. Results of biological experiments can be manually added and compared to the wealth of information retrieved from the integrated databases and the scientific literature. After confirming the experimental results, these could instantly be added to the analysis and possible changes in the prioritisation of genes could be observed. Maintaining such a structure, would allow evidence codes to be added to show where phenotypes were obtained and whether and how this affects the prioritisation.

To start with a manageable subset of data to be integrated and analysed, I chose MGD, OMIM and the scientific literature for my integration and prioritisation efforts. To enable the prioritisation of disease genes based on the integrated phenotype data, the following questions needed to be answered:

1. How are phenotypes semantically represented in existing resources?
2. What are advantages and disadvantages of these representations?
3. How do these phenotype representations inter-relate?
4. How can phenotypes be semantically integrated across species to allow for an improved gene prioritisation?
5. Can the existing data repositories be enriched?

My work was split into individual projects to create manageable sub-units of the overall work load, with the intention to combine all

individual parts once satisfying results have been achieved. While chapters 2 to 6 describe the achievements in the individual projects, chapter 7 provides an assessment with respect to the development of *InterPhen*. The content of each of the chapters is outlined here briefly.

Chapter 2

In chapter 2, I explore different existing phenotype resources as to how they represent phenotype data. Moreover, the advantages and disadvantages of the different ways of representing phenotypes are assessed. In addition, this chapter provides more details for both databases intensely used throughout this thesis: the mouse-specific MGD and the human-centric OMIM database.

Chapter 3

After assessing the different ways of representing phenotype data, chapter 3 focuses on the semantic integration of phenotypes across species and repositories, and on the application of the integrated data to prioritise disease genes. In a nutshell, mouse models that have been investigated in mutagenesis projects and described on a phenotype level have been compared to the signs and symptoms used for the description of human diseases. Mouse models exhibiting a high phenotype similarity with a disease constitute potential models for this disease. Hence, mouse models may give insights to the molecular basis of the disease. The suggested approach has been shown to outperform PhenomeNET and is especially valuable for diseases with unknown genetic cause, where "guilt-by association" approaches cannot be applied (see 1.3.1).

Chapter 4

The integration and prioritisation work described in the previous chapter relies on the availability of phenotype data. One bottleneck is the availability of comprehensive descriptions of signs and symptoms of genetically caused diseases in human. My 4th chapter introduces a data mining method to derive mouse-specific disease descriptions based on phenotype annotations assigned to mouse models. The derived disease descriptions are checked for their validity by applying them in a biological use case, corresponding to the gene-prioritisation task described in the previous chapter. Furthermore, examples are

provided that demonstrate the potential of the mouse-specific disease profiles to enrich the existing disease descriptions.

Chapter 5

In addition to comprehensive signs and symptoms of diseases, the work described in chapter 3 relies on phenotype data gathered from mice mutagenesis experiments reported in MGD. The data contained in MGD is manually curated from scientific literature and the process of manual curation is slow. In this chapter (chapter 5), I suggest two methods for phenotype extraction from scientific literature. Both methods aim to annotate mouse models with phenotype annotations and are evaluated on a gold standard corpus derived from MGD. As the evaluation of the results shows, both suggested method require improvements in order to reach the performance of existing generalised ontology concept recognition systems. Further to the methods' descriptions and evaluation results, this chapter also outlines possibles ways of improving the extraction of mouse phenotype data. The methods need significant improvement before an integration into the *InterPhen* system is possible.

Chapter 6

The integration and prioritisation method illustrated in chapter 3 does not only rely on the availability of phenotype data of mice and humans, it also requires an alignment of phenotypes between both the species. In chapter 6, I will demonstrate how text mining can be applied to the integration of phenotype data. Text mining is used to decompose phenotypes to achieve integration based on the decomposition.

Chapter 7

Chapter 7 summarises the main achievements of the individual projects, and how the achievements in the individual projects impact the development of the outlined system *InterPhen*. The chapter further reports about requirements that have to be met before *InterPhen* will be fully functional to support disease gene prioritisation based on integrated phenotype data. Finally, the chapter provides an outlook to the future of *InterPhen* and to other use cases in which phenotype may play a vital role for knowledge discovery.

1.5.1 *Additional guidance information*

Even though each chapter is designed to be an individual, self-existent part, it would be beneficial for the reader to first read through chapter 5 and then move to chapter 6. The topics covered in both chapters are closely related as one of the evaluation scenarios used in chapter 5 is the aim of the work presented in chapter 6.

Throughout my thesis, I will reference particular genes, diseases and phenotype information with showing the entry number of the corresponding database entry or an ontology concept. To reference to mouse specific markers, I will use the marker accession numbers from MGD, e.g. *MGI:95689* to reference *Gdf6*. For human genes however, I will use the Gene Product ID from UniProt, e.g. *Uniprot:Q9UBX0* for the human gene *HESX1*. For human diseases, I will provide the number of the corresponding OMIM entry, e.g. *MIM:#105830* for *Angelman syndrome*. Human phenotypes are represented with HPO concepts, e.g. *HP:0007370* for *Aplasia/Hypoplasia of the corpus callosum* while mouse phenotypes are represented with MP, e.g. *MP:0002196* for *absent corpus callosum*.

A detailed list about the contents of this thesis which have been or are submitted for publication is provided in appendix D and contributions are detailed on page xi.

2

THE PLETHORA OF PHENOTYPE DESCRIPTIONS

As outlined in section 1.5, the aim of my work was to integrate diverse phenotype data repositories and use the integrated data for data mining purposes, more specifically to mine gene–disease associations. Before exploring possible options to integrate and analyse phenotype data in different repositories, I assessed how phenotypes are described in existing phenotype data repositories. The result of this assessment was a categorisation, which is presented in this chapter. Each identified category of phenotype descriptions is further illustrated with an example.

In addition to the categorisation of available phenotype descriptions, this chapter features an overview of two model organism database (MOD)s possessing phenotype content (see 2.4). These two databases are introduced here as they are extensively used throughout the work presented in this thesis.

2.1 BACKGROUND

Since the discovery of the causal relationship of a genotype in a certain environment and a phenotype, a need evolved to systematically record genotypes with their respective phenotypes. Multiple resources, also referred to as model organism database (MOD)s (see Leonelli and Ankeny (2012)) for examples) evolved to gather this type of information in diverse species. For example, phenotype efforts such as the International Mouse Phenotyping Consortium (IMPC) (Collins et al., 2007), focus on the discovery of a species' phenome. Underlying those efforts are enormous mutagenesis projects (Abbott, 2010), generating large amounts of biological data. Genotype–phenotype associations are mostly stored in databases, e.g. the Sanger mouse database[1] or the Mouse Genome Informatics database (MGD) (Blake et al., 2011).

Even though the individual MODs aim to consistently represent phenotype content within their resource, little attention has been paid on how other resources represent phenotype data. As a consequence, a variety of phenotype descriptions evolved that do not necessarily

1 http://www.sanger.ac.uk/mouseportal/

facilitate the integration of phenotype data across resources. This lack of integration isolates knowledge inside databases and prevents cross-species translational research. Only once we are able to integrate these databases, will we able to unlock the full potential of these resources for translational research. To achieve this integration, an understanding of the diverse ways of describing phenotypes and how they relate to each other is required.

A table of existing phenotype resources, which is not comprehensive and mainly covers mouse and man, is provided in appendix A. A modified version of this chapter has been published as a contribution to the 2nd Ontologies in Biomedicine and Life Sciences (OBML) workshop in 2010 (Oellrich and Rebholz-Schuhman, 2010). For more details refer to page xi.

2.2 EXISTING CATEGORIES OF PHENOTYPE DESCRIPTIONS

As defined in section 1.1.1, a phenotype is a measurable or observable characteristic of an organism. The collection of all phenotypes in a species is called the phenome of that species. Phenotypes span over different levels and different resolutions, including molecular mechanisms and even the environment of an organism.

Investigating databases such as the MGD, the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2011) and OrphaNet (Aymé and Schmidtke, 2007), several broader groups of phenotype descriptions were identified (also illustrated in figure 6):

1. narratives
    a) free text (see 2.2.1.1)
    b) database specific vocabularies/terminologies (see 2.2.1.2)
2. ontologies
    a) pre-composed phenotype ontologies (see 2.2.2.1)
    b) post-composed phenotype ontologies (see 2.2.2.2)
        i. qualitative descriptions (Mungall et al., 2010) (see 2.2.2.2)
        ii. phenes (Hoehndorf et al., 2010b) (see 2.2.2.2)

The identified groups will be further explained in the following subsections. Each of the groups is illustrated with selected examples (also summarised in figure 6).

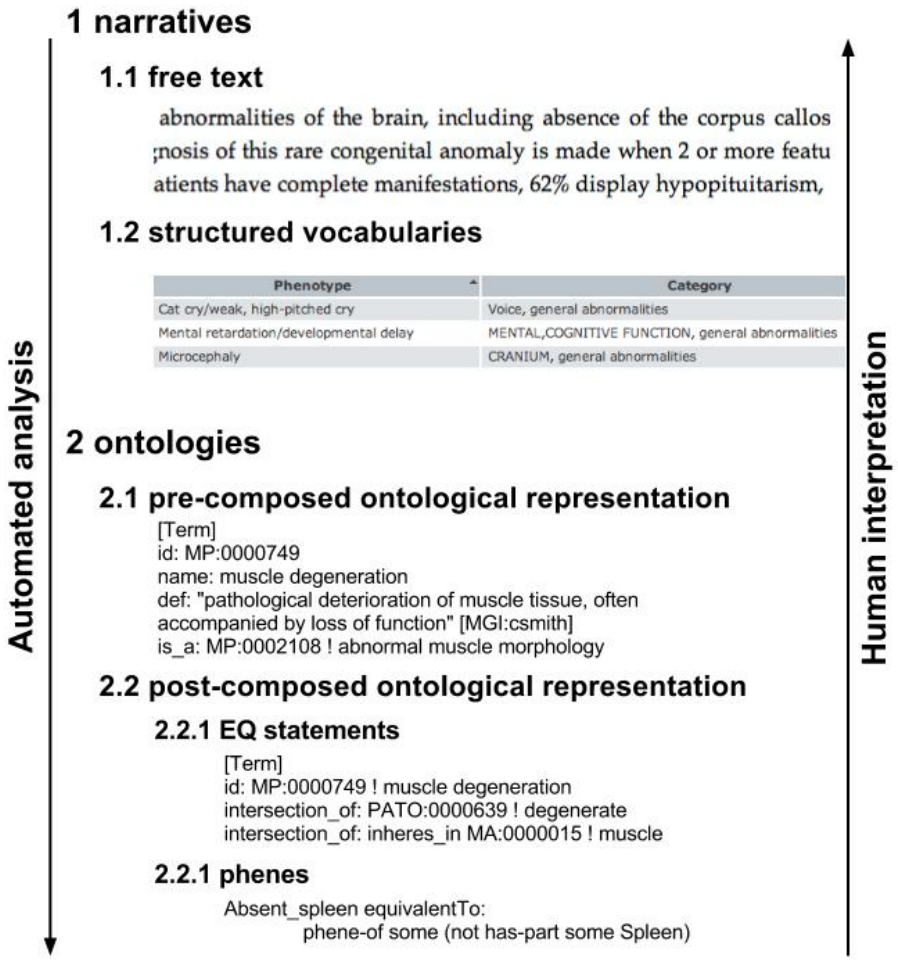


Figure 6: **Phenotypes are described using narratives and ontologies.** Human ease of interpretation decreases from top to bottom while machine ease of interpretation increases. Free text illustration is screen shot of OMIM's textual description, entry: Septo-Optic Dysplasia (SOD) (MIM:#182230). Illustration for structured vocabulary taken from DECIPHER's (Firth et al., 2009) *Cri du Chat Syndrome (5p deletion)*, see `https://decipher.sanger.ac.uk/syndrome/2`.

2.2.1 *Narrative phenotype description*

Narrative phenotype descriptions include both free text descriptions and predefined terminologies or vocabularies. Both types of narrative phenotype descriptions are explained individually and in more detail in the following subsections.

2.2.1.1 *Free text*

In a free text phenotype description, all information is provided as a narrative. For example, OMIM possesses for all its diseases a textual description of the prevalent phenotypes in affected individuals. Free text information has the advantages that the author of the description has full freedom about the words used to describe a phenotype. Moreover, the author does not have to learn about an existing vocabulary or ontology before providing descriptions. Consequently, the degree of detail can freely be chosen by the author as there is no limitation due to the availability of terms or concepts. Free text descriptions are specific to the person having written them – like a fingerprint – due to the use of certain words or combination of words.

Phenotypes described with free text constrain a potential integration of phenotypes due to possible ambiguity in descriptions and the differences in domain specific jargon (Groth et al., 2011). For that matter, integration is not only challenged when including different repositories possessing free text descriptions, e.g. OMIM and OrphaNet. Even the analysis of different entries of one resource is difficult, especially when different authors were involved in writing the free text descriptions.

Possibilities to automatically access the knowledge captured in text arise from the recent progresses in the area of text mining (TM). TM has advanced to identify and extract information of interest, but all the existing solutions are still not capable of achieving the same performance as human reasoning and interpretation.

2.2.1.2 *Structured vocabularies/terminologies*

Structured vocabularies or terminologies are agreed upon phenotype descriptions. They are limited to a selected subset of words and certain combinations of those words. Structured vocabularies are mostly used in databases, e.g. the London Dysmorphology Database (LDDB) (Winter and Baraitser, 1987; Guest et al., 1999) terminology, which was also used in the Database of Chromosomal Imbalance and Phenotype

in Humans using Ensembl Resources (DECIPHER) (Firth et al., 2009) to annotate patients and syndromes. However, in its newer version, DECIPHER moved from a structured vocabulary to an ontological representation (the pre-composed Human Phenotype Ontology (HPO) (Robinson et al., 2008), see 2.2.2.1) to represent both its patient and syndrome data (Corpas et al., 2012).

The chances of integration increase when phenotypes are represented using a structured vocabulary or terminology, since there is a limited subset of descriptions. However, this limitation is a disadvantage in as much as a curator is limited in choice and forced to learn the vocabulary. This may result in an non-ideal representation or even a loss of information, e.g. due to insufficient granularity of the applied terminology.

2.2.2 *Ontological resources for phenotype descriptions*

After the success of the Gene Ontology (GO) (Ashburner et al., 2000), standardisation efforts in the biomedical domain are ongoing (see 1.2), leading to the development of more and more ontologies. To handle the growing number of ontologies and increase their quality, communities such as the OBO foundry[2] (Smith et al., 2007) are evolving.

With the growing number of ontologies in the biological and biomedical domain, the number of ontologies covering phenotype information increases as well. Two major groups of ontological representations exist: pre- and post-composed phenotype ontologies, further described in the following sections.

2.2.2.1 *Pre-composed phenotype ontologies*

In a pre-composed phenotype representation, each phenotype is described as one concept in the phenotype ontology. Each concept of such an ontology can be readily used to, e.g. annotate experimental results stored in a database (see 1.2.2). Examples of pre-composed phenotype ontologies include the Mammalian Phenotype Ontology (MP) (Smith et al., 2005) and HPO. An example of a pre-composed phenotype concept is the MP concept *muscle degeneration* (MP:0000749) (represented in the OBO Flatfile format; see 1.2.1):

2 http://obofoundry.org/

```
[Term]
id: MP:0000749
name: muscle degeneration
def: "pathological deterioration of muscle tissue, often
      accompanied by loss of function" [MGI:csmith]
is_a: MP:0002108 ! abnormal muscle morphology
```

While pre-composed ontologies offer a good basis for data integration and analysis, they require a curator to know all available concepts and their meaning. Moreover, the annotator has to keep up with the development of the ontology as concepts may be added, amended or deleted over time. Furthermore, an ontology also limits the flexibility of the curator as the curator can only use concepts being defined in the applied ontology. Therefore, the applied ontology determines the level of granularity as well as the quality of the annotations.

Another problem arising with pre-composed phenotype ontologies is the limitation in either comprehensiveness or quality. If each possible phenotype is gathered in a pre-composed way, the resulting ontology becomes unmanageable. Due to the reduced manageability, the quality of the ontology decreases. If the aim is on a high quality, then a resulting pre-composed phenotype ontology will not be comprehensive due to the sheer amount of required phenotypes (even if restricted to only one species).

Even though a curator using a pre-composed phenotype ontology is limited by the choice of ontology, the possibilities to integrate knowledge across resources rise compared to the use of narratives to describe phenotype data. As long as phenotype repositories choose the same pre-composed phenotype ontology to annotate their content, results are comparable and an interchange of data between resources is facilitated. However, problems occur, when different pre-composed phenotype ontologies are used and no alignment (see 1.2.3) between those ontologies exist.

2.2.2.2 *Post-composed phenotype information*

Pre-composed phenotype ontologies can only be used for integration if data are annotated using the same pre-composed phenotype ontology or an alignment between those ontologies exist (see 1.2.3). Despite generalised alignment tools for ontologies (see 1.2.3), two alternatives evolved, particular to phenotypes. One possibility is the representation of phenotypes with entity–quality (EQ) statements (Mungall et al.,

2010). Thus, EQ statements take a qualitative approach to describe phenotypes. A second possibility is to extend from a pure qualitative description to distinguish differently formalised groups of phenotypes. Each group of phenotypes is formalised with a fixed description pattern (Hoehndorf et al., 2010b). However, both the approaches rely on species-independent bridging ontologies, such as GO.

Qualitative representation of phenotypes

Using EQ statements, the description of a phenotype is broken down into an *entity* and a *quality*. The *entity* is the affected component which is further described with the *quality*. Both *entities* and *qualities* are represented using concepts from other ontologies, e.g. Mouse adult gross Anatomy ontology (MA) (Hayamizu et al., 2005) or GO. While *entity* concepts are taken from a variety of ontologies, *quality* concepts are always taken from the Phenotype And Trait Ontology (PATO) (Gkoutos et al., 2009). Using the example from 2.2.2.1, the pre-composed phenotype concept *muscle degeneration* (MP:0000749) is decomposed into the *entity muscle* (MA:0000015) and the *quality degenerate* (PATO:0000639) (again represented in the OBO Flatfile format; see 1.2.1):

```
[Term]
id: MP:0000749 ! muscle degeneration
intersection_of: PATO:0000639 ! degenerate
intersection_of: inheres_in MA:0000015 ! muscle
```

EQ statement phenotypes allow for a better machine interpretation and integration of pre-composed phenotype ontologies. The process of building EQ statements to describe phenotypes certainly also allows for a greater flexibility of the curator than in the case of structured vocabularies or pre-composed phenotype ontologies. However, the process of developing and assigning EQ statements to any phenotype information requires the curator not only to know one particular ontology but rather all ontologies which can be used in an EQ statement. EQ statements still create limitations as compared to free text phenotype descriptions because the curator relies on concept availability in each of the applied ontologies.

Phenes

While the representation of phenotype descriptions with EQ statements takes a qualitative perspective of phenotypes, Hoehndorf et al. (2010b) extended the approach to a distinction of groups of phenotypes and

a formalisation for each group. EQ statements correspond to one group of phenotypes and are therefore a subset. Phenotypes – in this approach called *phenes* – are categorised into different groups, e.g. *structural phenes* or *process phenes*. For each group of phenes, fixed Web Ontology Language (OWL) patterns are suggested on how to represent those phenotypes. An example for a *structural phene* would be:

```
Absent_spleen equivalentTo:
              phene-of some (not has-part some Spleen)
```

Deriving EQ statements assumes that phenotypes can always be represented by *qualities* of *entities* and therefore only focuses on a strictly qualitative representation. With the phene OWL patterns suggested by Hoehndorf et al. (2010b), phenotype descriptions extend beyond a pure qualitative view. By applying the suggested OWL phenotype patterns, automated reasoning over body parts and processes is enabled.

This approach certainly possesses the highest complexity to describe phenotype information but due to its complexity enables the best representation to automatically process data and integrate resources. To use the method, a curator is required to know not only all the groups of phenotypes, but also the pattern for each group and the ontologies which are used inside those patterns, e.g. GO used for *process phenes*.

2.3 CONCLUSIONS

A variety of phenotype descriptions exist, each possessing advantages and disadvantages. While some of the descriptions are better suited towards automated integration and analysis, others are better suitable for human generation and interpretation. No overall standard exist that would allow easy transformation from one category of phenotype descriptions to another. Thus, alignment methods and guidelines for *common practice* are still required to achieve comprehensive and consistent data sets that can be easily integrated and facilitate automated analysis.

2.4 TWO SELECTED EXAMPLES OF PHENOTYPE DATABASES

To illustrate how phenotype data are embedded in databases, two community-approved databases are described here in more detail.

Both databases are extensively used in the studies presented in this thesis.

2.4.1 *Online Mendelian inheritance in man database*

The Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2011) database provides information about human genetic diseases. Besides describing relevant phenotypes for a disease (see 1.1.1), OMIM also contains entries for genes and links to their allelic variants. As of November 2010, OMIM contained 20,267 entries out of which 13,606 described genes and about 7,000 described genetic disorders (Amberger et al., 2011). Those numbers are statically increasing with the establishment of new syndromes, the discovery of new disease genes, and the identification of allelic variants of genes. The information contained in the database is curated from scientific literature (Thorisson et al., 2009).

The content of the database is available via a web interface[3] and in compressed text files. One such file is OMIM's MorbidMap (updated at regular time intervals) containing all gene–disease associations that have, to date, been manually evaluated. With the aim of computationally assessing large data sets, efforts are ongoing to assign phenotype annotations, amendable to computer-interpretation, for each OMIM record (Robinson et al., 2008; Köhler et al., 2009)[4].

While in the past the database had a focus on Mendelian diseases, it now also covers polygenic diseases (Hamosh et al., 2005).

2.4.2 *Mouse genome informatics database*

The Mouse Genome Informatics database (MGD) (Blake et al., 2011) provides an abundance of information concerning the model organism mouse, which is manually curated from scientific literature (Hancock and Mallon, 2007). Like OMIM, MGD also covers a great variety of information, ranging from a detailed description of the genotype over to the phenotype. Report files summarise subsets of the vast amount of data covered in the database. Those files are updated regularly and are provided for download[5].

3 `http://omim.org/`

4 download file (phenotype_annotation.omim) is available at `http://compbio.charite.de/svn/hpo/trunk/src/annotation/`

5 `ftp://ftp.informatics.jax.org/pub/reports/index.html`

In this database, phenotype information is represented using the pre-composed Mammalian Phenotype Ontology (MP) (Smith et al., 2005) and is assigned to mouse models. This assignment allows the connection of genotype and phenotype; in contrast, in OMIM (see 2.4.1) the emphasis is on the description of phenotypes in disease states.

3

USING MOUSE MODELS TO PRIORITISE GENETIC CAUSES OF HUMAN DISORDERS

The main focus of my work was the integration and analysis of phenotype data from different species with the aim to derive candidate genes for genetically caused human diseases (see 1.5). While the previous chapter described the different categories of existing phenotype descriptions, this chapter reports about my efforts to integrate two community-approved species-specific databases: the Mouse Genome Informatics database (MGD) (Blake et al., 2011) and the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2011). The integrated resources were then used to prioritise mouse genes for human disease.

3.1 BACKGROUND

One database reporting on human diseases is the OMIM database. This database holds, as of November 2010, entries for over 7,000 diseases (Amberger et al., 2011) (see 2.4.1). For only half of these diseases the underlying mechanisms are known, and further studies are required to elucidate the aetiology of the other half (Schofield et al., 2012).

A variety of automated gene prioritisation algorithms have been developed to help with the identification of disease causing genes (Tranchevent et al., 2010). The automated tools consider a range of features when prioritising disease genes (see 1.3). Since most of the available tools rely on known gene–disease associations and follow a "guilt-by association" approach (Tranchevent et al., 2010; Lee et al., 2011), they cannot be applied to the prioritisation of genes for diseases with yet unidentified molecular origins. However, information about phenotypes may be used to prioritise or predict candidate genes for diseases as well as functional relations between genes and proteins, even in the absence of knowledge about the molecular basis of a disease (van Driel et al., 2006).

Large-scale mutagenesis projects are ongoing, e.g. EuroPhenome (Morgan et al., 2009), aiming at the identification of a species phenome. With the progress of those projects, species-specific model organism

database (MOD)s evolve, gathering the phenotypic outcomes from the mutagenesis projects. Examples of MODs are the Mouse Genome Informatics database (MGD) (Blake et al., 2011) and the WormBase database (Yook et al., 2012). Due to species-specificity, those databases encapsulate knowledge which is hard to integrate (Smedley et al., 2008). However, only the integration of resources will enable us to get a comprehensive picture about the interplay of genotype and phenotype and allow us to derive knowledge about human diseases (Hoehndorf et al., 2011c; Chen et al., 2012; Washington et al., 2009) (see 1.1.3).

Mouse and human share a remarkable genetic overlap of 99%; therefore, mouse models are perceived as a reliable resource for investigations concerning human disease (Rosenthal and Brown, 2007). Due to the enormous amounts of data gathered in mouse and human, it is not feasible any longer to integrate the data manually. Automated methods are required which enable fast and reliable integration of species-specific descriptions and therefore facilitate cross-species investigations.

Here, we present a method to prioritise candidate genes in mice based on comparing experimentally derived phenotype data with phenotype descriptions of human diseases. We apply our method to the collection of phenotypes available from MGD (see 2.4.2) and compare those to the disease phenotypes available from the OMIM database (see 2.4.1). We evaluate our method by using a receiver operating characteristic (ROC) curve (see 1.4.2.2). Furthermore, we manually investigate a single disease together with its three highest ranked gene candidates.

The different parts of this chapter have been published in several ways: a journal publication in PLoS ONE (Oellrich et al., 2012), and an accepted publication in the Journal of Biomedical Semantics (extended contribution to the 3rd Ontologies in Biomedicine and Life Sciences (OBML) workshop 2011). For more details refer to page xi. Software and results are freely available online[1].

1 http://code.google.com/p/phenomeblast/wiki/CAMP

3.2 METHODS AND MATERIALS

3.2.1 *Overall work flow*

Before explaining the input data and each of the individual steps in more detail, a general overview of the approach is provided, which is also illustrated in figure 7.

Mouse models obtained from MGD (see 2.4.2) were used to prioritise genes for human diseases contained in OMIM. Before MGD's mouse models could be used for prioritisation, a unification step was required (see 3.2.4), assuring that both mouse models and human diseases would be represented with the same phenotype ontology. Prioritisation of each mouse model for each disease was executed based on their phenotype similarity (see 3.2.5). The obtained results were evaluated both automatically and manually (see 3.2.6).

3.2.2 *Input data*

In order to use models for gene prioritisation, we obtained three of the report files provided by MGD (all downloaded on 9 March 2011):

- `MGI_GenoDisease.rpt`,
- `MGI_GenePheno.rpt` and
- `HMD_Human5.rpt`.

The first report file contained gene–disease associations with 959 diseases, 1,135 genes and 2,088 model–disease associations. The second report file provided the information about 15,474 genotypes for 7,513 genes together with their observed phenotypes. The phenotypes are represented using 6,405 unique MP concepts. The third report file covers the information about human–mouse orthologous genes. In the version we downloaded, mappings for 17,828 orthologous genes were contained.

To incorporate the OMIM information into our study, we obtained the MorbidMap file on 1 March 2011, available via the database's download service. MorbidMap contains the information about known associations of human diseases and genes. The version we used contained 2,717 diseases that were linked to 2,266 genes with 3,463 distinct gene–disease associations (on average 1.27 genes per disease). The phenotypes associated with human diseases described

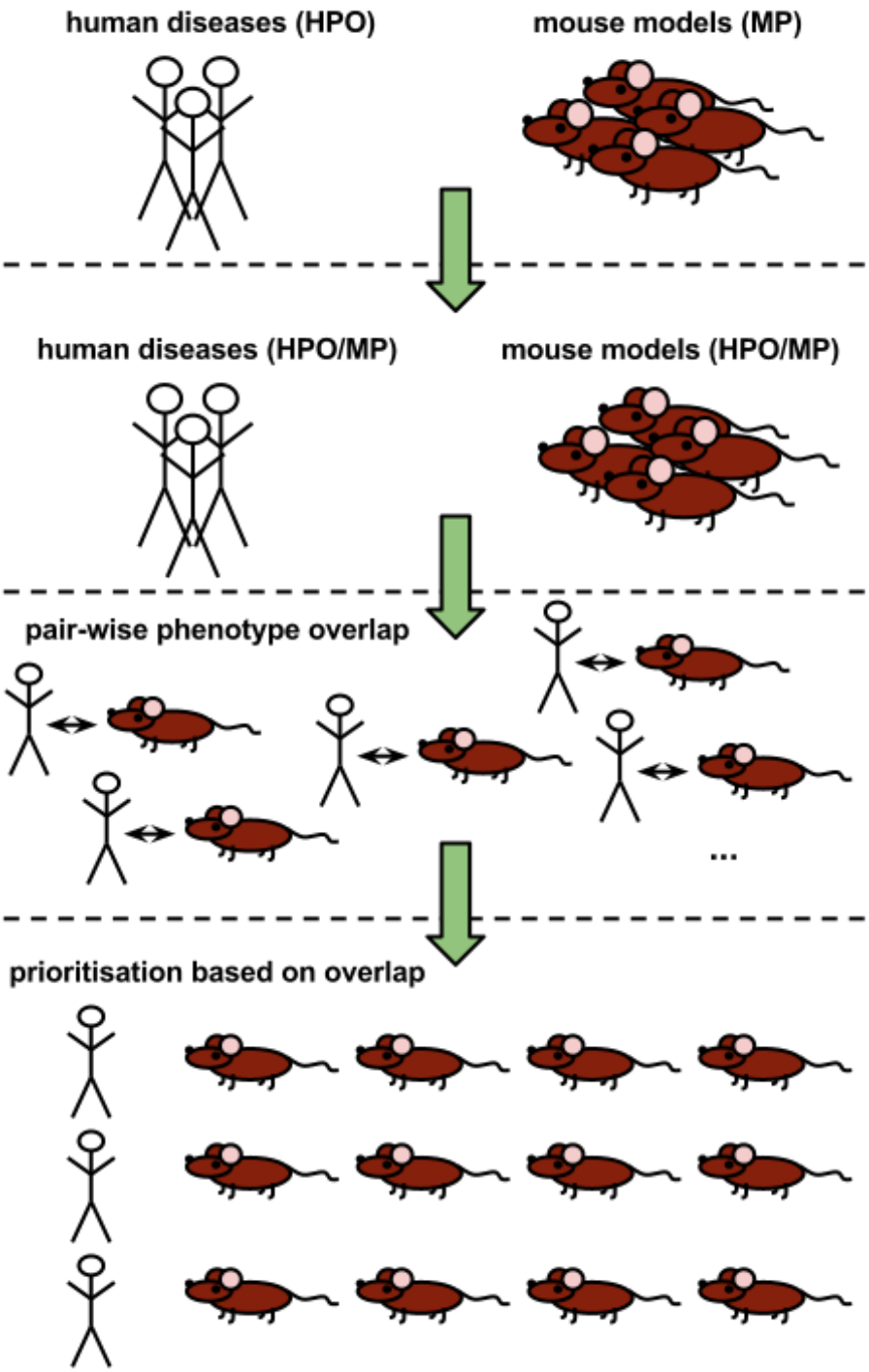


Figure 7: **Work flow for prioritising mouse models based on phenotypes.** Mouse models are obtained from MGD and are represented in MP. Human disease descriptions are collected from OMIM and represented in HPO. Before mouse models can be prioritised according to their phenotype similarity, a mapping of MP and HPO is required. Once mouse models and diseases possess the same phenotype representation, a pair-wise similarity measure can be applied, allowing each model to be ranked according to its phenotype similarity to the disease. High ranking models are assumed to be disease candidates.

in OMIM are available as HPO annotations from the HPO web site (`http://www.human-phenotype-ontology.org`). The downloaded file comprised annotations for 5,027 OMIM entries.

Human genes are referenced in OMIM's MorbidMap through gene symbols. To map between human and mouse genes, we first transferred the human gene symbols to Entrez Gene IDs and then used MGD's HMD_Human.rpt report file to map between orthologous genes based on Entrez Gene ID and MGD accession number. We used the gene symbols provided by the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) database[2] and their mapping to Entrez Gene IDs. The version employed was downloaded on the 27 October 2011 and contained 64,176 gene symbols mapped to 28,963 unique Entrez Gene IDs. 17,023 unique Entrez Gene IDs were shared between MGD's HMD_Human.rpt report file and HGNC's gene symbol mapping file.

3.2.2.1 *Ontology resources*

Mouse models obtained from MGD were phenotypically described using MP. On the other hand, OMIM's human diseases were described using HPO. To achieve a means of comparison between mouse models and diseases, a common representation between MP and HPO is required. Therefore, we obtained both MP and HPO from the OBO Foundry ontology portal[3] (Smith et al., 2007). MP was last modified on 21 June 2011 and comprised of 8,658 concepts. The applied HPO version was last modified on 26 June 2011 and contained 10,282 concepts.

3.2.3 *Phenotypic alignment of resources*

Ontological alignment is matching of concepts across ontologies to facilitate the integration of resources (see 1.2.3). To align MP and HPO, we generated mappings from one ontology to the other and *vice versa*. A *mapping* between two ontologies is a set of axioms that formally inter-relates the concepts belonging to both ontologies. We focus on mappings where the axioms relating concepts from two ontologies take the form of sub- and equivalent-classes axioms between atomic concepts. In particular, given the two concepts $A \in O_1$ and $B \in O_2$, a mapping involving both A and B will be of the form:

2 `http://www.genenames.org/`
3 `http://obofoundry.org/`

- `A SubClassOf: B`, or
- `B SubClassOf: A`, or
- `A EquivalentTo: B`.

For a concept $A \in O_1$, we will say that A maps to the concept $B \in O_2$, if A is either equivalent to B or a subclass of B. Reversely, B will be mapped to concept $A \in O_1$, if B is either equivalent or a subclass of A. Consequently, one concept can be mapped to multiple concepts, and the obtained mapping is a many-to-many instead of a one-to-one mapping. The general idea of an ontological mapping is illustrated in figure 8.

Extracting mappings with applying ontological relations yields non-symmetrical mappings between two ontologies O_1 and O_2 which consequently means that retrieving and analysing one mapping, e.g. the mapping of all concepts from O_1 to concepts from O_2, does not allow any conclusions for the reverse mapping, e.g. the mappings of all the ontological concepts in O_2 to concepts from O_1.

To achieve a mapping between HPO and MP, we employ a combination of two different approaches. The first approach relies on a lexical matching of concept labels and synonyms (see 3.2.3.1) while the second approach relies on species-agnostic ontologies and the availability of logical definitions (Gkoutos et al., 2009; Mungall et al., 2010) (see 3.2.3.2). Before combining the mappings derived from the application of either approach (see 3.2.3.4), the obtained mappings were compared for their commonalities and differences (see 3.2.3.3). The resulting mapping from the combination of both the approaches was then used to unify the representation of mouse models and disease.

3.2.3.1 *Generating mappings based on linguistic features*

Ghazvinian et al. (2009) showed that an efficient way of aligning biomedical ontologies is achieved by applying a textual alignment algorithm (see 1.2.3). The introduced textual alignment method matches concepts based on their labels and synonyms and is implemented in the Lexical OWL Ontology Matcher (LOOM) (Ghazvinian et al., 2009). LOOM matches two concepts if either their labels or any of the synonyms match exactly or deviate by only one character from each other. For example, LOOM generates a match between the HPO concept *Melena* (HP:0002249) and the MP concept *melena* (MP:0003292). However, it does not match the MP concept *abnormal pupil morphology* (MP:0001317)

and the HPO concept *Abnormality of the pupil* (HP:0000615) even though the concepts imply the same meaning and their synonyms are lexically close (*Pupillary abnormalities* and *pupil abnormalities*).

Applying LOOM to both phenotype ontologies yields a list of matched concepts. To extract mappings for the ontologies with LOOM's extracted matches, we assume that the LOOM extracted matches represent Web Ontology Language (OWL) (see 1.2.1) equivalence class axioms:

```
HPO_concept EquivalentTo MP_concept
```

In the case of the matching concepts *Melena* (HP:0002249) and *melena* (MP:0003292), the equivalence class statement is represented with the OWL axiom:

```
HP:0002249 EquivalentTo MP:0003292
```

For each of the matches generated with LOOM, the corresponding equivalence class axiom was added to a knowledge base consisting of both MP and HPO.

To extract the mappings from HPO to MP and *vice versa*, we used an automated reasoner (HermiT [4]) to classify the combined MP and HPO ontology incorporating all the equivalence class statements corresponding to LOOM's extracted concept pairs. A mapping from HPO to MP was obtained by iterating through all concepts in the HPO and performing a query for all classes that are equivalent to, or are a super-concept of the HPO concept, and belong to MP. For example, the HPO concept *Progressive hearing impairment* (HP:0001730) is mapped to the MP concepts *hearing loss* (MP:0006325), *abnormal hearing physiology* (MP:0001963), *abnormal ear physiology* (MP:0003878), *hearing/vestibular/ear phenotype* (MP:0005377) and *Mammalian Phenotype* (MP:0000001) based on the lexical match between the HPO concept *Hearing loss* (HP:0000365) and the MP concept *hearing loss* (MP:0006325) (see figure 8). The reverse mapping from MP to HPO was generated equivalently.

In the applied ontology files (see 3.2.2.1), a small number of terms were used more than once as a label or synonym: 103 terms in MP and 42 terms in HPO.

3.2.3.2 *Generating mappings based on species-agnostic ontologies and logical definitions*

Instead of generating a separate set of mappings, we used the ontology mappings applied in PhenomeNET (Hoehndorf et al., 2011c).

4 http://hermit-reasoner.com/

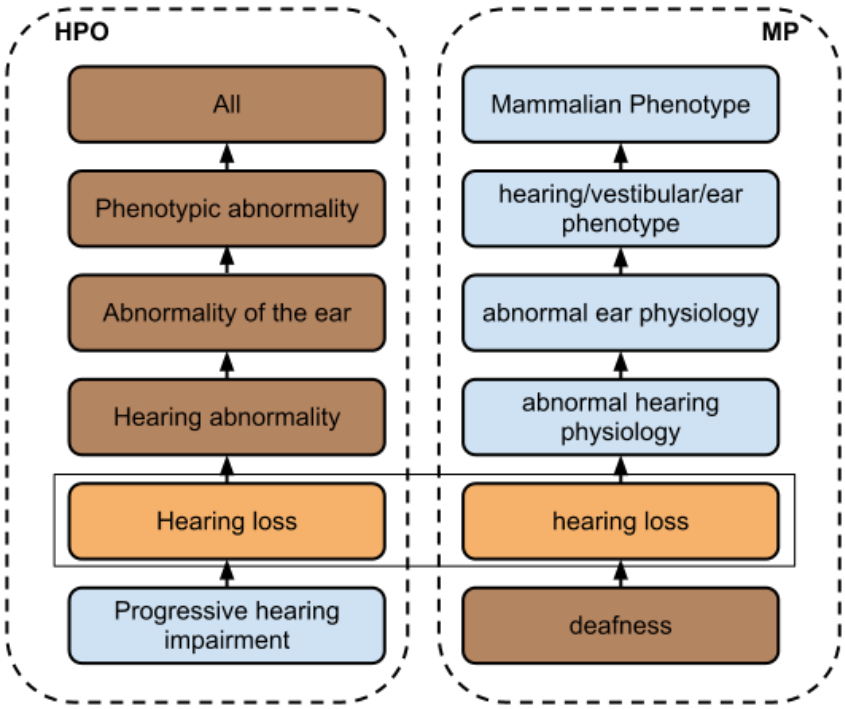


Figure 8: **Illustration of an example mapping based on lexical matching.** Concepts on the left side belong to HPO and all the concepts on the right side belong to MP. Applying LOOM to both ontologies extracted a lexical match between the HPO concept *Hearing loss* and the MP concept *hearing loss*. Based on this lexical match, the HPO concept *Hearing loss* is declared to be equivalent to the MP concept *hearing loss*, and the mapping for HPO's concept *Hearing loss* will include the MP concepts *hearing loss*, *abnormal hearing physiology*, *abnormal ear physiology*, *hearing/vestibular/ear phenotype* and *Mammalian Phenotype*. The HPO concept *Progressive hearing impairment* will be mapped to the same MP concepts as *Hearing loss*. Conversely, both the MP concepts *hearing loss* and *deafness* are mapped to the HPO concepts *Hearing loss*, *Hearing abnormality*, *Abnormality of the ear*, *Phenotypic abnormality* and *All*.

Underlying PhenomeNET is the PhenomeBLAST software. PhenomeBLAST matches concepts from HPO and MP by employing several OBO Foundry ontologies[5] (Smith et al., 2007):

- Gene Ontology (GO) (Ashburner et al., 2000)
- UBERON (Mungall et al., 2012),
- Mouse adult gross Anatomy ontology (MA) (Hayamizu et al., 2005),
- Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003),

5 http://obofoundry.org/

- Mouse PATHology ontology (MPATH) (Schofield et al., 2004)
- Chemical Entities of Biological Interest (ChEBI) (Degtyarenko et al., 2008) and
- Phenotype And Trait Ontology (PATO) (Gkoutos et al., 2009)

Most of the OBO Foundry ontologies are available as OWL files and with that allow for knowledge inference through reasoning over the ontologies. However, Mungall et al. (2010) as well as Golbreich et al. (2006) noted that there are serious issues with reasoners not scaling with the combination of OBO Foundry ontologies, preventing reasoning over integrated ontologies. The issues arise from the complexity and amount of knowledge contained in the ontologies.

In its latest version (OWL2), OWL supports three different profiles: OWL EL, OWL QL and OWL RL[6]. While OWL EL limits the expressivity in OWL, e.g. negations are not allowed, it enables reasoning over the combined OBO Foundry ontologies. Not all existing reasoners support all the three profiles defined in OWL2. A list of available reasoners for the different profiles is reported on the World Wide Web consortium (W3C) web page[7].

Schulz et al. (2009) demonstrated how reasoning for SNOMED Clinical Terms (SNOMED CT) (see 1.2.1) can be improved by transferring the ontology into OWL EL. Following on from this success, Hoehndorf et al. (2011a) showed that multiple OBO Foundry ontologies can sucessfully be integrated and reasoned over by transferring them into an OWL EL profile. From this study, the EL Vira software originated[8]. Hoehndorf et al. (2011a) also tested some of the existing reasoners and reported their performance on OBO Foundry ontologies being transferred to OWL EL with EL Vira.

To reason over the incorporated ontologies, PhenomeBLAST first converts all the ontologies into OWL EL using EL Vira (Hoehndorf et al., 2011a). The OWL EL profiles of the ontologies are then integrated into one combined ontology. The available logical definitions for MP as well as HPO concepts (Mungall et al., 2010) are added as axioms to this integrated ontology. PhenomeBLAST uses the CB reasoner (Kazakov, 2009) to reason over the combined ontology for the logical definitions and generate the mappings from MP to HPO. The CB reasoner is one of the reasoners capable to reason over ontologies formally described in OWL EL.

6 http://www.w3.org/TR/owl2-profiles/
7 http://www.w3.org/2007/OWL/wiki/Implementations
8 https://code.google.com/p/el-vira/

Then, PhenomeBLAST identifies all equivalent and super-classes of an MP concept in HPO, and *vice versa* for the mappings from HPO to MP. To apply the mappings generated with PhenomeBLAST and incorporated in PhenomeNET, we downloaded the mappings from `http://phenomeblast.googlecode.com` and transformed them into our required input format. Due to the incorporation of the ontologies' structures, the mappings provided by PhenomeBLAST are also non-symmetrical and mostly provide a list of mapped concepts instead of a single mapped ontological concept.

3.2.3.3 *Comparison of mappings from both mapping approaches*

To further analyse the obtained mappings by either method, we determined their overlap. Concepts which were not mapped with both methods were excluded from this comparison. Both methods generate a list of mapped concepts (see 3.2.3.1 and 3.2.3.2). Possible ways of comparison include either to compare both the lists or to compare the most specific concepts of either list. Here, we did not only compare the most specific concept, instead, we compared the lists of concepts for a single ontology concept with each other. Due to non-symmetrical mappings, we independently assessed both mappings: HPO to MP and MP to HPO.

While comparing the results, we could identify four different categories the results fall into:

1. exact overlap of the mappings (*exact*)
2. the lexical mappings are a subset of the ontological mappings (*lexical* $\subset$ *ontological*)
3. the ontological mappings are a subset of the lexical mappings (*ontological* $\subset$ *lexical*), and
4. the lexical and the ontological mapping overlap for a number of mapped concepts but each possesses also concepts not contained in the other (*overlap*).

An illustration of the four overlap categories is provided in figure 9.

3.2.3.4 *Combining both approaches into one consistent mapping*

Fung et al. (2007) suggested that lexical and ontological mapping for anatomy ontologies could be used in a complementary way. Thus, we assumed that combining the obtained lexical mappings between

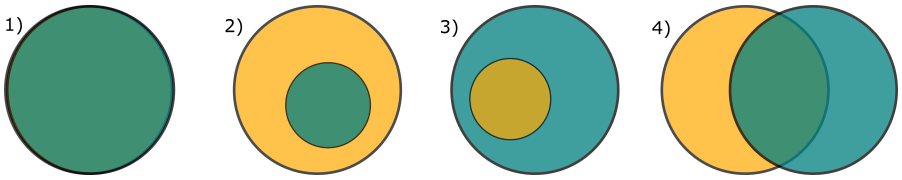


Figure 9: **Illustration of the obtained overlap categories.** Due to both mappings generating mostly lists of mapped concepts instead of only a single mapped concept, we obtained four overlap categories: exact overlap, lexically mapped concepts are a subset of the formally mapped, formally mapped concepts are a subset of the lexically mapped, and both mappings share common concepts but both also contain concepts not mapped by the other.

phenotype ontologies with PhenomeBLAST's mappings would lead to a richer ontology, enabling better results in the application for disease gene prioritisation. Therefore, we generated a novel mapping based on the formal definitions of the concepts in phenotype ontologies and the lexical matches between the concepts' labels and synonyms.

We modified the PhenomeBLAST software to add the additional equivalent class axioms derived from the lexical matches (see 3.2.3.1) to PhenomeBLAST's underlying ontology and used the modified PhenomeBLAST ontology to re-generate mappings between HPO and MP (see figure 10). Combining the mappings with the described procedure yields again non-symmetrical mappings and lists of mapped concepts instead of a pair of mapped concepts.

3.2.4 *Generating a common ontological representation of models and diseases*

While MGD's mouse model annotations are represented using MP, human diseases in OMIM are annotated with HPO. This leaves three options for a unified representation:

1. a common representation in MP
2. a common representation based on HPO, or
3. a common representation based on HPO and MP.

While other approaches use a combination of the ontologies, we limited the representation of phenotypes to one ontology, either MP or HPO. This consequently leads to the following six gene prioritisation tasks:

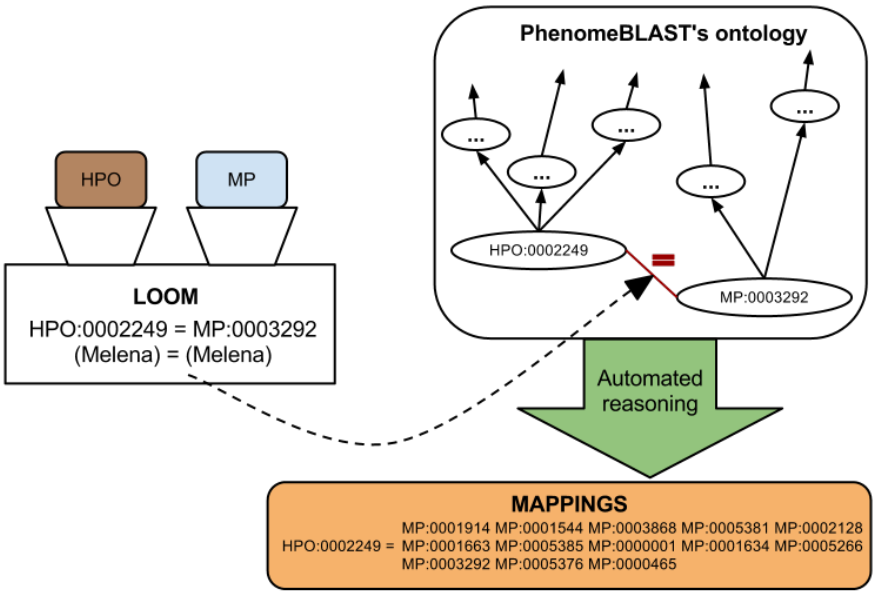


Figure 10: **Integration of lexical and ontological mapping.** LOOM is applied to MP and HPO to extract lexical matches of concepts (based on labels and synonyms). All concepts matched by LOOM are inserted as equivalence class axioms into PhenomeBLAST's ontology. Mappings are generated by reasoning over PhenomeBLAST's adapted ontology and extracting all equivalence and super classes for each concept.

1. **mouse models** converted applying the **lexical mapping** and **original disease** annotations
   (**HPO-based** similarity score and prioritisation)

2. **mouse models** converted applying the **ontological mapping** and **original disease** annotations
   (**HPO-based** similarity score and prioritisation)

3. **mouse models** converted applying the **combined mapping** and **original disease** annotations
   (**HPO-based** similarity score and prioritisation)

4. **disease annotations** converted applying the **lexical mapping** and **original mouse model** annotations
   (**MP-based** similarity score and prioritisation)

5. **disease annotations** converted applying the **ontological mapping** and **original mouse model** annotations
   (**MP-based** similarity score and prioritisation)

6. **disease annotations** converted applying the **combined mapping** and **original mouse model** annotations
(**MP-based** similarity score and prioritisation)

3.2.5 *Prioritisation based on ontology annotations*

After the alignment of the phenotype representations of human diseases and animal models, both resources can be represented in a unified way. The unified representation is obtained by applying the mappings to convert between ontologies. Once mouse models and diseases are represented within one ontology, a similarity measure (see 1.3.2) is applied to determine the phenotype overlap. Due to mouse models and diseases being annotated with a set of ontology concepts, we decided to apply the Jaccard index. The Jaccard index is a group-wise calculated similarity measure and is calculated as follows:

$$sim(D, M) = \frac{|P(D) \cap P(M)|}{|P(D) \cup P(M)|} \tag{3.1}$$

where P(D) are disease phenotypes and P(M) are mouse model phenotypes. The semantic similarity measure is calculated for all possible combinations of diseases and mouse models. This leads to a matrix of mouse models and diseases with the corresponding calculated similarity of model and disease as elements.

Mouse models are then ranked according to their phenotype semantic similarity with a disease. Mouse models showing a high phenotype overlap with the disease are assumed to be more relevant to this disease than mouse models showing less overlap in their representation. The mouse model possessing the highest phenotype similarity with the disease is assigned rank one and the model with the lowest phenotype similarity is assigned the lowest possible rank. As more than one mouse model may possess the same phenotype similarity score with the disease, our ranking algorithm also resolves ties: the average rank is calculated between all mouse models sharing the same phenotype similarity score. With applying the ranking algorithm, we obtained a ranked list of mouse models for a disease with the highest ranks (starting from or close to rank one) representing the best candidates for a disease.

3.2.6 *Evaluation of prioritised mouse models*

When prioritising mouse models for human genetic disorders, a large number of results are produced, due to the combination of all available phenotypically described mouse models and disorders. There are too many results for it to be practical for them to either be experimentally evaluated or manually curated. One possibility to establish a connection between a mouse model and a disease would be to choose a threshold. If a mouse model possesses a phenotype similarity with a disease that is equal or higher than the chosen threshold, then this mouse models would be considered relevant to this disease. However, given the number of diseases and a potential of incomplete phenotype annotations, finding a threshold which is representative for all the diseases assessed with the method, is not likely. We therefore chose to apply an automated evaluation using receiver operating characteristic (ROC) curves (see 1.4.2.2) and manually assessed a single disease and its highest ranked mouse models.

The ROC analysis was performed twice (for each of the six prioritisation tasks; see 3.2.4), using either a set of known gene–disease associations in humans (OMIM's MorbidMap) or using a set of gene–disease associations in mice (disease annotations available in MGD). In the absence of a reliable set of true negative gene–disease associations, we assume that only *known* gene–disease associations constitute *positive* examples while all *unknown* associations constitute *negative* examples.

3.2.6.1 *Transforming OMIM's MorbidMap into a mouse-specific representation*

Before OMIM's MorbidMap (see 3.2.2) could be used for the automated evaluation of the prioritisation of disease gene candidates, it required transformation in a mouse-specific representation. An entry in MorbidMap looks as follows:

```
Septooptic dysplasia, 182230 (3)|HESX1, RPX, CPHD5|601802|3p14.3
```

Database entry fields are separated by bars ("|") in MorbidMap. The first field corresponds to the disease, the second field contains the human gene names, the third contains OMIM identifiers referencing OMIM gene entries to the disease and the fourth field provides chromosomal regions.

Human genes were mapped to mouse genes via orthology using MGD and the Human Genome Organisation (HUGO) Gene Nomenclature Committee (HGNC) database [9]. This database contains MorbidMap's gene symbols as well as Entrez Gene IDs that can be used to map to MGD's mouse genes (from the HMD_Human5.rpt, see 3.2.2). The resulting mouse-specific MorbidMap looked like (MGI:96071 corresponds to the mouse gene symbol *Hesx1*):

```
182230 MGI:96071
```

Each gene was then extended to all mouse models possessing an allelic variant of this gene. Applying this step yielded a mouse specific MorbidMap also possessing model–disease associations (similar to MGD's model–disease associations):

```
182230  Hesx1<+>|Hesx1<tm1Icar>
        Hesx1<tm1(cre)Jpmb>
        Hesx1<tm1Icar>
        Hesx1<tm1Icar>|Hesx1<tm2Jpmb>
        Hesx1<tm1Icar>|Hesx1<tm3Jpmb>
        Hesx1<tm2Jpmb>
        Hesx1<tm3Jpmb>
```

3.2.6.2 *Comparison to PhenomeNET*

Two other solutions applying mouse models to the prioritisation of disease genes are PhenomeNET and MouseFinder (Hoehndorf et al., 2011c; Chen et al., 2012). Hoehndorf et al. (2011c) reported an AUC score of 0.68, measured on a combined evaluation set of gene–disease associations obtained from MGD and OMIM. MouseFinder only reports recall measures and therefore lacks the means of a direct performance measure comparison.

To assess the differences in performance to our approach, we applied PhenomeNET to both our benchmark sets (gene–disease associations from OMIM and MGD; see 3.2.6) and determined its performance as AUC score. Hanley and McNeil (1982) described a method to compare ROC curves with each other and provided this method as an online tool[10]. We applied this tool to determine the differences between our method and PhenomeNET.

9 http://www.genenames.org/
10 http://vassarstats.net/roc_comp.html

3.3 RESULTS

3.3.1 *Applying the mapping algorithms*

Since the mouse models in MGD and the OMIM diseases possess different phenotype representations, a mapping between HPO and MP was required. To map both the ontologies to each other, we used a combination of a lexical (realised with LOOM; see 3.2.3.1) and an ontology-based mapping (realised with PhenomeBLAST; see 3.2.3.2). Both mappings were obtained independently at first, then compared to each other and finally combined after comparison. All three mappings (lexical, ontological, combined) were applied in the gene prioritisation task.

3.3.1.1 *Obtaining lexical and ontological mappings*

Applying LOOM (see 3.2.3.1) to HPO and MP, we extracted 607 pairs of corresponding HPO and MP concepts (one-to-one mapping; see 1.2.3). These 607 initial pairs were manually verified by a domain expert. We then added the 607 pairs as equivalent class axioms to a knowledge base consisting of HPO and MP and applied a reasoner to extract the mappings. As a result, we were able to map 3,027 (29%) concepts for HPO and 1,254 (14%) MP concepts. Note that the resulting mappings are non-symmetrical and constitute a many-to-many alignment.

Applying PhenomeBLAST's mapping algorithm (see 3.2.3.2), yields mappings for 9,978 (91%) HPO and 7,985 (92%) MP concepts. Note that the application of PhenomeBLAST also yields a non-symmetrical mapping, which constitutes a many-to-many alignment.

3.3.1.2 *Comparison of both mappings*

Both mappings share 3,019 mapped concepts for HPO and 1,251 for MP, which leaves only a small remainder of the lexical mappings not being provided by the ontological mappings. When assessing the differences in both the obtained mappings, the outcomes were recorded according to the four groups defined in 3.2.3.3.

Table 2 shows that the mappings obtained with either method diverge from each other, even though both methods share some common information for all compared mappings. The lexical mapping method works well for clinical terminology, such as *osteoarthritis* (MP:0003560, HP:0002758) or microcephaly (MP:0000433, HP:0000252). However, this clinical terminology is difficult to represent with formal definitions

as they mostly constitute a collection of phenotypes. Concepts corresponding to those clinical terms could be a cause for the *ontological* $\subset$ *lexical* category.

The ontological mapping method has an advantage over the lexical mapping method whenever it is required to map to more than one concept. For instance, the HPO concept *Abnormality of the cardiovascular system* (HP:0001626) corresponds to both of the following MP concepts, *abnormal cardiovascular system morphology* (MP:0002127) and *abnormal cardiovascular system physiology* (MP:0001544). Those concepts requiring more than one mapped concept could explain the category *lexical* $\subset$ *ontological*.

	HPO to MP	MP to HPO
# exact	110	93
# lexical $\subset$ ontological	1367	502
# ontological $\subset$ lexical	226	88
# overlap	1316	568
# concepts	3019	1251

Table 2: **Illustrates the amount of mappings falling into each of the overlap categories when both methods are compared.** Due to the mappings between HPO and MP not being symmetrical, the mappings are independently compared, once for the mappings from HPO to MP direction and once for the mappings from MP to HPO.

To further assess the difference of the mappings, we also investigated the number of mapped concepts for each of the obtained mappings and averaged this number over all available ontological concepts for either HPO or MP. Using the ontological mappings generated through PhenomeBLAST, concepts from HPO are, on average, associated with 7.1 concepts from MP and MP concepts with 9.3 HPO concepts. Through lexical matching (using LOOM), HPO concepts are associated, on average, with 2.3 MP concepts and MP concepts with 1.0 concepts from HPO. When combining the mappings, the average number of mapped concepts increases to 7.8 concepts from MP that are associated with an HPO concept and 9.7 HPO concepts that are associated with an MP concept which also shows that the lexical mapping contains complementary information to the ontological mapping.

However, mapping through lexical matching produces, on average, significantly less concepts. For 71% of the concepts in HPO, we were unable to identify any corresponding MP concepts through the lexical matching approach. Similarly, for 86% of the MP concepts, no

corresponding HPO concept could be identified. By combining both approaches a richer – in terms of number of concepts mapped per concept – is obtained. This richer mapping can potentially improve candidate gene prioritisation based on the integration of phenotype information.

3.3.2 *Disease gene identification through gene prioritisation*

First, we applied the three different HPO to MP mappings (lexical, ontological, combined) to convert human disease phenotypes from an HPO- into an MP-based representation (see 3.2.4). Second, we used the three MP to HPO mappings (lexical, ontological, combined) to convert mouse phenotypes from an MP- into an HPO-based representation (see also 3.2.4). After converting the data into a common representation, we computed the phenotype similarity between all mouse models in MGD and all OMIM diseases which possess a phenotype description (see 3.2.5). Phenotype similarity was calculated once based on MP and once based on HPO.

To evaluate the performance of our prioritisation algorithm, we compared our results against known gene–disease associations contained in MGD and OMIM (twice for each of the six prioritisation tasks; see 3.2.6). As a result, we obtained 12 ROC curves with their associated AUC values. Figure 11 illustrates the resulting ROC curves.

Table 3 shows that the highest AUC measure is obtained, when mouse models are prioritised and evaluated against MGD's gene–disease associations. AUC scores are in general higher when evaluated against MGD's gene–disease associations and also are generally higher when using MP instead of HPO for gene prioritisation.

based on	HPO*			MP†		
mapping	lex	ont	comb	lex	ont	comb
OMIM	0.678	0.690	0.700	0.732	0.727	0.730
MGD	0.691	0.737	0.748	0.864	0.895	0.899

Table 3: **Areas Under Curve (AUC) measures for all gene prioritisation tasks.** First row: AUC values for evaluation against OMIM data; second row: AUC values evaluation against MGD data; * HPO- and † MP-based prioritisation. *lex*ical, *ont*ological, and *comb*ined mapping.

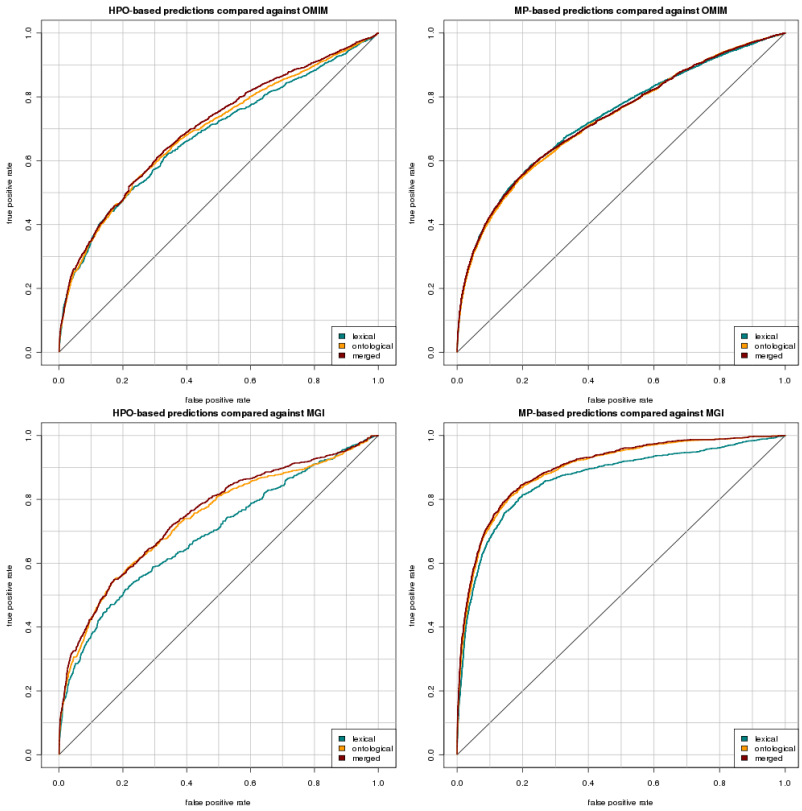


Figure 11: **ROC curves resulting from evaluation.** The left panel includes all the results for conversion of mouse models from an MP representation to an HPO representation and performing the gene prediction in HPO. The right panel includes all the results for conversion of human diseases into an MP-based representation. Each plot shows the evaluation results using each of the three mappings: lexical, ontological and combined. The two panels on the top are the results of the evaluation against OMIM and the two panels at the bottom are the results of the evaluation against MGD.

3.3.2.1 *Comparison to PhenomeNET*

To assess the differences in performance of our method and PhenomeNET, we evaluated PhenomeNET's performance using our evaluation sets (see 3.2.6.2). The AUC measure was recorded and its performance to our method assessed with the method described by Hanley and McNeil (1982). For comparison, we used our best performance measures obtained for MGD and OMIM (see 3), employing the *combined* mapping. Our approach achieves a significantly higher performance than PhenomeNET (p-values determined based on one-tailed Student's t-test)):

- $p = 3.2 \times 10^{-4}$, based on OMIM's gene–disease associations
- $p < 1 \times 10^{-6}$, based on MGD's gene–disease associations

3.3.2.2 *Prioritising candidate genes for orphan diseases*

Based on the results of our quantitative evaluation, we can apply our method's prioritisation results to suggest candidate genes for orphan diseases. Orphan diseases are diseases that occur rarely in a population. Due to their rare occurrences, orphan diseases are difficult to investigate. Consequently, little to nothing is known about their genetic causes. By prioritising disease gene candidates for orphan diseases, the design and priority of biological experiments could be influenced. Studies in mutagenesis projects such as the International Knockout Mouse Consortium (Austin et al., 2004) could be prioritised according to highly ranked genes.

To verify the potential of our method to correctly prioritise disease gene candidates, we manually assessed a small subset of the prioritisation results obtained when calculating phenotype similarity based on MP and using the combination of lexical and ontology-based mappings (corresponds to the scenario where we achieved the highest AUC score; see table 3).

For example, our method selects knockouts of *Gdf6* (MGI:95689), *Marcks* (MGI:96907) and *Vax1* (MGI:1277163) on ranks 1, 2 and 3 for Septo-Optic Dysplasia (SOD) (OMIM:#182230). Investigating further, we can suggest that *Vax1* could be a candidate gene for patients suffering from SOD.

SOD is a disorder characterised by any combination of optic nerve hyperplasia, pituitary gland hyperplasia, and midline abnormalities of the brain, including absence of the corpus callosum and septum pellucidum (Dattani et al., 1998). *Vax1* mutations in mice share remarkable phenotypic similarities with SOD in humans as illustrated in figure 12. For example, both the disease and the mouse models are annotated with *abnormal eye development* (MP:0001286), *abnormal optic nerve morphology* (MP:0001330), and *absent corpus callosum* (MP:0002196). Our results confirm a recent study in which *Vax1* has been suggested as a strong candidate gene for SOD when no *Hesx1* (MGI:96071) mutations are present (Bharti et al., 2011). Details on the steps involved in prioritising *Vax1* for SOD, and parts of our input data (full data available in appendix B), are illustrated in figure 12.

Furthermore, our method ranks *Gdf6* and *Marcks* on first and second place for SOD. *Gdf6* has previously been identified to implicate ocular and skeletal abnormalities (Asai-Coakwell et al., 2007), in particular abnormalities of the coronal suture between bones in the skull (Settle et al., 2003), while deficiency of the *Marcks* protein in mice has been

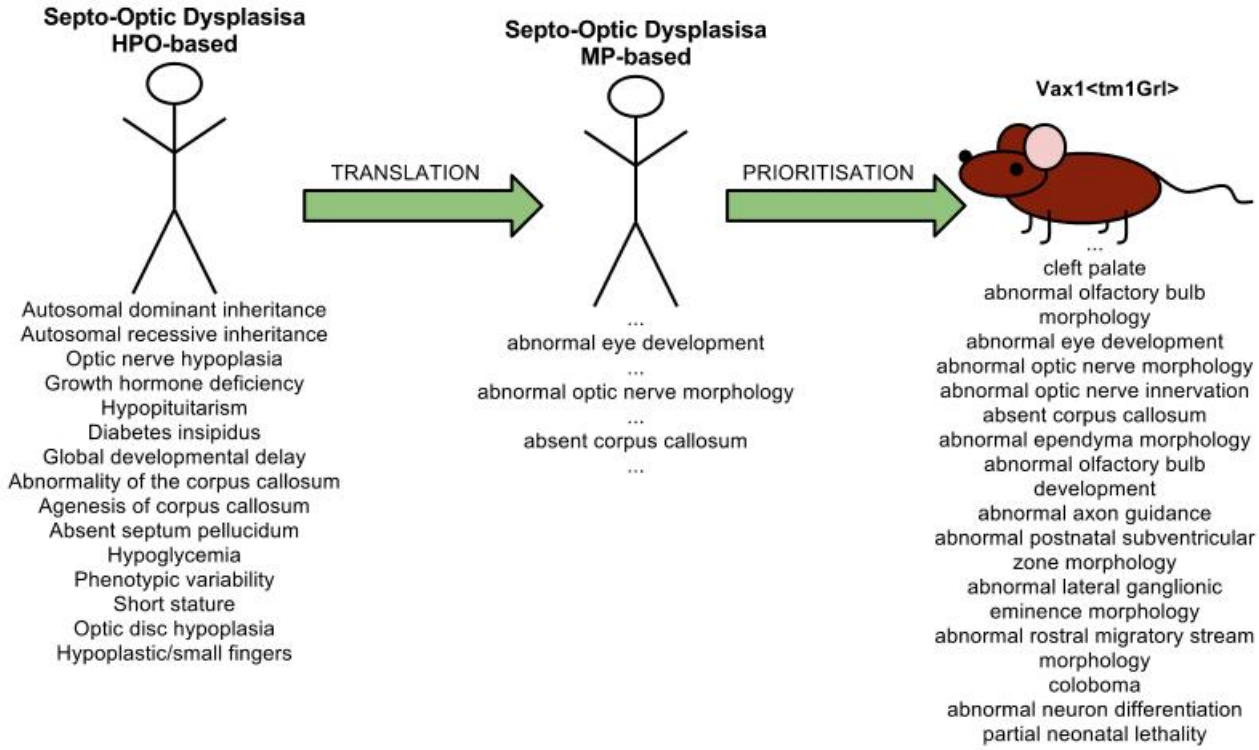


Figure 12: ***Vax1* is one of the highest ranked mouse model for SOD.** Human diseases converted to MP with combined mapping (results with highest AUC score; see 3). Manual verification of some MP-based prioritisation results (including SOD. Figure illustrates original annotation for SOD based on HPO and its conversion to MP. Also includes MGD's annotation for mouse models with *Vax1<tm1Grl>* (MGI:1859863) allele. Figure does not include all annotations, to reduce complexity (full data available in appendix B).

shown to result in an absence of the corpus callosum, and cortical and retinal abnormalities (Stumpo et al., 1995). Based on their phenotype similarity to SOD (full information provided in appendix B), *Gdf6* and *Marcks* are promising novel candidates for genes involved in SOD.

Both OMIM and MGD possess *HESX1* as causative gene for SOD. Our approach also identifies a *Hesx1* model on rank 22.

3.4 DISCUSSION

3.4.1 *Comparison to related work*

Several studies showed the importance and suitability of automatically exploring the integration of phenotype data (Hoehndorf et al., 2011c; Chen et al., 2012; Washington et al., 2009) (see 1.1.3). Moreover, PhenomeNET (Hoehndorf et al., 2011c) and MouseFinder (Chen et al., 2012) are both tools utilising mouse phenotypes to identify candidate genes for human diseases. Although the underlying data sources are the same, our approach differs from both solutions in the way the data is integrated and mouse models are prioritised.

Our approach differs from PhenomeNET, not only by expanding the mapping, but also in the phenotype similarity calculation. By applying ontological mappings between species-specific ontologies, PhenomeNET integrates five different species. To calculate phenotype semantic similarity, PhenomeNET maintains annotations of five different ontologies while in our approach, we only use one ontology at a time. The inclusion of multiple, often redundant (i.e., equivalent) phenotypes may introduce additional noise that affects the resulting similarity values. To calculate phenotype similarity, PhenomeNET also incorporates weighing scores for phenotypes (see 1.3.2) in their calculation which we left out. Applying a weighing mechanism may improve our results but this is subject to further tests. Comparing the performances of PhenomeNET and our approach based on MGD and OMIM shows that are our approach significantly outperforms PhenomeNET (see 3.3.2.1).

MouseFinder, similar to our approach, uses a combination of an ontological and a lexical mapping, but their mappings may still differ from ours. No data has been provided for MouseFinder on how the lexical mapping is achieved and how both lexical and ontological mappings are combined. Furthermore, MouseFinder makes a different assumption in the way phenotype similarity is calculated. Our approach

uses a group-wise phenotype similarity score while MouseFinder employs a pairwise phenotype semantic similarity (see 1.3.2). By applying a pairwise semantic similarity measure, MouseFinder picks up on significant phenotypes while our approach maintains the disease or mouse model phenotype description as a whole. Moreover, MouseFinder also employs a weighing before calculating the phenotype similarity (as does PhenomeNET) which we do not include in our approach.

In total, within the first 500 ranks, MouseFinder reports a recall of 58% when compared against OMIM's gene–disease associations and 65% when compared against MGD's associations. Since multiple mouse models with the same similarity measure share a rank, we cannot derive a precision for MouseFinder and consequently lack the means of a comparison of performance.

3.4.2 *Opportunities for orphan diseases*

Due to the rare occurrence of orphan diseases, it is difficult to investigate those diseases and establish common patterns of their genetic origin. This means that approaches employing identified gene–disease associations to profile a disease ("guilt-by association"; see 1.3.1) cannot be applied to orphan diseases.

However, phenotype descriptions do exist for orphan diseases and can be used to identify the genetic mechanisms. This makes our method especially valuable for orphan diseases as it only requires a phenotype description. The potential lying in our approach has been demonstrated with the example of SOD (see 3.3.2.2).

3.4.3 *Lexical versus ontological mapping*

In almost all cases, the ontological mapping performs better than the lexical mappings. There is only one case, where the lexical mapping performs better than the ontological. This case occurs when phenotype similarity is calculated based on MP and the results are evaluated against OMIM. In this case, applying the lexical mappings yields an AUC score of 0.732 while the ontology-based mapping leads to an AUC score of 0.727. It seems surprising that the lexical matching of initially 607 concepts (see 3.3.1.1) performs similarly to ontology-based mappings (based on more than 10,000 formal concept definitions in both the MP and HPO) when applied to the task of gene prioritisation.

A possible explanation lies in:

- the annotation depth of mouse models in the MGD database
- disease annotations in the OMIM database, as well as
- the depth of concepts that match exactly between the MP and HPO.

On average, mouse models in the MGD are annotated at a depth of 5 in MP (Espinosa and Hancock, 2011). The concepts that lexically match exactly between HPO and MP, however, are mostly specialised, clinical terms that are used for annotating disease-related phenotypes in OMIM. These terms denote complex concepts that carry substantial information about a disorder. As a result of their complexity, they are often not formally defined and would therefore not map completely across species when using ontology-based mappings. If an appropriate MP concept can be identified, all mouse models that are annotated with it or any of its super-concepts will share features with the clinical term. Consequently, mouse models will also share some phenotype overlap with the disease that is annotated with the clinical term.

3.4.4 *Suitability of benchmark sets*

Further to consistently showing higher performance scores when applying MP instead of HPO, our results also show consistently higher performance measures when evaluating against MGD's gene–disease associations instead of OMIM's gene–disease associations. One possible explanation for this behaviour is the difference in species underlying MGD and OMIM. While OMIM's MorbidMap contains gene–disease associations confirmed in human, MGD possesses gene–disease associations in mouse. Because of our approach ranking mouse genes, we could expect a better performance on MGD than on OMIM. To further address this aspect, we plan to include other benchmark sets in the future, also covering different species.

3.5 CONCLUSIONS

With the integration of two phenotype ontologies, HPO and MP, we allowed for the integration of data provided from MGD and OMIM. Due to the integration of both the resources, we could use the combined data set to identify promising mouse models for human genetic

disorders based on a phenotype similarity measure (see 3.3.2.2). We could demonstrate that our approach performs significantly better than PhenomeNET (see 3.4.1), a phenotype-based network for disease candidate prioritisation. Furthermore, by using the example of SOD, we could show the potential of the approach to be applied to diseases where only little to nothing is known about the genetic components underlying the disease. This opens possibilities to integrate such approaches in planning stages of biological experiments and with that guide biological discovery.

3.6 FUTURE WORK

Applying LOOM as the lexical matching component with its algorithm only allowing for one mismatching character, is a very stringent method to lexically map the concepts of ontologies. In further studies, we would like to explore other mapping algorithms based on lexical features of the phenotype ontologies and assess their performance. Possibilities include e.g. AgreementMaker (Cruz et al., 2009) or OntologyMatcher (Adamusiak et al., 2011).

Further to extending the mappings from HPO to MP and *vice versa*, we would like to explore phenotype similarity measures. Due to the existing variety of similarity measures (Pesquita et al., 2009) and the conclusion that better phenotype similarity measures are still to be found (Zhang et al., 2012), another scoring mechanism, may improve the identification of disease gene candidates. For example, employing frequency information of phenotypes may improve our prioritisation results.

To date, a variety of gene prioritisation tools exist, considering sometimes different and sometimes overlapping features to prioritise genes. Despite their existence, no generalised benchmark set was established to evaluate their performance and provide a means of comparison. Tranchevent et al. (2010) aim to generally compare the different solutions of gene prioritisation and work on the *Critical Assessment* of gene prioritisation tools. So far, we have benchmarked our solution only against other phenotype-based gene prioritisation tools. Another step in future work will be the comparison against more benchmark sets. We aim to use the *Critical Assessment* data described on the *Gene Prioritization Portal* [11].

11 http://homes.esat.kuleuven.be/~bioiuser/gpp/

As illustrated for SOD, our approach shows the potential to prioritise biologically meaningful gene candidates. Therefore, another avenue in future research will be the manual investigation of diseases together with their ranked genes. We intend to derive a list of promising candidate genes for a subset of the diseases. Those candidates can then be provided to experimental biologist for further verification if not reported in literature already.

4

MOUSE-SPECIFIC PROFILES TO ANNOTATE GENETIC DISEASES

In chapter 3, I described a method to prioritise disease gene candidates using phenotype data of human diseases and results of mutagenesis experiments in mice (see 1.1.3). A requirement for the prioritisation of human disease gene candidates based on phenotype information is the definition of diseases based on their signs and symptoms. In this chapter, I report about my work that focused on the definition of diseases using results from biological experiments and known gene–disease associations. Furthermore, deriving new definitions for diseases possessing a phenotype description may lead to the enrichment of the existing phenotype description.

4.1 BACKGROUND

The successful application of phenotype information to support biological discovery has been shown in many cases (Korbel et al., 2005; van Driel et al., 2006). Solutions exist using cross-species data analyses to support disease gene discovery (Hoehndorf et al., 2011c; Chen et al., 2012; Washington et al., 2009). Some of these solutions use phenotype descriptions provided by databases such as the Online Mendelian Inheritance in Man (OMIM) database (Amberger et al., 2011) or OrphaNet (Rath et al., 2012). Thus, the success of the studies employing phenotype information depends on the quality and the coverage of the phenotype data in the applied databases.

Investigating the coherence of a subset of human phenotype databases, Oti et al. (2009) showed that the phenotype data held in those databases is incomplete. Despite the best efforts, phenotype information may be missed when defining a disease or the phenotype might not yet be evident (Schofield et al., 2011). Incompleteness may also evolve when diseases are rarely occurring and it is hard to define common phenotype patterns (see 1.1.1). Furthermore, inheritable diseases change their phenotypic appearance over generations and may express more severe phenotype effects in succeeding generations (Warren and Nelson, 1993). Oti et al. (2009) suggested a way of enrich-

ing the information but only focused on integrating human-specific phenome databases.

Previously it has been shown that mouse models can be used to identify phenotype patterns also holding true in human (Lisse et al., 2008). With the increase in phenotype data available through large-scale mutagenesis experiments (Abbott, 2010), the subsequent data repositories constitute an enormous resource for building phenotype profiles. One database hosting phenotype information from large-scale mutagenesis experiments is the Mouse Genome Informatics database (MGD) (Blake et al., 2011) (see 2.4.2).

In this chapter, mouse-specific disease profiles are introduced which have been derived from mouse model-disease associations and phenotype annotations to mouse models. Underlying the built process of the mouse-specific disease profiles is the idea of "guilt-by association" approaches (see 1.3.1) to use previously identified gene candidates to profile diseases. The derived mouse-specific profiles are investigated for their validity by applying them in a gene prioritisation task. Furthermore, the mouse-specific disease profiles are used to generate a mouse-specific disease classification. With manual investigation of the obtained classification, it can be demonstrated that the classification can serve to enrich existing human-centred disease phenotype descriptions such as those contained in OMIM.

Results, source code and additional information are available online[1]. For more details refer to page xi.

4.2 METHODS AND MATERIALS

The underlying idea for building mouse-specific disease descriptions is taken from "guilt-by association" approaches, where the collective information about certain features of genes, e.g. Gene Ontology (GO) annotations, is taken to profile a disease (see 1.3.1). In this particular case, the characterising feature of the genes was the phenotype that is exhibited when this particular gene is mutated. To derive mouse-specific disease profiles for the enrichment of phenotype descriptions of human diseases, we used the phenotype descriptions available from MGD (Mammalian Phenotype Ontology (MP) (Smith et al., 2005) annotations) and two different sets of gene–disease associations: one set of mouse model–disease associations from MGD and one set of gene–disease associations from OMIM. Therefore, a model of a disease

1 https://code.google.com/p/phendis/

is a mouse model that has been associated either in MGD directly (based on phenotype data) or via orthology from OMIM (based on gene information). Following our approach, we obtained two different sets of disease profiles:

1. one set of disease profiles based on model–disease associations in MGD (further referred to as **MMM**)
2. one set of disease profiles based on gene–disease associations in OMIM (further referred to as **MOM**)

Due to the differences in how genes and models are assigned to diseases in MGD and OMIM, the derived disease profiles could look substantially different in both **MMM** and **MOM**. In MGD, it is not the gene that is associated with a disease, instead it is a certain allelic variant (see 1.1.2) in a specific mouse model that is assigned to the disease. This means that not all allelic variants of the gene may be associated with the disease, which may be advantageous in cases where only one function of a multi-functional gene is relevant to a disease. On the contrary, OMIM assigns genes instead of allelic variants to a disease. Furthermore, either database only assigns allelic variant or gene to a disease if it is reported in the respective species (mice for MGD and human for OMIM). This may also lead to differences in the profiles as a disease may not be associated with the same gene in both species.

4.2.1 *Input data*

4.2.1.1 *Mouse-specific data*

To include mouse phenome data in our study, we obtained three of the report files provided by MGD (all accessed on 9 March 2011):

- `MGI_GenoDisease.rpt`,
- `MGI_GenePheno.rpt` and
- `HMD_Human5.rpt`.

The first file provides gene–disease associations with 959 diseases, 1,135 genes and 2,088 model–disease associations. The second file contains phenotype information for mouse models and the third orthology information (required to use OMIM's gene–disease associations).

4.2.1.2 *Human-specific data*

OMIM's gene–disease associations are maintained in a file called MorbidMap. The MorbidMap file was downloaded on 1 March 2011, available via the database's download services. MorbidMap contains the information about known associations of human diseases and genes. In the version we used, 2,717 diseases were linked to 2,266 genes with 3,463 distinct gene–disease associations (on average 1.27 genes per disease).

OMIM's MorbidMap contains gene–disease associations based on human genes. To incorporate the information from OMIM's MorbidMap, we mapped diseases via orthologous genes and mouse models possessing allelic variants in those genes to mouse models (see 3.2.6.1).

4.2.1.3 *Evaluation set*

Building mouse-specific disease profiles based on gene–disease associations does not provide an insight into the validity of those profiles. Therefore, the validity of the mouse-specific disease profiles was assessed by applying those profiles in a disease gene prioritisation task. The prioritisation approach introduced in chapter 3 was used for this purpose.

Due to the profiles being built from MGD's and OMIM's gene–disease association, the data sets applied for evaluation in chapter 3.2.6 could not be used here. We used OrphaNet and Kyoto Encyclopedia of Genes and Genomes (KEGG) instead. We combined all available gene–disease associations from both data resources and removed those contained in OMIM and MGD. The resulting evaluation set covered 1,626 diseases and 7,027 model–disease associations which are neither contained in MGD nor OMIM.

4.2.2 *Building two sets of mouse-specific disease profiles*

Both the mouse-specific profiles **MMM** and **MOM** (see 4.2) were built based on mouse model–disease associations: one set of mouse-specific disease profiles based on MGD's (**MMM**) and one set of profiles based on OMIM's MorbidMap which was transferred to a mouse representation before (**MOM**) (see 4.2.1).

To build disease profiles from the lists of model–disease associations, we employed the phenotype data gathered in mouse model experiments (provided in MGD's MGI_GenePheno.rpt report file; see

4.2.1.1). Experimentally examined mouse models possess recorded phenotypes represented with MP annotations. A disease profile is then built by accumulating all the available phenotype information of all mouse models associated to this particular disease.

Figure 13 depicts how the mouse-specific phenotype profile is built for Beckwith-Wiedemann syndrome (BWS) (MIM:#130650) based on MGD's disease–model associations.

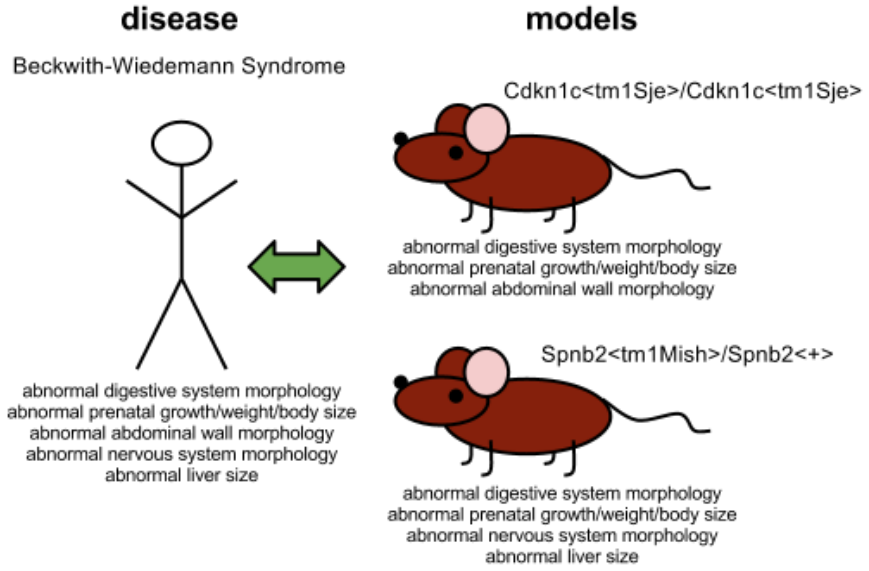


Figure 13: **Illustration of disease profile creation using known gene–disease associations.** Each disease profile is built based on accumulation of phenotype information associated with mouse models. Beckwith-Wiedemann syndrome (BWS) (MIM:#130650) is associated with models Cdkn1c<tm1Sje>/Cdkn1c<tm1Sje> (MGI:1861811) and Spnb2<tm1Mish>/Spnb2<+> (MGI:2450327) (only shown with subset of annotations here). BWS obtains not only the shared annotations, e.g. *abnormal digestive system morphology* (MP:0000462), but also annotations which are only related to one of the mouse models, e.g. *abnormal liver size* (MP:0004848) used to annotate Spnb2<tm1Mish>/Spnb2<+> (MGI:2450327).

In MGD, BWS is associated with two mouse models:

1. *Cdkn1c<tm1Sje>/Cdkn1c<tm1Sje>* (MGI:1861811) and
2. *Spnb2<tm1Mish>/Spnb2<+>* (MGI:2450327).

Mouse model Cdkn1c<tm1Sje>/Cdkn1c<tm1Sje> (MGI:1861811) is e.g. annotated with (among other phenotype data which is suppressed here):

- *abnormal digestive system morphology* (MP:0000462)
- *abnormal prenatal growth/weight/body size* (MP:0004196) and

- *abnormal wall morphology* (MP:0004196)

Mouse model Spnb2<tm1Mish>/Spnb2<+> (MGI:2450327) is e.g. annotated with (among other phenotypes suppressed here due to complexity)

- *abnormal digestive system morphology* (MP:0000462)
- *abnormal prenatal growth/weight/body size* (MP:0004196)
- *abnormal nervous system morphology* (MP:0003632) and
- *abnormal liver size* (MP:0004848)

The accumulation step then gathers phenotype information from the assigned mouse models and distinguishes three types of phenotype information:

1. required information: phenotypes common to all mouse models (corresponds to *abnormal digestive system morphology* (MP:0000462) and *abnormal prenatal growth/weight/body size* (MP:0004196) in figure 13)
2. optional information: all phenotypes from all mouse models even if not present in all mouse models
   (corresponds to all annotations shown for the disease in figure 13) and
3. frequency information: by counting phenotype occurrences across the associated mouse models
   (also corresponds to the annotations shown in figure 13 for the disease, but in addition a frequency score is maintained for each annotation, e.g. 1 for *abnormal digestive system morphology* (MP:0000462) – because both mouse models are annotated with this concept – and 0.5 for *abnormal liver size* (MP:0004848) – as only one mouse model is annotated with it).

The reason for distinguishing three different types of phenotype information is that Oti et al. (2009) showed that employing the *frequency* of phenotypes can improve results of genetic analyses based on phenotype data. Moreover, discussions about what types of phenotype data to in-/exclude in a disease profile are still ongoing.

The three different types of phenotype data for each disease are maintained in independent files, even though they are interrelated

and can be transferred from one to the other: *optional phenotypes* are all phenotypes also maintained in *frequency phenotypes*, just without frequency scores; *required phenotypes* are all phenotypes possessing a score of 1 among the *frequency phenotypes*. The only reason to keep those three types separate is an easier handling in further analyses and therefore avoiding a filtering step each time the different types of phenotype data are required to be distinguished. This consequently leads to six annotation files:

1. MMM_required
2. MMM_optional
3. MMM_frequency
4. MOM_required
5. MOM_optional and
6. MOM_frequency.

4.2.2.1 *Comparison of both sets of mouse profiles*

Both sets of mouse profiles (**MMM** and **MOM**) were gathered from two independent data resources, which means that both the obtained sets of profiles may substantially differ from each other. Even though the databases communicate about their content, there is no guarantee that data are propagated from one to the other. Furthermore, MGD reports mouse-specific model–disease associations while OMIM holds human-specific gene–disease associations. Therefore, we compared both obtained sets of profiles which each other. This comparison will enable us to assess potential differences in the mouse profiles due to the underlying data.

To compare the mouse-specific disease profiles MMM and MOM with each other, we compared each type of phenotype information (required, optional, frequency; see 4.2.2) individually, e.g. MMM_required and MOM_required. The files were compared using a semantic similarity measure (see 1.3.2). Even though the Jaccard index (applied for gene prioritisation before; see 3.2.5) would be sufficient for the binary representation of *required* and *optional phenotypes*, a semantic similarity was required, allowing to take *frequency* information of phenotypes into account. We therefore decided to base the comparison on the cosine similarity measure

$$\begin{aligned} sim(A, B) &= \cos(\theta) \\ &= \frac{A \cdot B}{\|A\|\|B\|} \\ &= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \end{aligned} \tag{4.1}$$

where A and B are vectors of numbers, in our case in the range of [0,1]. Each number corresponds to the probability of a phenotype occurring together with given a disease (determined based on the number of occurrences in mouse models). Cosine similarity reduces to a Jaccard index when applied to binary vectors. The cosine similarity measure delivers scores in the range of [0,1]. A perfect agreement in terms of similarity is symbolised by a score of 1 and a complete disagreement is indicated by a score of 0.

Similarity scores were accumulated for all diseases which can be compared across the two profiles. Due to differences in the input data, not every disease in MOM is also represented in MMM and *vice versa*. Consequently, similarity scores could only be calculated in the case of diseases having profiles in both MMM and MOM. Similarity scores were collected for all diseases in both representations and normalised for the number of data points.

4.2.3 *Application to gene prioritisation*

Due to the derivation of mouse-specific disease profiles, it is unclear how well those profiles characterise human diseases. Furthermore, no assumption is possible whether the generated profiles are an helpful extension to the existing phenotype description in databases such as OMIM. Therefore, we applied the obtained mouse-specific disease profiles in a gene prioritisation task. Each type of phenotype information (required, optional, frequency; see 4.2.2) was assessed individually for both the representations MMM and MOM.

We adapted the approach introduced in chapter 3 and instead of using OMIM's phenotype descriptions (see 3.2.2), we applied the mouse-specific disease profiles. Mouse models were again ranked per individual disease possessing a profile, and the obtained results were recorded for further evaluation.

Similar to evaluating the prioritisation results in chapter 3 (see 3.2.6), we used receiver operating characteristic (ROC) curves (see

1.4.2.2) to assess the performance of the prioritisation results. The prioritisation results were evaluated against the reduced set of the combined OrphaNet and KEGG gene–disease association (see 4.2.1.3). This procedure resulted in six ROC curves (one for each the phenotype annotation files; see 4.2.2) and their corresponding area under curve (AUC) score.

4.2.4 *Classification of diseases based on phenotype information*

We further assessed the generated mouse-specific disease profiles for their potential to extend annotations of existing clinical disease descriptions, such as those provided by OMIM. Therefore, we applied the MMM profiles (due to this being the smaller data set and our results showing that both mouse-specific profiles are similar; see 4.3.1.1) to a disease classification task.

To achieve a disease classification, the generated MMM disease phenotype profiles were integrated with MP into an Web Ontology Language (OWL) (see 1.2.1) representation. We incorporated the *required* and *optional phenotypes* maintained in MMM_required and MMM_optional (see 4.2.2). A disease was then defined as the equivalent to the intersection of the phenotypes. An illustration of the disease classification is shown in figure 14.

The resulting OWL-based ontology that incorporates MP and the disease profiles, was classified using the ELK reasoner (Kazakov et al., 2012). If one disease becomes a subclass of another, then all the phenotypes associated with the superclass are also associated with the subclass. Consequently, for every disease, the *optional* representation becomes a subclass of the *required* disease definition. If two diseases are annotated with the same phenotypes, they become equivalent classes.

Although the ontology is limited by the phenotypes that have been recorded from the various experimental studies, we manually investigated the results for biological meaningful classifications. Diseases classify according to their phenotype information and are assorted according to the taxonomy of MP, i.e. appear as subclasses of MP concepts as well as other diseases. Therefore, we further investigated the obtained classification for its potential to enrich existing human-specific disease descriptions. We hand-selected two examples for which we obtained evidence from scientific publications.

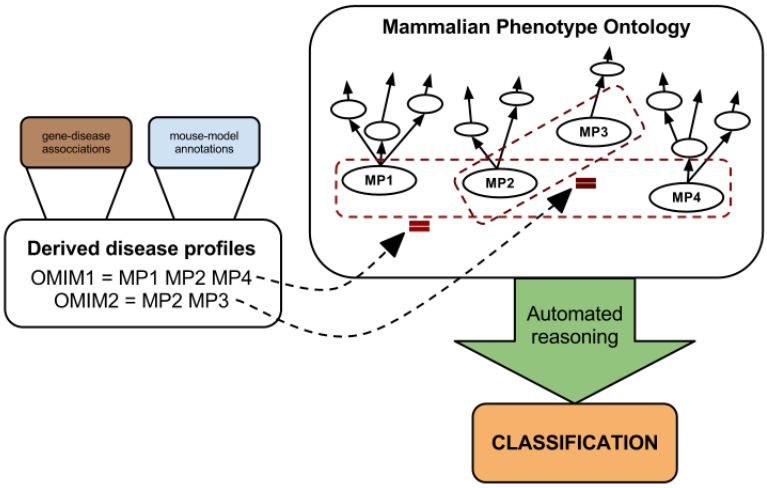


Figure 14: **Illustration of the disease classification process.** Mouse-specific disease profiles are integrated with MP using equivalence class statements. The disease profile corresponds to the conjunction of all the describing phenotypes. The enriched ontology is classified and a hierarchy of diseases is obtained based on their phenotype descriptions.

4.3 RESULTS

4.3.1 *Mouse-specific profiles for human genetic disorders*

Applying the described algorithm (see 4.2.2), we obtained two independent disease profiles: one based on MGD's (further referred to as **MMM**) and one based on OMIM's gene–disease associations (further referred to as **MOM**). Based on the known OMIM and MGD gene–disease associations, 2,460 MOM and 959 MMM MP-based disease profiles were generated. The resulting disease representations share 805 diseases, but the obtained profiles may still differ depending on the phenotypes of the mouse models associated with a disease (see 4.3.1.1).

Each disease profile can be characterised by the number of its phenotype annotations and can be further categorised in three phenotype groups: *required*, *optional* and *frequency* (see 4.2.2). An overview of the content of the generated profiles is provided in table 4.

Employing OMIM's gene–disease associations (MOM profiles), we obtain more diseases being described with mouse-specific profiles. The number of annotations is generally higher in the MOM profiles than in MMM profiles. However, our results suggest that the MMM

profiles	dis*	number of annotations min†	max†	avg†	frequency ranges min‡	max‡	avg‡
MMM	959	4 (1)	871 (355)	93 (40)	0.023	1.0	0.624
MOM	2460	3 (1)	1224 (190)	115 (13)	0.004	1.0	0.306

Table 4: **OMIM's gene–disease annotations lead to a higher coverage in number of diseases represented and number of assigned annotations.** First row: characteristics MMM profiles; second row: characteristics MOM profiles. * number diseases with profile; † *min*imum, *max*imum, *av*rage number of optional phenotypes, required phenotypes in parentheses ; ‡ *min*imum, *max*imum, *av*rage frequency of phenotypes.

profiles are more consistent than the MOM profiles. This is indicated by the average number of *required* phenotypes being higher in MMM profiles (see column three to five in table 4). Further to that, the higher consistency in MMM profiles is also supported by the range and average of the frequency values (see last three columns in table 4).

4.3.1.1 *Comparison of obtained mouse-specific disease profiles*

To identify whether substantial profile differences occur depending on whether human- or mouse-specific gene–disease associations are used, we compared both mouse profiles (see 4.2.2.1). The comparison shows that both the derived mouse-specific disease profiles are mostly in agreement (see 15). Some difference are more evident in *required* phenotype information than in the *optional* or *frequency* phenotype representation.

Differences in both the generated mouse profiles may occur, if a disease is associated with different genes in either human or mouse. For example, Barth Syndrome (BTHS) (MIM:#302060) is associated in mouse with the genes *Fkbp1a* (MGI:95541) and *Mest* (MGI:96968), while in human it is associated with the *TAZ* gene (UniProt:Q96F92). Phenotypes associated with *Fkbp1a* and *Mest* include *decreased cardiac muscle contractility* (MP:0005140), *decreased body weight/size* (MP:0001262, MP:0001265) and *abnormal mitochondrion morphology* (MP:0006035) which overlap with the phenotypes provided in OMIM[2]. However, phenotype annotations for the *TAZ* gene have an emphasis on kidney related phenotypes, such as *kidney cysts* (MP:0003675) or *dilated renal tubules* (MP:0002705). In this particular case, the profiles derived from

2 Phenotypes in OMIM: (Dilated) Cardiomyopathy, Growth retardation, Ultrastructural abnormalities in mitochondria on electron microscopy; `http://omim.org/clinicalSynopsis/302060`

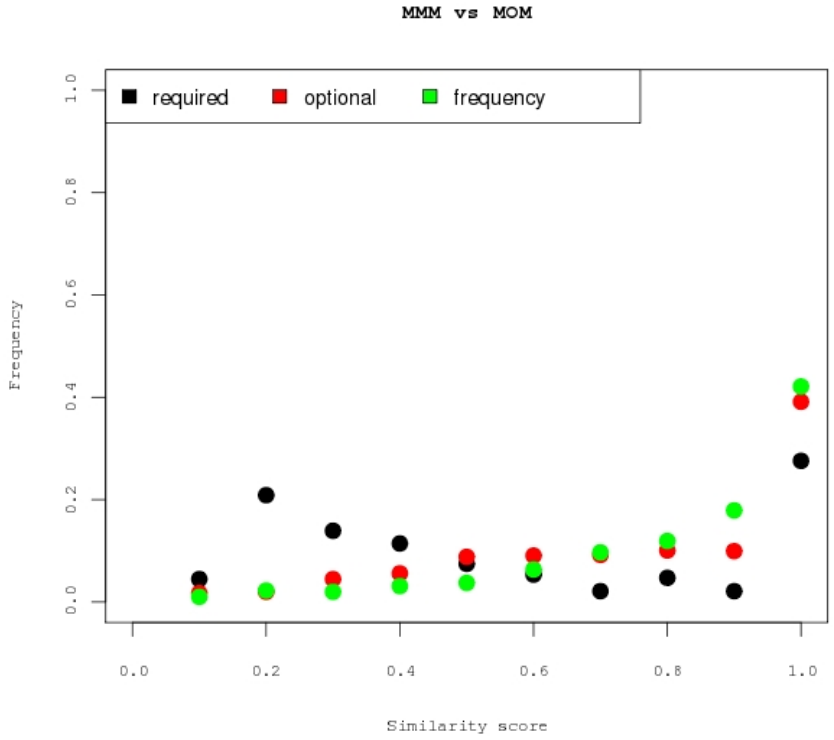


Figure 15: **Both generated mouse profiles or mostly in agreement.** Comparison of both mouse-specific disease profiles based on cosine similarity and separate for each type of phenotype information (*required*, *optional*, *frequency*).

the model–disease associations contained in MGD show a higher similarity with the defined signs and symptoms of the disease. We note here that there might be future tests on the *TAZ* gene that may reveal similar phenotypes to BTHS's signs and symptoms.

A high similarity of profiles occurs if a disease is associated with genes exhibiting similar phenotypes in mouse and human. In most cases, this happens when a disease is associated with the same gene in human and mouse. For example, Spondylocheirodysplasia, Ehlers-Danlos syndrome-like (SCD-EDS)(MIM:#612350) is associated with mutations in the *Gnat2* gene in mouse (MGI:95779) as well as the *GNAT2* gene in human (UniProt:P19087). The MP annotations in both generated mouse-specific profiles include the following phenotypes: *postnatal growth retardation* (MP:0001732), *malocclusion* (MP:0000120), *decreased bone mineral density* (MP:0000063), *abnormal dermal layer morphology* (MP:0001243), and *corneal thinning* (MP:0005543). Both the

mouse-based profiles show a high similarity with the phenotypes for SCD-EDS described in OMIM[3].

An agreement of both profiles is, however, no indicator whether the obtained mouse-specific disease profiles are a good or a bad representation of signs and symptoms of a disease. As mutagenesis experiments are ongoing, it may happen that the phenotype scans in mice are incomplete. Thus, the application of mouse phenotypes would lead to an incomplete phenotype representation of a disease. For example, the mouse-specific profiles for Alport syndrome, X-linked (ATS) show phenotypes related to kidney impairment, e.g. *abnormal renal glomerulus morphology* (MP:0005325) and *abnormal urine homeostasis* (MP:0009643), but do not mention any phenotypes related to eyes or hearing impairment, while OMIM mentions these phenotypes for ATS[4].

Differences to the existing human phenotype profiles may either be additional, missing or species-specific information. If differences constitute additional annotations, then they can be indicators for tests which have not been associated with a disease so far. If annotations are lacking, this may be an indication for the incompleteness of MGD or a sign of species diversity. Some of the annotations used are also particular to the phenotype resource. For example, OMIM provides annotations concerning the onset of a disease, e.g. *Juvenile onset* (HP:0003621) while MGD provides information respective to litter size, e.g. *decreased litter size* (MP:0001935). Consequently, if profiles are needed that support finding mouse genes, great care is required to build profiles which accommodate for species-specific aspects as well as representing all the relevant signs and symptoms of a disease. In conclusion, one has to be cautious as to what kind of gene–disease associations are used to profile diseases as this may lead to different results in further studies.

4.3.2 *Performance of mouse phenotype profiles in gene prioritisation*

To be able to judge on the validity and potential of the derived phenotype mouse profiles, both MMM and MOM annotation files were applied in a gene prioritisation task (see 4.2.3). The obtained results were gathered and evaluated according to known gene–disease associations from OrphaNet and KEGG and the AUC score of the corresponding ROC curve calculated.

3 http://omim.org/clinicalSynopsis/612350
4 http://omim.org/clinicalSynopsis/301050

The best AUC score is obtained when OMIM's gene–disease associations are employed including *frequency* information about the occurrences of phenotypes for each disease. In general, for both disease profiles it accounts that the more information is included, the better the performance: *frequency* is best, followed by *optional* and *required*.

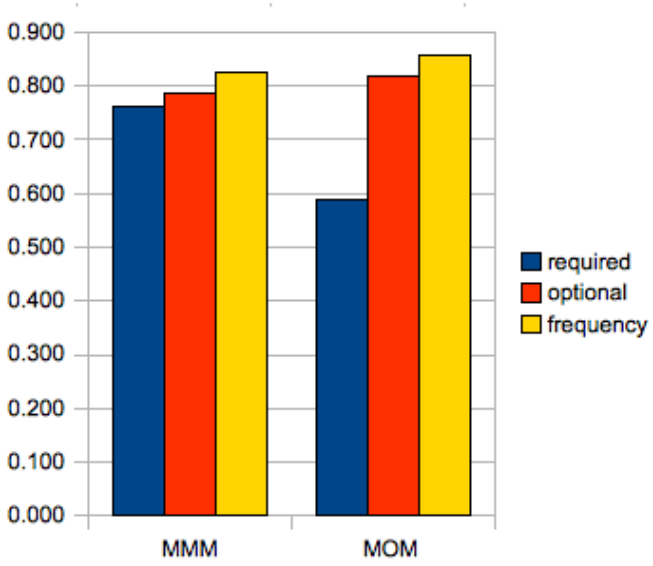


Figure 16: **Employing OMIM's gene–disease associations leads to better performance when employing all phenotype information.** Performances of both derived mouse profiles in gene prioritisation task when evaluated according to known gene–disease associations in OrphaNet and KEGG.

4.3.3 *Application of mouse profiles to disease classification*

After assessing the validity of the of the disease profiles and their performances in the gene prioritisation task (see 4.3.2), the profiles were applied to assess their potential for annotating diseases. Given the similarity of both the derived profiles, only the MMM disease profiles were employed as they are smaller in numbers (see 4). Diseases were classified based on their MMM profiles and a small number of interesting cases was manually verified by a senior biological expert.

Our manual investigation supports that the forming classification represents biological relatedness. For example, based on the mouse-specific experimental observations, diseases that affect the cone electrophysiology, such as Retinitis Pigmentosa (MIM:#611131) and enhanced S-core syndrome (MIM:#268100), are classified together.

Further to that, the results also support that the mouse-specific profiles can indeed be used to support curators and suggest phenotypes for inclusion into existing databases reporting phenotype information for human diseases, e.g. OMIM. For instance, investigating diseases classified with the mouse-specific observations related to *abnormal sleeping patterns* (MP:0001501), several diseases are retrieved with known associations to such patterns. One example classified with *abnormal sleeping patterns* (MP:0001501) is *Angelman syndrome* (MIM:#105830), which is also annotated in OMIM with *Abnormal sleep-wake cycle*.

Among those disease is also a number of diseases such as *schizophrenia* (MIM:#181500) and *Lesch-Nyhan syndrome* (MIM:#300322) which, currently, have no such association in the OMIM clinical synopsis or in the HPO-based annotations. The association of *abnormal sleeping patterns* (MP:0001501) to such diseases, however, has already been made at a clinical level (Cohrs, 2008; Saito et al., 1998).

Another example would be that neither the OMIM clinical synopsis nor the HPO-based annotations provide a relation between *excessive scratching* (MP:0001412) and Prader-Willi syndrome (PWS) (MIM:#176270). Our mouse disease representation reveals a link between them. *Excessive scratching* (MP:0001412) is a known characteristic of PWS patients resulting from the high pain threshold of those patients. It may be an indicator of the patient's stress levels and as such should be monitored (Buono et al., 2010; Jacob et al., 2009).

It is noted here, that in addition to those examples, a quantified evaluation of the enrichment potential will be necessary and also an assessments of the resulting disease classification as such.

4.3.4 *PhenDis – a web interface to browse disease profiles*

Further to providing the MMM and MOM mouse-specific disease profiles for download, they are also available through a simple web interface (PhenDis)[5]. PhenDis allows to search for diseases and shows phenotypes associated with each disease in MOM and MMM. In addition to the mouse-specific disease profiles, it also shows the human specific annotations contained in OMIM (see 3.2.2). Figure 17 shows a screen shot of the web interface.

5 http://phenomebrowser.net/PhenDis

PhenDis

disease ID: search

Disease List

MIM_ID	MMM (MP)	Similarity score MMM vs OH (HP-based)	MOM (MP)	Similarity score MOM vs OH (HP-based)	OH (HP)
277460	mammalian phenotype abnormal head movements tremors paresis abnormal skeletal muscle morphology abnormal spinal cord morphology abnormal retina morphology retinal degeneration disorganized retinal layers abnormal stationary movement abnormal locomotor behavior ataxia impaired coordination stereotypic behavior abnormal motor coordination/ balance abnormal embryogenesis/ development abnormal placenta morphology decreased trophoblast giant cell number placental labyrinth hypoplasia abnormal placenta labyrinth morphology abnormal homeostasis abnormal somatosensory cortex physiology abnormal reproductive system physiology infertility female infertility abnormal motor capabilities/coordination/movement abnormal extraembryonic tissue morphology abnormal eye morphology abnormal muscle physiology abnormal muscle morphology abnormal lipid homeostasis abnormal fertility/fecundity neurodegeneration muscular atrophy head shaking	0.22440969726240098	mammalian phenotype abnormal head movements tremors paresis abnormal skeletal muscle morphology abnormal spinal cord morphology abnormal retina morphology retinal degeneration disorganized retinal layers abnormal stationary movement abnormal locomotor behavior ataxia impaired coordination stereotypic behavior abnormal motor coordination/ balance abnormal embryogenesis/ development abnormal placenta morphology decreased trophoblast giant cell number placental labyrinth hypoplasia abnormal placenta labyrinth morphology abnormal homeostasis abnormal somatosensory cortex physiology abnormal reproductive system physiology infertility female infertility abnormal motor capabilities/coordination/movement abnormal extraembryonic tissue morphology abnormal eye morphology abnormal muscle physiology abnormal muscle morphology abnormal lipid homeostasis abnormal fertility/fecundity neurodegeneration muscular atrophy head shaking abnormal ocular fundus morphology	0.22440969727312549	All Mode of inheritance Autosomal recessive inheritance Phenotypic abnormality Abnormality of head and neck Abnormality of the head Abnormality of the face Abnormality of the ocular region Abnormality of the eyelid Abnormality of the periorbital region Abnormality of the nervous system Abnormality of the skin Xanthomatosis Xanthelasma Ataxia Areflexia Reduced reflexes Abnormality of the cerebellum Abnormality of the integument Abnormality of metabolism/homeostasis Abnormality of the central nervous system Incoordination Cysts of the eyelid Nodular changes affecting the eyelids Abnormality of movement skin papules skin nodules Skin cysts

Figure 17: **PhenDis web interface.** PhenDis web interface to enable browsing through obtained mouse-specific disease profiles.

4.4 DISCUSSION

In this study, two independent mouse-specific phenotype descriptions for human genetic disease were derived from gene–disease associations available in MGD and OMIM. Both mouse-specific disease profiles combined allow for a representation of 2,614 diseases but the integration of profiles from MMM and MOM is still to be investigated.

In total, the number of disease being represented with mouse-profiles (2,614 diseases) is less than the number of diseases described in the OMIM annotation file (5027 diseases as of March 2011; see 3.2.2). The overlap between the existing human-specific annotations and the derived mouse-specific profiles is 1,792 diseases. This means that there are at least 822 diseases which are only described with mouse-specific profiles. Those 822 disease profiles may be used to serve as the basis for annotating these diseases in human.

When applied to the gene prioritisation task, disease descriptions derived from OMIM (MOM profiles) perform better than the MGD-derived descriptions (MMM profiles) for *optional* and *frequency* information but not *required* phenotypes. This suggests that for the identification of necessary signs and symptoms of a disease, the MMM disease profiles are better suited. The high performance of either set of profiles suggests that both derived profiles possess valuable information that can be used to describe human diseases on a phenotype level. Integration of both profiles is, however, subject to further studies.

Applying the mouse-specific disease profiles in a gene-prioritisation task leads to an AUC score of 0.857. This AUC score shows that the obtained profiles are suited to prioritise disease gene candidates; thus, the generated profiles are valid. With showing the validity of the profiles, the derived disease descriptions constitute good candidates for inclusion into reference databases, upon evaluation. Furthermore, the obtained results suggest that the frequency information derived from the occurrences of allelic variants per disease can be used to estimate frequency attributes for phenotypes in human. However, it is not clear yet how those frequency counts can be transferred across species.

To date, no investigations have happened as to whether those profiles can help to elucidate species-specificity. Both the disease profiles MOM and MMM share 688 diseases with diseases being annotated with HPO annotations in OMIM. A systematic investigation of those 688 diseases and their assigned phenotype profiles may lead to further

insights about the differences between mouse and man. However, this is also subject to further investigations.

4.5 CONCLUSIONS

Mouse-specific disease profiles were derived using the gene–disease associations contained in MGD and OMIM and phenotype descriptions assigned to mouse models. While both the derived mouse-specific disease profiles mostly agree, MOM disease profiles possess a higher performance when applied in a gene prioritisation task. The application to gene prioritisation also supports previous findings that including *frequency* information of phenotypes leads to better performances in biological use cases (Oti et al., 2009). Applying the profiles to classify diseases and manually assessing the obtained classification, it could be demonstrated that the mouse-specific disease profiles harbour the potential to enrich existing human-specific annotations of those diseases.

4.6 FUTURE WORK

A first step in future research will be the integration of both obtained mouse-specific disease profiles (MMM and MOM). Therefore, a more systematic analysis is required to identify when both the profiles differ and how the differences can be overcome.

Further to that, the mouse-specific disease profiles may provide insights into species-specificity of diseases. 688 diseases are shared between between MMM, MOM and the OMIM diseases annotated with HPO (see 3.2.2). Systematically investigating a subset of these diseases and manually assessing the assigned phenotypes, alongside with the distribution of the phenotypes, could potentially elucidate cases, were orthologous genes do not exhibit similar phenotypes and lead to species-specificity.

While the classification has so far only been manually explored, a systematic analysis may yield a number of annotations which could be provided to databases for potential inclusion. In order to achieve that, the results need to be quantified in a more systematic and automated evaluation than before.

5

MINING PHENOTYPES FROM SCIENTIFIC LITERATURE

In chapter 3, I described a method to prioritise disease gene candidates based on phenotype information. In the previous chapter, I reported about my data mining efforts to further improve the definition and description of signs and symptoms of human diseases. In this chapter of my thesis, I report about my efforts to text mine phenotype information from scientific literature using the Mammalian Phenotype Ontology (MP) (Smith et al., 2005). My aim was to enrich the phenotype descriptions of mouse model experiments contained in the Mouse Genome Informatics database (MGD) (Blake et al., 2011).

5.1 BACKGROUND

Phenotype studies, such as the gene–disease prioritisation method shown in chapter 3, PhenomeNET (Hoehndorf et al., 2011c) or MouseFinder (Chen et al., 2012), rely on the availability of phenotype annotations. If the amount and quality of those annotations changes, the outcomes of those studies change as well. Analysing the biological coherence of human phenome databases, Oti et al. (2009) discovered that phenome databases do not cover sufficient information and require further improvements in terms of coverage. Thus, automated methods are sought, enabling access to more phenotype information.

Due to most results of biological experiments being published in academic journals, the scientific literature constitutes a tremendous resource for phenotype information. Curation is one possibility to identify pieces of knowledge contained in scientific literature and populate databases with it. For example, both databases MGD and the OMIM database (Amberger et al., 2011) are curated from scientific literature. Despite the accuracy curation provides, the process is slow and costly (Jaeger et al., 2008).

In addition to curation, text mining (TM) has been successfully applied to enrich and populate databases (Krallinger et al., 2008). However, in the area of TM, little attention has been paid to the extrac-

tion and normalisation (see 1.4) of disease related entities other than genes and their products (Jimeno et al., 2008). Even though it has been shown that text mined phenotype data can be used for further biological analysis (Korbel et al., 2005; van Driel et al., 2006), phenotypes were not extracted systematically and provided as annotation sets.

Despite extraction systems targeted towards the extraction of one particular type of entity, more generalised solutions exist. Those generalised solutions annotate free text with diverse terminologies and ontologies, e.g. Open Biomedical Annotator (OBA) (Jonquet et al., 2009) or Whatizit (Rebholz-Schuhmann et al., 2008). Even though these tools can be also applied to the extraction of phenotype data, they target a broad coverage and thus possess a low performance.

Existing TM tools have been tested to be integrated into the work flow of the curation work required to populate MGD (Dowell et al., 2009). Dowell et al. (2009) found that the investigated tools for phenotype extraction, among which OBA was analysed, did not show satisfying results. Furthermore, in a survey about the suitability of available TM tools for curation, a requirement for reliable phenotype extraction tools was identified (Hirschman et al., 2012).

In this chapter, two methods (Phentomine$_{LEX}$ and Phentomine$_{EQ}$) are described, which aim at the extraction of MP phenotypes to annotate mouse models, such as those described in MGD. Both methods have been implemented and tested separately so far. While the first method uses an information theory approach, the second employs the manually assigned EQ statements for MP concepts (see 2.2.2.2).

Our first method (*Phentomine*$_{LEX}$) was inspired from the GO tagger described in (Gaudan et al., 2008) which has later been adapted to the recognition of disease names (Jimeno et al., 2008). Applying this approach to the recognition of disease names in a test corpus led to a competitive recall of 76.28% (see 1.4.2.1). This method uses solely textual features, such as the words contained in the ontology's concept labels and synonyms together with their frequency.

In our second method (*Phentomine*$_{EQ}$), the manually assigned EQ statements (see 2.2.2.2) of MP concepts are employed to extract phenotype information from free text. In an EQ statement, a phenotype is broken down into an *entity* and a *quality*. Thus, the challenge of identifying phenotypes in scientific literature is altered to the identification of its constituents (*entity* and *quality*) in free text.

By interrogating curators about their experiences with TM solutions, Alex et al. (2008) found that TM solutions help to speed up the curation

process even if felt otherwise by curators. Furthermore, Alex et al. also showed that a high recall of a method is perceived to be better than a high precision (see 1.4.2.1). A high recall means that there is a likelihood to retrieve a lot of false positive results but the chances of missing information is less likely. We therefore aim on developing a phenotype extraction solution which possesses a high recall, even though this could potentially lead to a reduction in precision.

Due to linguistic differences between abstracts and full text papers and the difference in content covered (Cohen et al., 2010), we investigated both abstracts from MEDLINE and full text papers from PubMed Central (PMC) (Open Access subset)[1] (see 5.2.1.1) to provide phenotype annotations. The occurrences of phenotype information across the different sections of a scientific paper were recorded to identify the sections with the highest density of phenotype features. This will allow to either filter extracted phenotypes or assign evidence scores in later stages.

Additional information, source code and results are available online[2]. For more details refer to page xi.

5.2 METHODS AND MATERIALS

5.2.1 *Input data*

Both phenotype extraction methods use textual data from ontologies. We refer to the concept labels and synonyms as terms. The ontologies used in this part of the study are OBO Foundry ontologies[3] (Smith et al., 2007) and have been developed and maintained in the OBO Flatfile format (see 1.2). The OBO Foundry website does not only support the OBO Flatfile format but also provides alternative formats[4]. In this study, we used the .tbl files since these can be easily processed and integrated.

Even though the OBO Flatfile distinguishes between different types of synonyms[5], this distinction is not maintained in the .tbl file. In an first attempt, we used all the data provided by the .tbl file. In later stages, it may be required to further filter the synonyms depending on

1 http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/
2 https://code.google.com/p/phentomine/
3 http://www.obofoundry.org/
4 http://www.berkeleybop.org/ontologies/
5 http://www.geneontology.org/GO.format.obo-1_2.shtml#S.2.2.2

their scope to increase the performance of the phenotype extraction system.

A version of the MP ontology was downloaded as .tbl file on 16 May 2012 from the OBO Foundry web page . The MP .tbl file comprised 9,127 concepts with labels, and 5,701 out of these also possessed synonyms. Textual definitions comprised in this file are, thus far, not taken into consideration but leave room for further improvements of the textual approach in future work.

Our second method works based on EQ statements. Therefore, we further downloaded the corresponding EQ statements for MP concepts. EQ statements were available for 6,575 MP concepts (72%). MP's EQ statements employ a number of other OBO Foundry ontologies, such as Gene Ontology (GO) (Ashburner et al., 2000) or Phenotype And Trait Ontology (PATO) (Gkoutos et al., 2009). As covering all the ontologies employed by EQ statements would be out of scope of an initial study, we limited the phenotypes to structure and process phenotypes, i.e. phenotypes which can be represented with Mouse adult gross Anatomy ontology (MA) (Hayamizu et al., 2005), GO and PATO. Therefore, we also downloaded (all on 16 May 2012) a .tbl file version for of the three ontologies.

All 6,575 concepts available with an assigned EQ statement possess at least one concept of MA, PATO, and GO in their EQ statement. However, only 3,761 MP concepts can be represented using those ontologies solely (the subset we are working with). A small fraction of structural phenotypes could not be covered with this assumption as those are constructed using UBERON (Mungall et al., 2012), a multi-species metazoan anatomy ontology, instead of MA.

5.2.1.1 *PubMed Central corpus*

To assess the occurrence and extract phenotype information, we used the Open Access subset of PMC. This corpus comprises not only abstracts but also the full text of journal articles. The scientific publishers decide which articles from which journals will be made available in PMC. We downloaded the Open Access subset in June 2011 and by then, the corpus comprised over 2 million articles. However, to ensure that the articles we investigate contain phenotype information, we only used those articles that are referenced in MGD's MGI_PhenoGenoMP.rpt report file. The reduced PMC corpus contained 621 documents (see 5.2.3).

5.2.2 *Two approaches to mine phenotypes from text*

To extract phenotype information from scientific literature, two different methods, *Phentomine*$_{\text{LEX}}$ and *Phentomine*$_{\text{EQ}}$, have been implemented. *Phentomine*$_{\text{LEX}}$ focuses on the extraction of phenotype data by employing textual characteristics while *Phentomine*$_{\text{EQ}}$ uses manually defined EQ statements (see 2.2.2.2). Both methods have been developed and evaluated independently, even though we aim to combine them in later stages. It is still to be determined whether and how those two approaches can complement each other. The overall workflow of both methods is illustrated in figure 18.

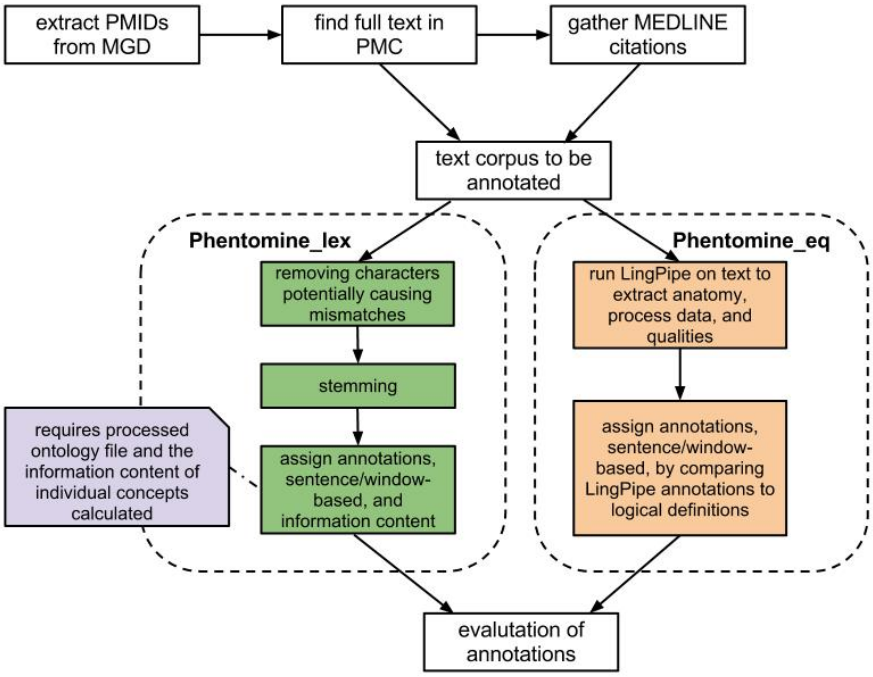


Figure 18: **Overview of the several steps involved in *Phentomine*.** *Phentomine*$_{\text{LEX}}$ illustrated in green; *Phentomine*$_{\text{EQ}}$ coloured in orange. Before annotations can be assigned, pre-processing steps for ontologies are required: *Phentomine*$_{\text{LEX}}$ requires the information content (see 5.2.2.1) calculated for each individual ontology concept; *Phentomine*$_{\text{EQ}}$ requires filtered and processed EQ statements.

5.2.2.1 *Phenotype extraction based on textual features (Phentomine*$_{\text{LEX}}$*)*

In our phenotype extraction method, we focus on a high recall at an acceptable rate of precision (see 1.4.2.1). Gaudan et al. (2008) published a NER system which has been shown to possess a high recall for the recognition of GO concepts and disease names. We amended the

algorithm to not only support phenotypes (currently only MP) but also extended the applied natural language processing (NLP) techniques.

While Gaudan et al. (2008) only uses the words contained in the concepts names of an ontology, we also take the synonyms into consideration. Furthermore, we apply a pre-processing step to remove potentially error-causing characters and words possessing a low information content. In addition, we applied a Porter stemmer (Porter, 1980) before the calculation of the information content of each ontological concept. All steps mentioned here, are explained in more detail in the following (see *Implementation*).

Methodology

All uniquely stemmed words used to build the names and synonyms of the MP concepts (derived from the ontology's .tbl file; see 5.2.1) constitute the corpus. Based on this corpus, the occurrence probability is calculated and consequently the information content. Applying this formula, rarely occurring words in the corpus will have a high information content, while words occurring frequently possess a low information content score. The information content for each individual concept or synonym (*I(t)*) is stored to use it when assigning phenotype annotations to free text.

For each of the concept labels and synonyms (collectively called terms), the information content is calculated as described in (Gaudan et al., 2008). The information content for a term t is the sum of all the individual information content scores for each token (in our case stemmed words) w within the term:

$$I(t) = \sum_{w \in t} I(w). \tag{5.1}$$

The information content is calculated as the negative logarithm of the stemmed word occurrence probability $P(w)$ within a text corpus:

$$I(w) = -\log(P(w)). \tag{5.2}$$

We note here that the corpus used to calculate the information content, is the MP ontology file. This file is specifically targeted to describe mouse phenotypes and hence possesses only words describing those phenotypes. Consequently, the words contained in this file are ex-

pected to be mostly information rich and thus allow for the application of an information content.

To decide whether a phenotype is contained in a text span or not, we used specificity and evidence as suggested by Gaudan et al. (2008). Evidence is calculated as the ratio of the number of words of the text span to be annotated (either in a sentence or sliding window) overlapping with the term, and the number of words used to build the term:

$$evidence = \frac{|(words\ in\ term)\ \cap\ (words\ in\ text\ span)|}{number\ words\ in\ term}, \quad (5.3)$$

Specificity in this context is the ratio of the information content of a text span and a term

$$specificity = \frac{I_{ts}}{I_c}, \quad (5.4)$$

where I_{ts} is the information content of the text span, which can either be a sentence or a subset of words from the sentence (obtained when applying a sliding window).

Implementation

The first step (see figure 18) is to process the ontology's .tbl file. The .tbl file was filtered for special characters and stop words. Special characters, such as "%" or "-", refer to any textual component which potentially causes errors during the extraction of phenotypes. Special characters are neither alphabetic characters nor numbers. A comprehensive list of the applied character filters including their replacements is provided in appendix C.1. In addition to filtering special characters, we also removed a list of stop words which is also illustrated in appendix C.2. Stop words, such as "in" or "the", are considered not to carry any meaningful information. Thus, they can be removed before analysis to reduce noise and potential errors evolving from them.

After filtering special characters and removing stop words from the concept labels and synonyms, the remaining textual content was stemmed using a Porter stemmer (Porter, 1980). We used the Porter stemmer implementation provided by Snowball[6]. Snowball is a Java library providing several different stemmer.

6 http://snowball.tartarus.org/

Based on the stemmed content, the information content for all stemmed words contained in the .tbl file were calculated. We only used the stemmed words contained in labels and synonyms but not the textual definitions. The information content of each stemmed word was recorded before all the text content was processed.

Furthermore, before assigning phenotype annotations to both MEDLINE abstracts and PMC full text papers, the textual content of the scientific publications was pre-processed. The text of abstract or full text was, similar to the pre-procession of the ontology file, filtered for special characters and stop words (see appendix C), and stemmed using a Porter stemmer.

The stemmed and filtered text content was then processed once sentence-based and once based on a sliding window and the *Phentomine*$_{\mathrm{LEX}}$ algorithm applied to it. The information content of either sentence or sliding window was calculated by using equation 5.1. In this equation, t corresponds to the textual representation of the sentence or sliding window and $\mathrm{I}(w)$ is the information content of the stemmed word in the sentence or sliding window. A stemmed word which is part of a sentence or sliding window, but is neither contained in a label nor in any synonym of any concept, possesses an information content of zero.

A phenotype annotation is assigned to a sentence or sliding window, and consequently abstract or full text, if the evidence is greater than 0.5 and specificity is greater than 0.8. Those measures have been found suitable for GO concept extraction (Colgiu, 2012). Phenotype annotations are uniquely recorded for either abstract or full text and stored for evaluation purposes as described in section 5.2.3.

5.2.2.2 *Phenotype extraction based on EQ statements (Phentomine*$_{\mathrm{EQ}}$*)*

Methodology

Besides identifying phenotype concepts in literature by applying an information theory approach, our second method (*Phentomine*$_{\mathrm{EQ}}$) utilises ontological features. More specifically, we applied the manually defined EQ statements (Mungall et al., 2010) provided for a subset of MP concepts (see 2.2.2.2). In an EQ statement, phenotypes are broken down into an *entity* and an *quality* constituent. Thus, the quest of extracting phenotype information is altered to the identification of the constituents of phenotypes, assuming that the constituents can be

identified with a high F-measure (see 1.4.2.1). Phenotype annotations are then assigned based on the recognised constituents in a sentence.

For instance, the MP concept *muscle degeneration* (MP:0000749) possesses the following EQ statement (following the OBO Flatfile format; see 1.2.1):

```
[Term]
id: MP:0000749 ! muscle degeneration
intersection_of: PATO:0000639 ! degenerate
intersection_of: inheres_in MA:0000015 ! muscle
```

In this example, the MA term *muscle* (MA:0000015) corresponds to the *entity* constituent and the PATO term *degenerate* (PATO:0000639) corresponds to the *quality* constituent of this phenotype concept.

EQ statements for MP concepts are composed employing several other OBO Foundry ontologies such as the MA or PATO. To reduce the number of ontologies to a manageable subset for a prototype, we decided to cover three of the ontologies used for composition: MA, PATO and GO. We consider the resulting phenotypes to correspond to structural and process phenotypes (see 5.2.1).

Implementation

Even though OBA was used to establish a baseline for phenotype extraction from scientific literature (see 5.2.3.3) and could have also been used for the extraction of the constituents, we found the annotation process handling restrictive. Instead of OBA, we used the LingPipe library[7] (Carpenter, 2007) to build annotation servers for the constituents of phenotypes. LingPipe had previously been tested within our research group for fast and reliable annotations (results not shown here).

For each of the ontologies required to represent structure and process phenotypes in EQ statements (MA, GO and PATO), an individual server was set up. Those servers annotate text spans with ontology concepts and therefore identify the EQ constituents. Underlying those servers are annotation dictionaries which reference text spans to be annotated to ontology concepts used for annotation. Each dictionary was created from the concept labels and synonyms contained in the .tbl files.

Both MEDLINE abstract and PMC full text papers were assessed with those annotation servers and all MA, GO, and PATO annotations were

7 http://alias-i.com/lingpipe/

recorded on a sentence base. Phenotype annotations were assigned by going through the sentence annotations and matching them to the manually assigned EQ statement of an MP concept. We chose two different set-ups to assign the phenotype annotations to sentences:

1. either all constituents which are contained in the manually assigned EQ statement had to be present in the sentence or
2. only 80% of the constituents.

As in the case of *Phentomine*$_{\text{LEX}}$ (see 5.2.2.1), MP phenotype annotations were uniquely recorded for either abstract or full text and stored for evaluation purposes (see 5.2.3).

5.2.3 *Evaluation of the extracted phenotypes*

A common procedure to evaluate NER systems is the application of a text corpus and the calculation of performance measures, such as precision, recall and F-measure (see 1.4.2.1). In the past decades, the main focus of text mining challenges (competitions for TM tools) was on gene, gene product mention and their normalisation. As a consequence, no corpus exist which covers phenotypes, and in particular phenotypes represented in MP.

As the development of such a corpus would be very labour-intense and costly, we searched for alternative options and decided to use the phenotype information available in MGD's MGI_PhenoGenoMP.rpt report file[8]. The MP phenotype annotations provided by this file are not only assigned to a mouse model but also referenced with a PubMed Identifier (PMID). The PMID indicates which publication has caused the mouse model to be annotated with this MP annotation.

To use these annotations, we downloaded the report file in July 2011. First, we extracted all contained PMIDs but then limited the PMIDs to those possessing a full text in PMC, as performance was assessed for both abstract and full text (see 5.1). This subset comprised 621 scientific publications. We then went through the annotation file again and collected referenced MP annotations for each of the 621 documents. Those 621 documents together with their recorded MP annotations extracted from the report file, constitute our reference corpus (further referred to as *gold standard*) containing 7,353 (unique 2,390) MP annotations.

8 ftp://ftp.informatics.jax.org/pub/reports/index.html

Even though those annotations are not assigned to a text span, they still can be used for the assessment of a text mining solutions. Phenotype annotations held in the MGD have been established through manual curation (Hancock and Mallon, 2007). Therefore, they are deemed highly trustworthy. Especially in the light of replacing manual curation with automated solutions, using those annotations for evaluation purposes will elucidate the still existing gap between manual and automated annotation of text.

5.2.3.1 *Reduction of the annotation corpus for Phentomine$_{EQ}$*

While *Phentomine$_{LEX}$* is theoretically capable of extracting all MP annotations in the applied version of the ontology file, *Phentomine$_{EQ}$* is limited to MP's structure and process phenotypes (see 5.2.2.2). If the *gold standard* would also be used for the evaluation of *Phentomine$_{EQ}$*, this method would obtain a lower recall and consequently also a lower F-measure score due to not working with MP's full version. Therefore, we created a limited version of the *gold standard*, only covering the 3,761 structure and process phenotypes used for extraction (see 5.2.2.2). This lead to the reduction of the evaluation corpus to 499 documents (out of the 621), comprising 3,464 (47% of *gold standard*; unique 1,149) MP annotations (further referred to as *restricted gold standard* corpus).

5.2.3.2 *Evaluation with pre-composed phenotype ontologies*

Using MGD's annotation file as *gold standard* poses the limitation that annotations cannot be traced back to the text span in a paper which caused the annotation (see 5.2.3). This makes the analysis of potential flaws in either of the methods difficult. Especially in the case of *Phentomine$_{EQ}$*, thorough analysis methods are required as flaws can be introduced through a variety of sources, e.g. errors in the recognition of either *entity* or *quality*, or the EQ statement used to assign MP annotations. Therefore, we decided to incorporate an independent evaluation step for *Phentomine$_{EQ}$*.

This independent evaluation step comprised the automated generation of EQ statements for MP's structure and process phenotypes[9]. The automatically generated EQ statements are then compared to the manually defined EQ statements (see 2.2.2.2) and mismatches are recorded. Mismatches are recorded for any of the following five groups:

1. manually and automatically assigned EQ statements are identical,

9 Note that this is the work aim in chapter 6.

2. *Phentomine*$_{EQ}$'s constituents only subset of manual EQ statement
3. *Phentomine*$_{EQ}$ extracts more constituents than only those from manually assigned EQ statement
4. *Phentomine*$_{EQ}$ does not cover all constituents of EQ statements and also assigns additional
5. manually and automatically assigned EQ statement have nothing in common

Note that this evaluation was also based on MP's 3,761 structure and process phenotypes.

5.2.3.3 *Establishing baseline with Open Biomedical Annotator*

Even though we aimed at a solution possessing a high recall, we also aimed at a higher F-measure then existing solutions (see 5.2.2.1 and 1.4.2.1). Dowell et al. (2009) reported an unsatisfactory performance of the Open Biomedical Annotator (OBA) (Jonquet et al., 2009) system. Therefore, we used OBA as a reference system and calculated precision, recall and F-measure for it based on our test corpora.

To obtain comparable measures for both *Phentomine*$_{LEX}$ and *Phentomine*$_{EQ}$, we calculated the scores once based on the 621 documents contained in our *gold standard* and a second time based on the 499 documents contained in the *restricted gold standard* (see 5.2.3). However, the phenotypes were extracted only once for all documents and depending on the corpus included in, or excluded from, the calculation.

We used the web service component of OBA to extract phenotype data from the 621 documents contained in the *gold standard* corpus. An example client was downloaded from the help pages[10] and modified according to our needs. The client was used to annotate both, the MEDLINE citations and the full text documents taken from PMC. Even though OBA is capable of supporting annotations of all the in BioPortal (Noy et al., 2009) contained ontologies, we reduced the output annotations to MP (BioPortal identifier: 46,433). We also did not include the possibility of annotation enrichment provided by OBA, as enrichment usually leads to the inclusion of noise. It would also partially distort the comparability to both methods introduced in this thesis (see 5.2.2) as neither supports enrichment features yet.

10 https://bmir-gforge.stanford.edu/gf/project/client_examples/scmsvn/?action=browse&path=/trunk/Java/Annotator/

Performance measures calculated for OBA are precision, recall and F-measure (see 1.4.2.1). The obtained results are shown in table 5.

	# docs	# anno	prec (%)	rec (%)	f-meas (%)
abstracts	621	362	38.1 (19.5)	3.5 (3.23)	6.3 (5.5)
full text	621	594	21.5 (12.2)	10.4 (20.8)	14.0 (11.5)

Table 5: **Performance measures for OBA on the *gold standard* and the *restricted gold standard* corpus in parentheses.** # *docs:* number of documents sent for annotation. # *anno:* number of documents assigned at least one annotation. *prec, rec, f-meas:* performance measures calculated on *gold standard* and (*restricted gold standard*).

Independently from the corpus and disregarding whether it is an abstract or full text, OBA in general has a higher precision than recall in the chosen configuration. Notably, the precision drops and recall increases when shifting from the abstract to the full text which supports our hypothesis that some relevant phenotype information is only available in the full text of a paper. Thus, it is not sufficient to only analyse the abstract. The overall best performance of a 14% F-measure is achieved when evaluating on the *gold standard* corpus and analysing full text papers.

Furthermore, OBA's precision drops for extracting structure and process phenotypes from the *restricted gold standard* corpus (see table 5, values shown in parentheses) while recall increases. However, the overall F-measure drops together with the precision. This behaviour may be either due to the types of phenotypes (structure and process) or due to the documents in the *restricted gold standard* corpus. The documents in the *restricted gold standard* corpus may text-wise significantly differ from those in the *gold standard* corpus, but this needs further investigations.

5.2.4 *Phenotype contents in papers*

Focusing on a high recall when extracting phenotype information most likely increases the number of falsely identified phenotype annotations. Therefore it is necessary to identify ways of reducing potential noise in the assigned annotations. One possibility is to employ knowledge about the location of a phenotype in a paper. For example, while the introduction reports about the study investigated in the paper, it also provides background information for this study and speculations about further experiments. Due to that, an introduction likely intro-

duces some noise due to providing additional information which can be neglected as only the pure results are relevant. However, before such a section-based filter can be applied, it is necessary to identify the distribution of phenotypes in scientific publications. Based on the assessment of the distribution, sections can be identified which can potentially be omitted during the extraction process.

To assess the occurrences of phenotypes, it is necessary to investigate a number of full text papers which are preferably randomly chosen from a complete collection. The documents contained in the *gold standard* corpus (see 5.2.3) fulfil those requirements: they constitute a selection of the papers contained in PMC and are randomly chosen from the MGD database as to whether or not they possess a full text in PMC. Therefore, an assessment of phenotype occurrences was done on the 621 full text documents obtained from PMC, which have been used as *gold standard* corpus for the phenotype data extraction.

PMC as a whole, but also the subset of chosen documents, covers a range of journals which leads to differences in sections contained in a paper as each journal has their own section requirements. Therefore, a next step was the unification of the sections covered in the analysed papers.

To unify the different paper sections in the 621 documents, we extracted all possible section names contained in the full text of those 621 documents. We obtained 76 different section titles, but this list contained multiple references to the same type of section, just phrased differently. Differences in section titles is attributed to singular/plural and capitalisation. For instance, section titles included *Results and discussion* and *Results/discussion* which refer to the same type of section. We manually reduced the list of sections to the following nine sections:

1. introduction (*intro*),
2. methods and materials (*mm*),
3. results (*res*),
4. results and discussion (*res_dis*),
5. discussion (*dis*),
6. conclusions (*concl*),
7. discussion and conclusions (*dis_concl*)
8. additional files (*addf*) and

9. others (*other*).

A full list of all the 76 different section titles and their mapping is provided in appendix C.3. The obtained list also comprises types of sections which constitute the combination of other types of sections, e.g. *results and discussion* is the combination of *results* and *discussion*. In this analysis, we kept those combined section types, as it is unclear how to distribute the obtained annotations over each of the contained section types.

This way of sectioning a full text paper is only a preliminary attempt but provides a general idea of how phenotypes are distributed across the different sections of a paper. We intend to repeat this experiment with a more elaborate version of sectioning provided by Core Scientific Concept (CoreSC) annotations (Liakata et al., 2012). CoreSC annotations determine the scientific discourse of a paper on a sentence level and include sentence annotations such as hypothesis or background. The full text will be split into sentences and each sentence will be classified with a CoreSC annotation. Phenotypes are then attributed to each type of CoreSC annotations.

Continuing from collecting the different sections of a paper, we then used an annotation server set-up to assign MP annotations to the 621 full text documents. The annotation server was again used in its exact matching configuration, not supporting overlap of annotations but supporting synonyms of the MP concepts (see 5.2.2.2). The dictionary of the MP annotation server was derived from the MP .tbl file and comprised all concept labels and synonyms.

The content of the different sections was compared by collecting all unique MP concepts assigned to each of the unified section types. After the collection, we determined the pairwise overlap between each of the unified sections and recorded how many MP annotations are shared among two sections. If in general some sections are covered by others, we assume that to be visible from the determined overlap.

5.3 RESULTS

In this section, all the results obtained by applying Phentomine$_{LEX}$and Phentomine$_{EQ}$respectively to the *gold standard* and *restricted gold standard*, are described. Furthermore, the results of the assessment of the distribution of phenotypes in a scientific publication are provided (see 5.3.3).

5.3.1 *Extraction of phenotypes using lexical features (Phentomine$_{LEX}$)*

By applying *Phentomine$_{LEX}$* to the 621 documents in the *gold standard* corpus, we obtain a recall (see 1.4.2.1) in the best case of 86.2% (see table 6). This recall score is obtained when extracting phenotypes from the full text of a paper and assigning annotations based on sentences. By using this method, recall increases immensely (by approximately 50%) when applying the method to full text instead of abstracts, which suggests that some phenotype information is only accessible from the full text of a paper. Recall also increases when synonyms are incorporated confirming the use of synonyms for phenotype extraction.

The precision is in either case (abstract or full text) low. Best precision of 1.4% is obtained when analysing abstracts and annotations can be assigned either based on sentences or a sliding window. Low precision scores are generally obtained, when an applied method generates a lot of noise. The inclusion of synonyms also reduces precision, indicating that some of the synonyms produce noise instead of extracting valuable information.

As a consequence of the low precision scores, all the obtained F-measure scores are low. In the best case, an F-measure of 2.6% is obtained. This F-measure correlates with the best case of precision and is obtained when only abstracts are analysed and annotations are either assigned based on a sentence or a sliding window.

a/f*	s/w†	# doc	# anno	tp	prec (%)	rec (%)	f-m (%)
a	s	621	620	1893	**1.4**	25.7	**2.6**
a	s	621	620	2580	1.1	35.1	2.1
a	w	621	620	1889	**1.4**	35.7	**2.6**
a	w	621	620	2568	1.1	34.9	2.1
f	s	621	618	5461	0.5	74.3	1.1
f	s	621	618	6337	0.5	**86.2**	0.9
f	w	621	618	5444	0.6	74	1.1
f	w	621	618	6319	0.5	85.94	0.9

Table 6: **Using the information theory approach to extract phenotypes from literature results in a recall of 86.2%.** Annotations assigned based on * abstract (*a*) or full text (*f*), † sentence (*s*) or sliding window (*w*), without (white background) or with synonyms (light blue background) . # *docs:* number of documents sent for annotation. # *anno:* number of documents assigned at least one annotation. *prec, rec, f-meas:* performance measures calculated on *gold standard*.

5.3.2 *Extraction of phenotypes using EQ statements (Phentomine$_{EQ}$)*

Even though *Phentomine$_{EQ}$* was also run on the entire 621 documents, the performance measures have been calculated on the *restricted gold standard* (see 5.2.3.1). Best recall (see 1.4.2.1) of 17.1% is obtained when analysing full text papers and allowing some discrepancy (80% of constituents) between the manually assigned EQ statement and the text span to be annotated.

The best precision of 4% is however obtained when *Phentomine$_{EQ}$* is applied to abstracts only and allowing some discrepancy (only 80% of constituents required) between manually defined EQ statement and text to be annotated. This indicates that even though some information is only available from full text, the incorporation of the complete document leads to noise. The required discrepancy for an improved precision suggests that some of the constituents of phenotypes cannot be identified in text spans.

Even though the precision drops when shifting from abstracts to full text, the F-measure increase with the increased recall. The best F-measure of 3.5% is obtained when analysing full text papers, regardless of how much overlap is required to assign MP annotations.

a/f*	o†	# doc	# anno	tp	prec (%)	rec (%)	f-meas (%)
a	100%	621	155	72	3.7	2.1	2.7
a	80%	621	155	82	**4.0**	2.4	3.0
f	100%	621	565	576	1.9	16.6	**3.5**
f	80 %	621	565	594	1.9	**17.1**	**3.5**

Table 7: **Using EQ statements to extract phenotypes from literature results in a recall of 17.1%.** Annotations assigned based on * abstract (*a*) or full text (*f*), † sharing 80% or 100% of annotations with EQ statements. # *docs:* number of documents to be annotated. # *anno:* number of documents assigned at least one annotation. *prec, rec, f-meas:* performance measures calculated on *restricted gold standard*.

5.3.2.1 *Extracting EQ statements for MP concept labels*

Further to applying *Phentomine$_{EQ}$* to the *restricted gold standard*, the method was also evaluated by automatically generating EQ statements from the concept labels. This evaluation was executed on the 3,761 structure and process phenotypes *Phentomine$_{EQ}$* is limited to. Out of those 3,761, we obtain an EQ statements for 3,459 concept labels

when applying our MP phenotype extraction method which are split as follows over the five groups recorded (see 5.2.3.2):

1. 842 (23.7%) EQ statements could be recovered from concept labels
2. in 591 cases (16.6%) *Phentomine*$_{EQ}$ assigned only a subset of the constituents required to form the EQ statement
3. in 494 cases (13.9%) *Phentomine*$_{EQ}$ assigned not only the constituents required for the EQ statements but also adds additional concepts
4. for 1,274 concepts (35.9%) *Phentomine*$_{EQ}$ assigned not all constituents required for the EQ statement and instead provided constituents not relevant to the EQ statements
5. 348 cases (9.8%), the extracted EQ statement had nothing in common with the manually assigned EQ statement

Assuming that additional constituents can be manually filtered as long as all required constituents are present, this method provides the correct EQ statement for 37.6% of the structural and process phenotypes contained in MP.

Due to the relation between this evaluation task and the work aim of chapter 6, we consider this to be the baseline, improvements can be compared to.

5.3.3 *Spread of phenotypes across sections*

Phenotype annotations were uniquely collected for nine pre-defined sections in full text papers. Even though not all phenotype data may be retrieved this way, we still expect to see trends for sections. Collecting MP annotations showed that in the investigated subset, the results sections possesses the most phenotype content with 625 unique annotations, followed by the discussion (509 unique annotations) and the introduction section (375 unique annotations) (see table 8). The remaining six sections have comparatively little MP annotations. Taking the structure of a research paper into consideration, those results confirm its structure: the introduction provides the main findings, the results section provides a detailed view on all results, and the discussion provides a discussion of the results.

Determining the pairwise overlap between sections (see table 8) yields that the results section shares the greatest overlap with most

of the other section, followed by the discussion section. While all other sections share the greatest overlap with the results section, the introduction shares its greatest overlap with the discussion section. Relating this back to the structure of a scientific publication, this suggests that the main findings mentioned in the introduction are further assessed in the discussion section.

5.4 DISCUSSION

5.4.1 *A long way to achieve reliable phenotype extraction*

Even though the goal of obtaining a high recall (at best 86.2 % measured on *gold standard*; see table 6) has been achieved by applying *Phentomine*$_{\text{LEX}}$ to the extraction of phenotype information, the goal of achieving a reasonable precision was not met in either approach (at best 4% measured on the *restricted gold standard* corpus; see table 7). Despite the recall of *Phentomine*$_{\text{LEX}}$ being higher than achieved with OBA (comparison based on *gold standard*; see table 5), both methods cannot yet compete with the F-measure scores obtained with OBA. Given that OBA was deemed unsatisfactory for MGD purposes (Dowell et al., 2009), improvements to either method are required before results can be provided to curators or used for the automated population of phenotype databases.

Both methods are impacted by the way scientific findings are reported. For example, when reporting about scientific findings, references to and reviews about previous work are provided. This means that there is a likelihood that a scientific paper also reports additional phenotypes that are not relevant for annotating mouse models. This additional information leads to false positive annotations and decreases not only precision and recall, but also the F-measure for both methods.

One example falling into this category, and which is likely when considering mouse models, is a report about the study of a particular gene. In most cases, the study of a gene involves the study of a particular phenotype or a small set of related phenotypes. If there have been studies about this particular gene before, then the paper will also report about the phenotypes which have been associated with the gene before. To illustrate this further, two sentences from the

	#	intro	mm	res†	res_dis	dis	concl	dis_concl	addf	other
intro	375	**X**	125	274	109	**275***	35	2	64	22
mm	191	125	**X**	**164***	76	140	29	2	53	17
res	625	274	164	**X**	134	**363***	36	2	80	20
res_dis	168	109	76	**134***	**X**	127	27	2	49	11
dis	509	275	140	**363***	127	**X**	36	2	73	21
concl	38	35	29	**36***	27	**36***	**X**	2	19	13
dis_concl	2	**2***	**2***	**2***	**2***	**2***	**2***	**X**	**2***	**2***
addf	87	64	53	**80***	49	73	19	2	**X**	12
other	22	**22***	17	20	11	21	13	2	12	**X**

Table 8: **Distribution of phenotype data across different sections of full text paper.** * highest number of annotations covered by other section; † From all sections, *results* section covers most annotations also contained in other sections.

paper entitled *"Impaired wound healing in mice deficient in a matricellular protein SPARC (osteonectin, BM-40)."* (PMID:11532190)[11] are provided:

```
In large (25 mm) wounds, SPARC-null mice showed a
significant delay in healing as compared to wild-type
mice (31 days versus 24 days).
[...]
SPARC (secreted protein acidic and rich in cysteine) is an
extracellular glycoprotein expressed in elevated levels in
actively proliferating cells and organs such as developing
embryos and adult tissues associated with remodeling
(bone and gut), wound healing, angiogenesis (reviewed
in refs. [3,4]), and tumorigenesis [5,6].
```

The first sentence, taken from the abstract of the paper, reports about the scientific finding: the *delayed wound healing* due to deletion of the *Sparc* gene (MGI:98373). However, the second sentence (taken from the Background section) reports about previous phenotype-related findings about the Sparc gene (namely *remodeling (bone and gut), wound healing, angiogenesis, and tumorigenesis*). Despite *wound healing*, all the other mentioned phenotypes would cause false positives annotations.

Another source of false positive annotations is the report about initial expectations on an biological experiment, e.g. in what ways a phenotype could change due to the modification of a certain gene. For example, the paper entitled *"Generation and characterization of the Anp32e-deficient mouse."* (PMID:21049064) [12] reports that for the mutagenesis experiments executed, no abnormal phenotypes could be observed. At the same time, the abstract of the paper also states:

```
No defects in thymocyte apoptosis in response to various
stresses, fibroblast growth, gross behaviour, physical
ability, or pathogenesis were defined.
```

Apart from the improvements to the individual method, some of the false positive annotations produced by either method can be addressed by the inclusion of scientific discourse (see 5.2.4, 5.3.3 and 5.4.2), negation and speculation identification. Additional sources of error and potential improvements with respect to the individual methods are discussed in the following sections.

11 http://www.ncbi.nlm.nih.gov/pubmed/11532190
12 http://www.ncbi.nlm.nih.gov/pubmed/?term=21049064%5Buid%5D

5.4.1.1 *Limitations of the current Phentomine$_{LEX}$ implementation*

As our *gold standard* corpus (see 5.2.3) does not provide the exact phenotype locations in text, the identification of possible flaws in both methods is difficult. While it was possible to evaluate Phentomine$_{EQ}$ with an additional procedure (see 5.3.2.1), we did not find an additional evaluation case for Phentomine$_{LEX}$ yet. One possibility will be to manually annotate a small subset of sentences to assess the performance of the method.

A high recall and a low precision as performance measures suggest that there are a number of false positive annotations among the extracted phenotype data. Further to the sources of false positive annotations discussed in section 5.4.1, there are also method-related sources of false positive annotations. Due to the fact that the method assigns phenotype annotations based on the information content (IC) of term labels, in combination with evidence and specificity (see 5.2.2.1), a potential source of noise are term labels which share a high overlap in words and possess a similar overall IC, e.g. *increased fatty acid level* (MP:0005281) and *decreased fatty acid level* (MP:0005282) as well as *abnormal fatty acid level* (MP:0005280).

To illustrate this further, one of the example sentences provided before (see 5.4.1), can serve as example again:

```
In large (25 mm) wounds, SPARC-null mice showed a
significant delay in healing as compared to wild-type
mice (31 days versus 24 days).
```

The desired annotation (according to the gold standard) from this sentence would be *delayed wound healing* (MP:0002908). However, due the way the method is currently implemented, this sentence would also obtain the annotations *enhanced wound healing* (MP:0002724), *impaired wound healing* (MP:0001792), and *abnormal wound healing* (MP:0005023). While *abnormal wound healing* could be filtered out using the hierarchy of the ontology, other ways to remove the other unwanted annotations have to be found. Filtering according to the hierarchy may have its own caveats if more generalised annotations are required.

One possibility is that the measures for evidence and specificity (see 5.2.2.1) are not suitable for the extraction of phenotype concepts. These measures were obtained for GO terms and therefore may be ontology specific. To determine the best measures for evidence and specificity, the application of Phentomine$_{LEX}$ to a small set of manual

annotated abstracts and full text papers is required. Based on this reference corpus, ideal values for both parameters can be determined, which may even differ for abstract and full text of a paper. In addition, tests are required as to how the IC scores of individual concepts look like and whether they possess sufficient discriminatory power.

5.4.1.2 *Limitations of the current Phentomine$_{EQ}$ implementation*

For 348 out of 3,459 concepts labels (9.8%, see 5.3.2.1), *Phentomine$_{EQ}$* did not identify any constituents of the manually defined EQ statement. This means that both the identification of the *entity* and the *quality* failed at the same time. A subset of 40 concepts (out of the 348) was manually investigated to identify the common causes of mismatches. The groups of mismatches identified here may either concern *entity* or *quality* of the phenotype but not the entire EQ statement. A combination of two or more errors is required for a complete mismatch to happen.

Derived from those 40 further investigated cases, three common groups of errors have been identified. The first group of errors results from MP concepts possessing the word *absent* in their label. While the automated label-based method assigns as *quality* the PATO concept *absent* (PATO:0000462), a manual curator assigns as phenotype *quality* the PATO concept *lacks all parts* (PATO:0002000). This group of errors can be easily avoided by simply defining a replacement rule which requires the *quality* constituent to be replaced, whenever the MP concept label contains the word *absent*.

A second group of errors has its origins in the way the constituents are identified in concept labels. While the annotation servers work sequentially, annotations are not propagated from one server to the other. This leads to parts of concepts obtaining more than one annotation and consequently causes too many constituents being assigned to one concept. For example, with the automated method, the MP concept *enlarged dorsal root ganglion* (MP:0008490) would not only obtain the MA annotation *dorsal root ganglion* (MA:0000232) but also the PATO annotation *dorsal* (PATO:0001233).

The third group of errors lies in the composition of PATO and how PATO is used for building the EQ statements. Some PATO concepts can be built as a composition of other PATO concepts but yet, they still exist as individual concepts. For instance, the concepts *decreased depth* (PATO:0001472) can be built by using the PATO concepts *decreased* (PATO:0001997) and *depth* (PATO:0001595). Subsequently, automati-

cally identifying constituents of the MP concept *decreased palatal depth* (MP:0003766) extracts two PATO concepts (due to the longest consecutive span being annotated) while it is only one PATO concept in the manually designed EQ statements. A potential way of eliminating those errors is to derive a term-based composition of PATO.

5.4.1.3 *Creating a combined Phentomine method*

Thus far, the performances of the individual methods $Phentomine_{LEX}$ and $Phentomine_{EQ}$ do not yet allow a decision of whether or not both methods can be eventually integrated. Further investigations are required, not only to improve the individual methods, but also to assess the potential of both methods to complement each other. However, the determination of the complement makes little sense as long as both the methods are subject to change (to correct for flaws identified in either method). Only once both methods are stable, it can be determined whether and how those methods can complement each other.

5.4.2 *Potential filtering of annotations with sections*

When extracting phenotype information from papers and recording the location of the occurrences of the phenotypes, the highest contribution of phenotype information is in the results section (see table 8). Furthermore, the results section shares the biggest overlap in terms of phenotype information with most of the other sections. Only the introduction section shares its biggest overlap with the discussion. From the obtained numbers, two possible filtering options could be derived, aiming at the reduction of noise in the extracted phenotype data:

1. only using results sections and
2. using introduction, results and discussion section in conjunction.

If a filtering according to the results sections is applied, only phenotype data occurring in this section will be reported. This still may result in a huge number of extracted phenotype information as this section is the biggest contributor of phenotypes at the moment. Due to all results being reported in the results section, it may as well be that also minor findings or limitations are reported in this section.

Phenotypes reported as minor findings or in limitations would add to the noise level and would be better left out.

Therefore, a second way of filtering could be more promising. In the second option, phenotype data is only reported if it is consistently mentioned across introduction, results, and discussion. With this filtering concepts, it is ensured that major findings are reported back, assuming that those are the most relevant to be gathered. However, this approach may cause seemingly minor phenotype information not to be reported even though it might be crucially relevant if put in another context.

Both these approaches are one attempt to reduce noise when extracting phenotype information automatically and further studies are required as to which is more favourable over the other.

5.5 CONCLUSIONS

Applying an information theory approach (*Phentomine*$_{\mathrm{LEX}}$) to the extraction of phenotype information enables the identification with high recall. However, the method produces a lot of annotation which seem not necessarily relevant to be reported. Using an ontology-based approach instead (*Phentomine*$_{\mathrm{EQ}}$), confirms previously identified gaps between unstructured text and the way ontologies are built (Travillian et al., 2011). Even though neither method can compete yet with the F-measure obtained from OBA on the test corpora, we identified a number of errors (see 5.4.1) which will potentially improve the extraction of phenotype data.

Even though neither method has a high performance yet, previous findings about the differences in the information attainable from either abstract or full text could be confirmed (Cohen et al., 2010). When extracting phenotype information, the full body of a publication is required to gather the data which is relevant to a curator and consequently a database and the studies relying thereupon.

Analysing the occurrences of phenotype data in a scientific publication also suggests potential to reduce noise by discriminating the occurrences of phenotype information. From our results, two options arise on how to filter phenotype information according to sections and can be tested in further studies.

5.6 FUTURE WORK

As discussed in 5.4.1, both methods Phentomine$_{LEX}$and Phentomine$_{EQ}$ are equally impacted by the way scientific findings are reported. In addition to reports about previous findings, authors of scientific papers also include experimental hypotheses and report absences of expected outcomes. Therefore, directions for future work include more analysis to develop a discourse-based filter (see 5.4.2), as well as the inclusion of negation and speculation identification software. Software tools such as NegEx (Chapman et al., 2001) or BiographTA[13] (Morante and Blanco, 2012) for negation and speculation, as well as CoreSC annotations are strong candidates for the inclusion into a phenotype text mining pipeline.

In addition to the overall improvements, each of the methods has its own additional caveats (see 5.4.1.1 and 5.4.1.2). To improve Phentomine$_{LEX}$, the development of a manually annotated corpus is required to estimate what good measures for specificity and evidence (see 5.2.2.1) are. This corpus will also help to analyse further problems in Phentomine$_{EQ}$. Assessing the performance of both methods on a more suitable corpus to the performance leak identification may highlight additional problems common to both approaches apart from these identified here. Phentomine$_{EQ}$has been improved according to the suggestions in 5.4.1.2 while applying it to the decomposition of phenotypes (see chapter 6), but further improvements are required, in line with suggestions made in 5.4.1.2. The suggestions mainly cover the extension of the underlying dictionaries as well as finding a way to deal with clinical labels.

Another step for the improvement of extracting phenotypes from literature includes the filtering of synonyms, which have been treated all equally at the moment. The OBO Flatfile format (see 1.2.1) provides a detailed distinction of synonyms and in future work, we intend to use this distinction and with that provide the means to reduce the falsely identified phenotypes in literature. Further to that, we plan to make more use of the textual definitions, also provided within the ontology file, to be able to filter false positives from correctly identified phenotype annotations.

A visual representation of extracted phenotypes is critical its acceptance its potential integration into workflows. Therefore, a required extension to the phenotype extraction pipeline is the visual represen-

13 http://www.clips.ua.ac.be/BiographTA

tation of the obtained results and a possible verification through a curator. Phenex[14] (Balhoff et al., 2010) or Brat[15] (Stenetorp et al., 2012) are good candidates to provide the extracted results to a curator.

14 http://phenoscape.org/wiki/Phenex
15 http://brat.nlplab.org/

6

"EQ-LISING" PRE-COMPOSED PHENOTYPE ONTOLOGIES

Some of the disease gene prioritisation methods that use logical definitions to bridge between organisms, such as the one described in chapter 3, rely on the availability of logical definitions for pre-composed phenotypes. In this chapter, I report about my work that focused on the decomposition of pre-composed phenotypes to derive EQ statements.

6.1 BACKGROUND

Due to a lack of integration, the variety of co-existing phenotype descriptions (see 2) creates encapsulated knowledge in databases (Gkoutos et al., 2012; Schofield et al., 2012; Oellrich and Rebholz-Schuhman, 2010). In addition to ontology alignment algorithms (see 1.2.3), one other bridging mechanism found increasing application: the entity–quality (EQ) representation of phenotypes (Mungall et al., 2010) (see 2.2.2.2). In this representation, a phenotype is decomposed into an affected *entity* which is further described with an *quality*. For example *decreased body weight* can be split into the following parts: *body* as *entity*, *weight* as *quality* and *decreased* as a modifier for the *quality*.

EQ statements have been successfully applied in a number of studies, focusing on cross-species phenotype integration (Washington et al., 2009; Chen et al., 2012; Hoehndorf et al., 2011c). Those studies have shown promising results even though only a subset of pre-composed phenotype concepts possess an EQ statement yet. Extending the number of available EQ statements could potentially improve the results of these studies as more data becomes accessible.

In the current ontology development cycle, first pre-composed phenotypes are created and once finalised, the corresponding EQ statements are generated. Both pre-composed phenotypes and corresponding EQ statements are created manually. Due to the number of available pre-composed phenotypes and the creation of EQ statements being time-consuming, only a subset of concepts is available in EQ. For example, in the case of the Human Phenotype Ontology (HPO) (Robinson

et al., 2008), only 4,783 out of 9,795 possess a manually assigned EQ statement (as of June 2012).

A pre-composed phenotype ontology is, as well as other ontologies, a community effort and thus it is subject to change. Concepts evolve, get obsolete or simply change over time. As EQ statements are used to represent the pre-composed phenotype concepts, after each change in the dependent ontology, the EQ statements may have to be amended. Keeping the EQ statements updated is a requirement to ensure correctness and quality in integration processes. Therefore, an automated method capable of keeping pace with the ontologies' development cycle is required, assuring the availability and quality of EQ statements.

In previous studies, it has been shown that the lexical structure of ontological concepts can give insights to the formal definitions of the concepts (Wroe et al., 2003; Aranguren et al., 2008; Mungall, 2004; Ogren et al., 2004, 2005). The applied methods included regular expressions and grammar rules to derive logical definitions from Gene Ontology (GO) concept labels. However, applying either regular expressions or grammar rules requires a prior knowledge about the textual structure of the labels and how this may correspond to the formal representation. Users that are less aware about formal representations, may struggle with defining rules or expressions required to derive a formal representation.

This chapter reports about the development of the *EQ-liser* method, to decompose pre-composed phenotype ontologies into a post-composed phenotype representation. A prototype of the *EQ-liser* method was implemented and applied to structural and process phenotypes contained in MP and HPO. From these experiments, we were able to draw conclusions for a generalised decomposition method. Further to the decomposition of phenotype concepts, *EQ-liser* also facilitate the discovery of inconsistencies in the manually assigned EQ statements. Moreover, the approach can also be applied to elucidate inconsistencies in the concept labels of pre-composed phenotype ontologies.

A modified version of this chapter has been submitted as a contribution to the 4th Ontologies in Biomedicine and Life Sciences (OBML) workshop 2012. For more details refer to page xi. Prototype source code and results are available online[1].

1 https://code.google.com/p/eqliser/

6.2 METHODS AND MATERIALS

To decompose pre-composed phenotypes, a method is required to identify the constituents of the phenotype, in this case an *entity* and a *quality*. For instance, the MP concept *muscle degeneration* (MP:0003235) is manually defined as:

```
[Term]
id: MP:0000749 ! muscle degeneration
intersection_of: PATO:0000639 ! degenerate
intersection_of: inheres_in MA:0000015 ! muscle
```

Consequently an automated method would have to assign two constituents here: one for the *entity* (*muscle*; MA:0000015) and on for the *quality* (*degenerate;* PATO:0000639).

6.2.1 *Input data*

In the existing, manually derived EQ statements, the *entity* constituent is represented with a number of OBO Foundry ontologies[2] (Smith et al., 2007). The *quality* constituent is always represented using the Phenotype And Trait Ontology (PATO) (Mungall et al., 2010; Mabee et al., 2007). Ontologies used to describe *entities* may differ from species to species. Supporting all these ontologies would be out of scope of our preliminary study. Therefore, we limited our approach to two species-specific ontologies, HPO and MP. Further to that, we also limited those further to phenotype concept being represented in EQ using the Mouse adult gross Anatomy ontology (MA) (Hayamizu et al., 2005), the Gene Ontology (GO) (Ashburner et al., 2000), the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) and PATO. We consider this to be corresponding to structural and process phenotypes.

A version of the two phenotype ontologies was downloaded as .tbl file, together with their corresponding EQ statements on the 3 May 2012. HPO comprised 9,795, while MP included 9,127 MP concepts. 4,783 HPO and 6,579 MP concepts possessed a manually assigned EQ statement. After a reduction to their structural and process phenotypes, the MP concepts were reduced to 3,761 and the HPO concepts were reduced to 3,268. Furthermore, we obtained the .tbl files for MA, FMA,

2 http://obofoundry.org/

GO, and PATO which are transformed into dictionaries for the NER system (see 5.2.2.2).

6.2.2 *Implications from text mining phenotypes*

In the previous chapter (see 5), a method (Phentomine$_{EQ}$; see 5.2.2.2) was described to extract phenotype information from scientific literature. Due to the relation of the tasks – extracting EQ statements from text spans – the method was modified to become *EQ-liser*'s prototype.

In the course of evaluation, the phenotype extraction method was applied to MP concept to automatically derive EQ statements. The automatically derived EQ statements where then compared to those which had been manually assigned. A correct EQ statement was assigned for only 23.7%. Identified flaws of the method were discussed in 5.4.1.2 and are summarised here:

1. *absent* (PATO:0000462) versus *lacking all parts of type* (PATO:0002000) (addressed in 6.2.2.1)
2. identification of too many constituents (addressed in 6.2.2.2)
3. split PATO concepts (addressed in 6.2.2.3)

In order to reuse the Phentomine$_{EQ}$ method, we addressed those flaws as described in the following subsections and built the *EQ-liser* prototype out of it.

6.2.2.1 *PATO concept replacement*

When extracting phenotypes from literature, a first type of mismatches occurred, when concept labels contained the word *absent*. Despite the existence of an *absent* (PATO:0000462) PATO concept, these MP concepts possess the PATO concept *lacks all parts of type* (PATO:0002000) as *quality* in their manually assigned EQ statements. We note here that this is in correspondence with the formalised phenotype description patterns suggested by Hoehndorf et al. (2010b). However, our automated method assigned the *absent* (PATO:0000462) PATO concept to all those concepts due to the term name.

To accommodate for the correct representation of those concepts, our method would have to be altered. A simple solution would be a replacement rule. This replacement rule requires that, whenever the *quality absent* (PATO:0000462) would be assigned, a replacement happens with the PATO concept *lacks all parts of type* (PATO:0002000).

6.2.2.2 *Removing overlapping constituents*

The second type of mismatches was the identification of too many constituents due to overlaps (see 6.2.2). The servers applied to identify the *entity* and the *quality* inside a pre-composed concept work independently (see 5.2.2.2). This means that none of the individual servers possesses information about the constituents assigned from other servers. Consequently, two or more of the employed servers may identify constituents in the same text span of the concept.

For example, one identified constituent in *enlarged dorsal root ganglion* (MP:0008490) would be *dorsal root ganglion* (MA:0000232). Another identified constituent in this term would be *dorsal* (PATO:0001233). In this particular example, the second constituent *dorsal* (PATO:0001233) is redundant as it is entirely covered by *dorsal root ganglion* (MA:0000232).

Overlaps are not specific to any of the employed ontologies, i.e. anatomy and process constituents may overlap as well as process constituents may overlap with the *quality* constituent. As those redundant annotations create noise, it is desired to filter those annotations out.

A solution to this problem is a position-based filtering of constituents. After all constituents have been identified together with their positions, those are removed, which are entirely enclosed by other constituents. Applying this filter mechanism will provide the minimum number of constituents to decompose the concepts based on their label.

However, filtering is not possible for the process constituents in the current implementation of the prototype. This is due to the implementation of the applied GO server. The employed GO server is implemented according to the method described in (Gaudan et al., 2008), requiring a defined length of the text span the annotations are assigned to. Thus, in its current implementation, the GO server assigns constituents to complete concept labels instead of spans thereof. Consequently, process constituents are not included when filtering, as it would lead to the removal of all other constituents.

6.2.2.3 *Deriving a term-based decomposition of PATO*

A third type of mismatches occurred due to PATO concepts constituting a term-based combination of other PATO concept. For instance, the concept *decreased depth* (PATO:0001472) could be represented using the PATO concept *decreased* (PATO:0001997) and *depth* (PATO:0001595). An example of a pre-composed phenotype where this term-based com-

position leads to problems is e.g. the MP concept *decreased endocardial cushion size* (MP:0000301). It possesses the EQ statement:

```
[Term]
id: MP:0000301 ! decreased endocardial cushion size
intersection_of: PATO:0000587 ! decreased size
intersection_of: inheres_in MA:0000078 ! endocardial cushion
```

Due to the applied servers annotating the longest consecutive string, this MP concept would obtain two *qualities*: *decreased* (PATO:0001997) and *size* (PATO:0000117). However, a manually assigned EQ statement summarise both into one concept: *decreased size* (PATO:0000587). A solution to avoid these mismatches is a term-based decomposition of PATO. The term-based decomposition can be applied to "summarise" multiple *qualities* into one. We note here that this might not hold always true but may account for the majority of the cases.

To achieve a representation of PATO concepts together with their constituents, we applied a textual matching approach based on stemmed words in concept labels and synonyms. The following four steps were applied:

1. download .tbl file from OBO Foundry web page[3]
2. special character and stop word removal (see 5.2.2.1 and appendix C)
3. applying Snowball's Porter stemmer[4] (Porter, 1980) (see 5.2.2.1)
4. pairwise match of concepts based on stemmed labels and synonyms

The downloaded .tbl file comprised 2,290 PATO concepts. By applying the aforementioned steps, we derived term-based decomposition of 1,453 (63% from all) PATO concepts[5].

6.2.3 *Resulting workflow*

To accommodate the required changes described in the previous section, we adapted and extended the phenotype extraction pipeline (Phentomine$_{EQ}$; see 5.2.2.2). An illustration of the resulting *EQ-liser* prototype is shown in figure 19. Each of the individual steps are explained in the following.

3 http://www.berkeleybop.org/ontologies/
4 http://snowball.tartarus.org/
5 available online: http://code.google.com/p/eqliser/

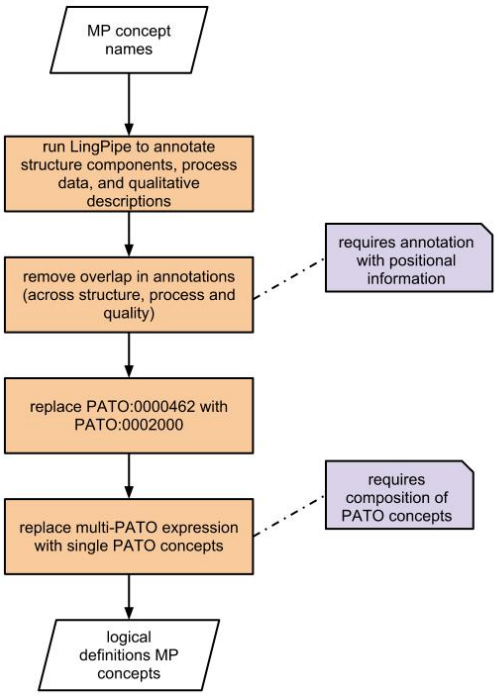


Figure 19: **Illustrates the workflow to decompose pre-composed phenotype concepts into EQ statements.**

The first step (see figure 19) in processing the ontology's downloaded .tbl file was the filtering for special characters ("%","-") and stop words ("in","or"). Special characters are all characters which are neither alphabetic nor a number. Those special characters, often punctuation, potentially cause problems when matching differently punctuated concept labels from other ontologies. Stop words are part of the common English language, but considered not to carry any discriminatory information. Consequently, stop words can be removed before analysis to reduce noise and potential errors resulting from their inclusion.

After character filtering and stop word removal from all the concept labels, we used annotation servers to identify the constituents of the ontology concepts. Underlying those servers is the LingPipe[6] (Carpenter, 2007) library, configured to perform exact matches. The dictionaries for LingPipe were compiled by using the labels and synonyms provided by the ontology files of FMA, MA and PATO. An individual tagging server was set up for each ontology. To recognise processes, we used an alternative approach described by Gaudan et al. (2008). After identification, overlapping (based on position in term) constituents were removed (see 6.2.2.2).

6 http://alias-i.com/lingpipe/

The last step in the workflow is the replacement of multiple identified *qualities* by applying the derived, term-based PATO compositions (see 6.2.2.3). This step also included the replacement rule for the *quality absent* (PATO:0000462) (see 6.2.2.3).

6.2.4 *Evaluation*

To evaluate our results, we introduced a two-step evaluation process. We first evaluated the obtained EQ statements for HPO and MP to the available manually assigned EQ statements of structural and process phenotypes. Comparing the automatically generated EQ statements with manually assigned, leads to five groups (see table 9):

1. manual and automated method identical (*equ*),
2. *EQ-liser* constituents only subset of EQ statement (*lp<ld*)
3. *EQ-liser* assigns more than only EQ statement constituents (*lp>lp*)
4. *EQ-liser* does not cover all constituents of EQ statements and assigns additional (*lp∩ld*)
5. constituents of *EQ-liser* and manually assigned EQ statements have nothing in common (*zero*)

In a second step, we manually investigated a subset of 100 EQ statements where the *EQ-liser* prototype does not identify any constituent correctly (corresponds to group 5, *zero*). The 100 concepts were equally split between HPO and MP. Potential errors in the prototype were identified independently for each ontology. Those 100 EQ statements were randomly chosen, using R's sample function.

6.3 RESULTS

6.3.1 *EQ-lising MP*

We measured the performance of the *EQ-liser* prototype against the manually defined EQ statements of the concepts (see 6.2.4). Applying the *EQ-liser* prototype to the MP ontology concept labels, we obtain automatically assigned EQ statements for 3,549 concept (94.4%) out of 3,761 structural and process phenotypes contained in MP (see 6.2.1).

1,079 concepts (30.4%) of the pre-composed phenotype concepts obtain a correct EQ statement, while 200 concepts (5.6%) obtain an

entirely wrong EQ statements. In the remaining 64% of the cases, the *EQ-liser* prototype assigns either a partially or entirely correct EQ statement but also adds additional constituents. If we relax the criterion to the method assigning the correct constituents contained in the manually assigned EQ statements but also allow additional, we achieve an improvement of 14.6% and 52.2% in total. 50 concepts out of the 200 concepts not sharing any overlap with the manually assigned EQ statements were further evaluated and results are discussed in 6.4.1.

Table 9 shows the numbers of concepts falling into either of the five distinguished evaluation groups (see 6.2.4).

6.3.2 *EQ-lising HPO*

To determine whether the prototype possesses the potential to be generalised into a species-independent method, we also assessed the performance of the *EQ-liser* method on a second pre-composed phenotype ontology. Therefore, we applied the *EQ-liser* prototype to the 3,268 structure and process phenotypes contained in HPO (see 6.2.1). When applying the *EQ-liser* method to HPO, 2,731 HPO concepts (84%) are assigned an EQ statement.

From those 2,731 concepts, in 249 cases (9.5%) the correct EQ statement is assigned. If we relax the criterion and only expect the correct manually assigned constituents among the *EQ-liser* constituent, we obtain the correct constituent for 363 concepts (13.3%). In 25.8% of the cases, the automated method assigns only constituents which do not occur in the manually assigned EQ statements. 50 randomly chosen concepts from this group are manually investigated and the results are discussed in 6.4.2.

Table 9 shows the numbers of concepts falling into either of the five distinguished evaluation groups (see 6.2.4).

6.4 DISCUSSION

To identify flaws in our method, we manually investigated concepts for which the *entity* and the *quality* identification fails (see column *zero* in table 9, also 6.2.4). 50 MP (out of 200) and 50 HPO (out of 705) were randomly chosen for manual investigation. We were able to identify common patterns why *entity* and *quality* recognition fails.

	no†	equ	lp<ld	lp>ld	lp∩ld	zero	total	%‡
MP	3761	1079	448	775	1047	200	3549	**52.2** (30.4)
HPO	3268	249	571	113	1093	705	2731	13.3 (9.1)

Table 9: **Performance measures when applying *EQ-liser* to MP and HPO concept labels.** † number of structural/process phenotypes in ontology; ‡ percentage of concepts with correctly identified constituents, no additional constituents in parentheses. *equ:* manual and automated method identical; *lp<ld: EQ-liser* constituents only subset of EQ statement; *lp>lp: EQ-liser* assigns more than only EQ statement constituents; *lp∩ld: EQ-liser* does not cover all constituents of EQ statements and assigns additional constituents; *zero:* constituents of *EQ-liser* and manually assigned EQ statements nothing in common; *total:* number of concepts automatically decomposed.

6.4.1 *Mismatches in MP concepts*

For entire mismatches to occur, both the identification of *entity* and *quality* fails. Instead of grouping the mismatches to combinations of *entity* and *quality* failures, we grouped them by ontology. Due to the limitation to structure and process phenotypes, *entities* are determined by MA and GO annotations; *qualities* are described in either case with PATO annotations.

One common group of mismatches in PATO annotations is due to particular extension/replacement patterns in the manually assigned EQ statements. Those manual extension/replacement patterns cannot be recognised with the *EQ-liser* method yet. For instance, *increased mitochondrial proliferation* (MP:0006038) possesses *increased rate* (PATO:0000912) as *quality* in the manually assigned EQ statements. Applying the *EQ-liser* method in its current state assigns as *quality increased* (PATO:0000470), which causes the mismatch in the *quality* part of the EQ statement. Similarly, any concept label possessing the phrase *increased activity* will be annotated in the manually assigned EQ statements with *increased rate* (PATO:0000912) as *quality* while *EQ-liser* assigns *increased* (PATO:0000470) as *quality* and therefore causes a mismatch. Furthermore, any concept label possessing the phrase *increased ... number* will be represented in the manual EQ statements with *has extra parts of type* (PATO:0002001) while in *EQ-liser* assigns *increased* (PATO:0000470) and *number* (PATO:0000070) as *quality*. Despite all those examples were now only shown with *increased*, the same is true for all concept labels possessing the word *decreased*. All those examples provided here could be changed with conditional,

hand-crafted replacement rules for PATO concepts and consequently lead to a reduction of the contradictory cases.

Mistakes in the identification of anatomical components occur when the naming of the affected anatomical structure differs in both MA and MP. This is mostly due to different singular/plural representations. For instance, the MA annotation *lumbar vertebra* (MA:0000312) is not automatically assigned to the MP concept *increased lumbar vertebrae number* (MP:0004650) due to *vertebra* being used in singular in MA but in MP in plural. Another source of mismatches in the anatomical structure is due to shortened expressions, e.g. MP uses *coat* while MA uses *coat hair*. The described mismatches could be addressed by adding additional terms to the dictionary underlying the MA annotation server.

Mismatches in the process information of pre-composed phenotype ontologies were caused e.g. due to not covering synonyms in the current version of the GO server. An example falling into this category are concept labels containing *salivation*, which are not recognised as *saliva secretion*. Other mismatches were caused by differences in words meant to express the same, e.g. *smooth muscle contractility* and *smooth muscle contraction*. A small fraction of mismatches in the process constituent is also caused by singular/plural-conflicts, e.g. MP uses plural *cilia* while GO uses singular *cilium*. Both synonyms and singular/plural-conflicts could be addressed by extending the dictionary underlying the current GO server.

Among the 50 manually investigated concepts, we could also identify a small proportion of wrongly manually assigned EQ statements. Those cases were reported to the developers of the EQ statements and got corrected. The wrongly assigned manual definitions were mainly due to old construction patterns and newly added concepts in the constituent ontologies.

6.4.2 *Mismatches in HPO concepts*

While the number of absolute mismatches in the case of MP is small (5.6%), it increases to 25.8% in the case of HPO (see column *zero* in table 9). Those numbers suggest that the structure of the textual data used to describe MP concept labels is better suitable for textual decomposition than HPO. One reason for MP's concept labels working better than HPO's is that HPO employs clinical terms as labels, while MP uses multi-term descriptions. The usage of clinical terms masks the *entity* as well as the *quality* of a concept. For example, HPO uses the term

Brachycephaly (HPO:0000248) as label instead of e.g. *decreased anterior-posterior diameter of the skull*[7]. MP describes the same phenotype as *shortened head* (MP:0000435) and uses *brachycephaly* only as a synonym.

To evaluate mismatches occurring in the decomposition of HPO, we again manually investigated 50 randomly chosen examples. The examples were taken from those concepts where *EQ-liser* and manually assigned annotations do not share any overlap (see column *zero*, table 9). Again, entire mismatches occur when both *entity* and *quality* recognition fail. As for mismatches in the MP decomposition, observed patterns were grouped according to the ontologies used: PATO for *qualities*, FMA and GO for *entities*. We note here that even though the samples were chosen at random, most of the concepts concerned either hand or foot and hence may constitute a bias.

One source of causing mismatches on the *quality* defining part of a phenotype, are due to the types of words (noun, verb, adjective, et cetera) used for description. For example, all HPO concepts containing the nouns *abnormality* and *abnormalities* are not automatically annotated with *abnormal* (PATO:0000460) due to the differences in word types. In addition to that, all concepts containing *abnormality* or *abnormalities* in their labels are manually assigned the PATO annotation *quality* (PATO:0000001), if they do not possess any other qualifier than *abnormal*. This annotation cannot be derived automatically and can only be covered with the addition of a special rule. Similar to the mismatches in *qualities* when decomposing MP terms, some errors occur due to certain extension/replacement patterns in the manually assigned EQ statements. For instance, *Irregular epiphysis of the middle phalanx of the 4th finger* (HP:0009219) is manually annotated with *irregular density* (PATO:0002141). Those identified mismatches can be addressed by adding special handling rules to the concept decomposition process.

Mismatches in the structural components of HPO phenotypes were partially due to differences in the naming of the anatomical structure in HPO and FMA. For example, while FMA chose to name fingers (*index finger* or *ring finger*), HPO assigned numbers to fingers (*2nd finger* or *fourth finger*). Moreover, HPO is not consistent with their numbering, e.g. it is *thumb* in all concepts concerning the *first finger* and it is *second toe* versus *2nd finger*. In addition to inconsistent numbering/naming of body parts, the HPO is inconsistent in the usage of singular/plural

7 derived from the textual definition of this concept, see `http://www.human-phenotype-ontology.org/hpoweb/showterm?id=HP:0000248`

forms of nouns, e.g. (*phalanges* versus *phalanx*). Another group of mismatches in anatomical structures is arising from the usage of shortened descriptions as compared to the FMA concept labels, e.g. *premolar* instead of *premolar tooth* or *metatarsal* instead of *metartarsal bone*. Most of the currently identified mismatches in the structural constituents of HPO concepts can be addressed by adding terms to the dictionary of the FMA annotation server.

In the chosen subset, and as partially the case for MP concepts, mismatches in process constituents were due to the current implementation of the GO server not supporting synonyms. For instance, *Abnormality of valine metabolism* (HP:0010914) does not obtain the GO annotation *valine metabolic process* (GO:0006573). This type of mismatches can potentially be corrected by including synonyms in the dictionary underlying the GO annotation server.

One group of mismatches which did not occur in the MP concept name decomposition, is the co-existence of identical concepts in different ontologies. Even though the OBO Foundry (Smith et al., 2007) aims at the orthogonality of the ontologies, this criterion is not fulfilled in all cases. For instance, both FMA and GO possess a concept *Chromosome* (GO:0005694, FMA:67093) (labelled identically). Depending on the understanding and knowledge of the EQ statement developers, it is chosen either for one or the other. Another example of a double existing concepts is *Anosmia* (HP:0000458, PATO:0000817). As those concepts should be removed during the process of quality assessment through the OBO Foundry, no action is required to include this aspect into the decomposition method.

In a number of investigated concepts, the manually assigned EQ statement were supposedly inconsistent. Those inconsistencies were reported to HPO's EQ statements developers. Some inconsistencies were confirmed and corrected by the developers, and the corrections are provided as a new version to the user community.

6.4.3 *Towards a generalised phenotype decomposition*

Even though the decomposition for HPO concepts does not yet work as well as automatically generating EQ statements for MP concepts, the changes required for either ontology are similar. To address most of the mismatches, the underlying annotation server dictionaries need to be extended and special extension rules are required to accommodate for specific patterns in the manually assigned EQ statements.

By addressing those required changes, we expect the performance to significantly improve for both the ontologies.

However, covering the correct set of annotations for 52% of the structural and process phenotypes contained in MP is a promising start to develop an automated method capable of deriving EQ statements from pre-composed phenotype ontologies. Due to the close development of MP's and HPO's EQ statements, the method has to be further tested on other pre-composed phenotype ontologies, such as the Worm Phenotype Ontology (Schindelman et al., 2011), for which also a decomposed subset of concepts exist. Once evaluated on another pre-composed phenotype ontology, we hypothesise that the performances of our suggested method will increase and we will be able to successfully decompose phenotype statements into their constituents.

6.5 CONCLUSIONS

Applying the *EQ-liser* prototype to generate EQ statements from MP concept labels for structural and process phenotypes yields a strictly correct EQ statement in 30% of the cases. Assuming that a curator will approve the EQ statements before they are used community wide, additionally assigned constituents can be easily removed from a correct EQ statement. With that assumption, we can identify the correct subset of constituents for the EQ statements in over 52% of the structural and process phenotypes. To achieve a similar rate for the decomposition of HPO concepts based on their names, the identified problems have to be addressed, which will enable a better identification of the EQ statements from their concept labels. Once the flaws have been corrected the method can be implemented within a prototype to derive EQ statements from pre-composed phenotype statements which will ease the integration of species-specific pre-composed phenotype information into a species-independent framework.

Aside from deriving decomposed phenotype expressions, applying the method to the concept labels also allows for the identification of inconsistencies within the naming of concepts. While MA and MP follow a rigorous naming scheme and hence facilitate the decomposition, HPO and FMA diverge from each other. Furthermore, HPO does not consistently name its own concepts which hinders a decomposition based on lexical attributes, confuses users of the ontology, and prevents easy integration of human data into other frameworks based on a decomposed presentation.

Despite allowing for the decomposition of concepts which have no EQ statement yet, the method has also proven useful to identify flaws in the manually assigned EQ statements. Based on the application of the method, inconsistencies have been identified which could be corrected and hence improve the quality of the existing EQ statements and consequently of all methods applying the EQ statements, such as PhenomeNET and MouseFinder (Chen et al., 2012; Hoehndorf et al., 2011c).

6.6 FUTURE WORK

Correcting the identified required changes when applying the method to MP and HPO concept labels, will be the first step in future work. This mainly means the extension of the underlying dictionaries for constituent recognition. Furthermore, the GO server will be amended to accommodate for the annotation of smaller text spans, not entire concepts labels. This will also allow to filter potential overlap in process constituents (see 6.2.2.2) and consequently reduce noise in the automatically assigned EQ statements.

One aspect of future work is the identification of a method to deal with clinical labels, mostly contained in HPO (see 6.4.2). Clinical terms are one word expressions hiding *entities* and *qualities*. However, the containment of *entities* and *qualities* in term labels is crucial to the flawless application of the approach. One possibility to extend the clinical terms to multi-word expression, is to use another ontology, e.g. MP. MP uses clinical terms mostly as synonyms and multi-term expression as concept labels instead. However, species-specific *entities*, such as anatomical structures, would have to be replaced. Another option would be to investigate the applicability of HPO's textual definitions for terms with single-word labels.

Despite correcting some of the patterns specific to the extraction of EQ statements of HPO and MP concepts, we intend to apply the approach to more species-specific ontologies such as the Worm Phenotype Ontology (Schindelman et al., 2011), to identify further problems and more generalised patterns to allow the integration of all available species-specific phenotype data. We aim to support the easy generation of decomposed phenotype expressions from further potentially evolving pre-composed phenotype ontologies.

A potential helpful extension could be the application of a part-of-speech tagger allowing for the identification of language patterns

inside the pre-composed phenotype concepts. Those patterns can then be used to further eliminate errors in the recognition of *entities* and *qualities*.

Furthermore, the current version of the method only covers structural and process phenotypes, neglecting all other types of phenotypes potentially incorporated inside pre-composed phenotype ontologies. Therefore, we plan to extend the approach to also cover other phenotypes, which requires the support of other types of *entities*, e.g. chemical components described with ChEBI.

Further to extending the coverage of the method, we intend to implement a stable version of the method as a command line tool, which will generate an OBO Flatfile and OWL file containing the descriptions of the phenotype concepts. The OWL file will be based on the phenotype patterns described by Hoehndorf et al. (2010b). This version will also include a validity check with existing EQ statements and support the report of conflicting EQ statements.

7

CONCLUSION

As outlined in 1.5, my work was geared towards the development of a phenotype mining solution that would facilitate the integration of several resources and enable its users to prioritise candidate genes for heritable diseases. The anchor point for the integration of the data repositories as well as the underlying basis for the gene prioritisation, was the available phenotype information. The overall problem has been divided into smaller parts, and each part was individually addressed in a sub-project (see chapter 3 to chapter 6). Once all the individual projects deliver satisfying results, they can be combined to build *InterPhen*. The following section (7.1) summarises the content and main achievements on the smaller projects, while section 7.2 assesses the overall status of *InterPhen*. Section 7.3 illustrates how *InterPhen* could be extended to other use cases.

7.1 SUMMARY OF MAIN ACHIEVEMENTS

Chapter 2 – The plethora of phenotype descriptions

Despite acknowledging the need of standardisation to integrate knowledge across resources and domains (see 1.2), several ways of describing phenotype information exist. In chapter 2, I provided an overview of the existing categories to describe and define phenotype data. Furthermore, I summarised advantages and disadvantages intrinsic to each of the categories with respect to generation and integration.

Chapter 3 – Using mouse models to prioritise genetic causes of human disorders

I obtained mouse genotype and phenotype data from MGD as well as phenotype descriptions (signs and symptoms) for human heritable diseases from OMIM. I integrated both data repositories by aligning both the ontologies used for representing phenotype data: the Mammalian Phenotype Ontology (MP) (Smith et al., 2005) and the Human Phenotype Ontology (HPO) (Robinson et al., 2008). The integrated data were

used to prioritise disease gene candidates by calculating and ranking the semantic phenotype similarity of disease–mouse model pairs. Obtained results were evaluated on currently known gene–disease associations, showing that the suggested approach significantly outperforms existing tools. By using solely phenotype information, the suggested method is particularly valuable for diseases where nothing or only little is known about the molecular mechanisms and the application of "guilt-by association" approaches is impossible.

Chapter 4 – Mouse-specific profiles to annotate genetic diseases

By applying known gene–disease associations from MGD and OMIM as well as gene–phenotype data from MGD, I derived mouse-specific phenotype descriptions (signs and symptoms) for human genetic diseases. It could be shown that these mouse-specific descriptions can extend the existing human-centred disease descriptions by manually investigating a small number of the mouse-specific disease descriptions. Extending the existing human-centred phenotype descriptions bears the potential to achieve a better cross-species integration and, as a consequence, could improve the prioritisation of disease genes. I validated the mouse-specific disease profiles through application in a gene prioritisation task.

Chapter 5 – Mining phenotypes from scientific literature

I proposed two approaches ($\text{Phentomine}_{\text{LEX}}$ and $\text{Phentomine}_{\text{EQ}}$) for the extraction of phenotypes from scientific literature. Both approaches use features of a pre-composed phenotype ontology to extract phenotype information from scientific literature. Both methods $\text{Phentomine}_{\text{LEX}}$ and $\text{Phentomine}_{\text{EQ}}$ have been evaluated against a gold standard. The evaluation of their performance against the gold standard showed that they both need performance improvements in order to be competitive with existing solutions for generalised ontology concept recognition. Finally, I assessed the section-wise distribution of phenotype information in scientific publications. This work could lead into a section-based filter for phenotype information as well as using scientific discourse as *evidence* for phenotype information.

Chapter 6 – "EQ-lising" pre-composed phenotype ontologies

I described a method allowing the decomposition of pre-composed phenotype ontologies into a post-composed presentation. The method is so far limited to the decomposition of structure and process phenotypes in mouse and human. To date, the suggested solution shows best performance for pre-composed mouse phenotype concept labels. However, I suggest modifications to enable species-independent decomposition of pre-composed phenotype concepts. The more post-composed phenotypes are available, the better an integration (in terms of coverage and precision) can be achieved. A better integration would lead to a higher performance of automated cross-species research such as described in chapter 3.

7.2 OUTLOOK: INTERPHEN – THE COMBINED PHENOTYPE MINING SOLUTION

The integration of the different proposed components will lead to a combined phenotype mining solution, called *InterPhen*. This system is inspired by PhenEx (Balhoff et al., 2010), EL Vira (Hoehndorf et al., 2011a), PhenomeNET (Hoehndorf et al., 2011c), and the formalised representation of phenotypes (Mungall et al., 2010; Hoehndorf et al., 2010b). At the moment, however, not all the individual parts are mature enough to be integrated into *InterPhen*. Currently, the biggest bottleneck is the integration with the scientific literature by means of recognising and extracting phenotypes from publications (see 5). In addition, the phenotype extraction was targeted towards mouse phenotypes and would have to be extended to cover other species, e.g. human. Furthermore, the decomposition of phenotypes (see 6) needs improvements before it can facilitate a better integration of resources.

Further to the prioritisation of genes based on integrated resources, the *InterPhen* system would not only provide phenotype information extracted from scientific literature but also allow the immediate integration with existing phenotype resources (based on their formal representation). This would support curators in their efforts and show immediate consequences underlying curation decisions. The extracted phenotype information would be provided with evidence scores obtained from either literature or derived through integration with existing data repositories. This procedure would enable to iden-

tify and report inconsistencies with existing knowledge repositories, which can then be further assessed.

In a way, *InterPhen* will be an interactive PhenomeNET. PhenomeNET covers several species and allows their alignment based on phenotype information. The PhenomeNET interface in its current implementation provides a static view on the data contained in its knowledge base. To interact with the knowledge base, queries can be submitted to PhenomeBLAST. PhenomeBLAST allows to obtain a ranked list of entities (described by their phenotypes), such as diseases or animal models. However, the exchange of data is based on the exchange of files. *InterPhen* would extend this to be a truly interactive system, where extracted phenotypes can be submitted and tested instantly.

Furthermore, *InterPhen* will support reporting mechanisms such as inconsistencies or extensions to existing databases or knowledge bases. This means that, e.g. after the extraction of phenotypes for a mouse models from a paper, the phenotypes can be checked for their validity and either be exported or reported for curation. To export phenotype data, various formats will be supported, such as NeXML (Vos et al., 2012), the OBO Flatfile format and OWL. NeXML is useful for integration with the PhenoScape Knowledgebase[1], while the other formats are supported by e.g. MGD and OMIM. Thus, *InterPhen* will also facilitate the integration with other knowledge bases.

7.3 BEYOND DISEASE GENE PRIORITISATION

With my work, I aimed to support the discovery of novel disease gene candidates by developing and using integration and analysis techniques for phenotype data. However, the application of phenotype data goes beyond human diseases. As illustrated by the PhenoScape Knowledgebase, phenotypes can also give insights to the evolution of species. Following on from PhenoScape's success, the Assembling, Visualizing and Analyzing the Tree Of Life (AVATOL) project[2] aims to facilitate the high-throughput analysis and crowd-sourcing of phenotype data in an evolutionary context. As more and more phenotype data become available, clustering species according to phenotypes may give entirely new insights to evolutionary aspects of species. One of the future avenues of *InterPhen* could be the integration with

1 http://kb.phenoscape.org/
2 http://avatol.org/

PhenoScape or AVATOL and apply evolutionary aspects to the origin of diseases.

Furthermore, despite the acknowledged triangular relationship of genotype and its environment leading to a particular phenotype, little work has been done to systematically gather and analyse environmental data. For most of the large-scale phenotyping projects, the environmental aspects were eliminated by pre-defining the experimental set-up in a way that it is the same for all laboratories performing the mutagenesis experiments. Therefore, another possible direction for *InterPhen* could be the extension to environmental factors. The incorporation of environmental information in *InterPhen* would allow to study the impact of external biotic and abiotic factors on an organism. More importantly, it would also allow to derive environmental components facilitating disease.

A

EXAMPLES FOR CLASSES OF PHENOTYPE DESCRIPTIONS

The following table provides some selected examples of phenotype resources. Phenotype resources are asorted according to the categories introduced and explained in 2.2. The table is not comprehensive, instead it aims to illustrate the variety of existing phenotype resources.

Type	Application examples
free text	OMIM FlyBase OrphaNet
structured vocabulary terminologies	London Dysmorphology Database (LDDB) DECIPHER
pre-composed ontologies	MGD – MP OMIM – HPO Mouse portal – MP Mouse Phenotype database – MP WormBase – WBPhenotype
EQ statements	Morpholino database ZFIN PhenoScape Knowledge base also applied to represent MP and HPO
phenes	decomposition of MP and HPO

B

EXAMPLE OF INPUT DATA FOR GENE PRIORITISATION

The data provided here corresponds to the example given in chapter 4. It covers the annotations for

- the three highest ranked alleles (Gdf6, Marcks, and Vax1)
- Septo-Optic Dysplasia (SOD) (MIM:#182230) and also
- includes original and converted of the disease.

Disease annotations were converted using the combined lexical and ontological mapping for a conversion from Human Phenotype Ontology (HPO) to Mammalian Phenotype Ontology (MP) (see 3.2.4).

VAX1<TM1GRL>

```
MP:0000111 cleft palate
MP:0000819 abnormal olfactory bulb morphology
MP:0001286 abnormal eye development
MP:0001330 abnormal optic nerve morphology
MP:0001332 abnormal optic nerve innervation
MP:0002196 absent corpus callosum
MP:0002653 abnormal ependyma morphology
MP:0002739 abnormal olfactory bulb development
MP:0002961 abnormal axon guidance
MP:0004275 abnormal postnatal subventricular zone morphology
MP:0004277 abnormal lateral ganglionic eminence morphology
MP:0004279 abnormal rostral migratory stream morphology
MP:0005262 coloboma
MP:0009937 abnormal neuron differentiation
MP:0011088 partial neonatal lethality
```

GDF6<TM1LEX>

```
MP:0001262 decreased body weight
MP:0001732 postnatal growth retardation
MP:0001963 abnormal hearing physiology
MP:0002082 postnatal lethality
```

MP:0005508 abnormal skeleton morphology
MP:0008259 abnormal optic disc morphology

MARCKS<TM1PJB>

MP:0000433 microcephaly
MP:0000774 decreased brain size
MP:0000780 abnormal corpus callosum morphology
MP:0000788 abnormal cerebral cortex morphology
MP:0000792 abnormal cortical marginal zone morphology
MP:0000822 abnormal brain ventricle morphology
MP:0000913 abnormal brain development
MP:0000914 exencephaly
MP:0000929 open neural tube
MP:0001265 decreased body size
MP:0001325 abnormal retina morphology
MP:0001436 abnormal suckling behavior
MP:0002185 ectopia
MP:0002199 abnormal brain commissure morphology
MP:0003052 omphalocele
MP:0003104 acrania
MP:0008221 abnormal hippocampal commissure morphology
MP:0008225 abnormal anterior commissure morphology
MP:0011090 partial perinatal lethality

SEPTO-OPTIC DYSPLASIA

HP:0000006 Autosomal dominant inheritance
HP:0000007 Autosomal recessive inheritance
HP:0000609 Optic nerve hypoplasia
HP:0000824 Growth hormone deficiency
HP:0000830 Hypopituitarism
HP:0000873 Diabetes insipidus
HP:0001255 Global developmental delay
HP:0001273 Abnormality of the corpus callosum
HP:0001274 Agenesis of corpus callosum
HP:0001331 Absent septum pellucidum
HP:0001943 Hypoglycemia
HP:0003813 Phenotypic variability
HP:0004322 Short stature
HP:0007890 Optic disc hypoplasia
HP:0009381 Hypoplastic/small fingers

```
MP:0000001 mammalian phenotype
MP:0000188 abnormal circulating glucose level
MP:0000189 hypoglycemia
MP:0000428 abnormal craniofacial morphology
MP:0000432 abnormal head morphology
MP:0000545 abnormal limbs/digits/tail morphology
MP:0000572 abnormal autopod morphology
MP:0000778 abnormal nervous system tract morphology
MP:0000780 abnormal corpus callosum morphology
MP:0000781 decreased corpus callosum size
MP:0000783 abnormal forebrain morphology
MP:0000787 abnormal telencephalon morphology
MP:0000913 abnormal brain development
MP:0000934 abnormal telencephalon development
MP:0000936 small telencephalic vesicles
MP:0001056 abnormal cranial nerve morphology
MP:0001071 abnormal facial nerve morphology
MP:0001253 abnormal body height
MP:0001255 decreased body height
MP:0001259 abnormal body weight
MP:0001262 decreased body weight
MP:0001265 decreased body size
MP:0001286 abnormal eye development
MP:0001325 abnormal retina morphology
MP:0001330 abnormal optic nerve morphology
MP:0001672 abnormal embryogenesis/ development
MP:0001731 abnormal postnatal growth
MP:0001746 abnormal pituitary secretion
MP:0001764 abnormal homeostasis
MP:0001985 abnormal gustatory system physiology
MP:0002078 abnormal glucose homeostasis
MP:0002085 abnormal embryonic tissue morphology
MP:0002089 abnormal postnatal growth/weight/body size
MP:0002092 abnormal eye morphology
MP:0002109 abnormal limb morphology
MP:0002110 abnormal digit morphology
MP:0002152 abnormal brain morphology
MP:0002164 abnormal gland physiology
MP:0002196 absent corpus callosum
MP:0002199 abnormal brain commissure morphology
MP:0002395 hemolymphoid system abnormalities
MP:0002396 abnormal hematopoietic system morphology/development
MP:0002544 brachydactyly
MP:0002752 abnormal somatic nervous system morphology
MP:0002864 abnormal ocular fundus morphology
```

```
MP:0003232 abnormal forebrain development
MP:0003348 hypopituitarism
MP:0003631 nervous system phenotype
MP:0003632 abnormal nervous system morphology
MP:0003633 abnormal nervous system physiology
MP:0003743 abnormal facial morphology
MP:0003861 abnormal nervous system development
MP:0003953 abnormal hormone level
MP:0003956 abnormal body size
MP:0005195 abnormal posterior eye segment morphology
MP:0005266 abnormal metabolism
MP:0005371 limbs/digits/tail phenotype
MP:0005376 homeostasis/metabolism phenotype
MP:0005378 growth/size phenotype
MP:0005379 endocrine/exocrine gland phenotype
MP:0005380 embryogenesis phenotype
MP:0005382 craniofacial phenotype
MP:0005385 cardiovascular system phenotype
MP:0005390 skeleton phenotype
MP:0005391 vision/eye phenotype
MP:0005394 taste/olfaction phenotype
MP:0005397 hematopoietic system phenotype
MP:0005418 abnormal circulating hormone level
MP:0005560 decreased circulating glucose level
MP:0005646 abnormal pituitary gland physiology
MP:0006216 abnormal optic disc size
MP:0006217 small optic disc
MP:0006221 optic nerve hypoplasia
MP:0008026 abnormal brain white matter morphology
MP:0008219 abnormal dorsal telencephalic commissure morphology
MP:0008259 abnormal optic disc morphology
MP:0008540 abnormal cerebrum morphology
MP:0009642 abnormal blood homeostasis
MP:0009772 abnormal retinal development
MP:0010771 integument phenotype
```

C

ADDITIONAL INFORMATION CONCERNING PHENTOMINE

C.1 REMOVAL OF SPECIAL CHARACTERS

The special characters were defined by manually investigating a small subset of abstracts and full text papers. As special characters are considered:

,
<
>
{
}
(
)
[
]
/
\
%
+
:
;
.
'
*
‘‘
’’
-
_

C.2 STOP WORDS LIST

The stop words were defined by manually investigating the .tbl files of Mammalian Phenotype Ontology (MP) and Human Phenotype Ontology (HPO). As stop words in phenotype ontologies are considered:

- and

- or
- with
- the
- of
- i
- ii
- iiiv
- VI
- VII
- VIII
- V
- XI

C.3 SECTION TITLES IN PMC SUBSET

Based on the 621 documents contained in our *gold standard* for evaluation of our phenotype extraction method, we identified section titles in full text papers (see 5.3.3). We retrieved 76 different section titles which were manually reduced to 9. The mapping of the section titles to the *unified* sections is provided in the following table.

unified section name	obtained section name
intro	1. Introduction, Background, CONTEXT AND CAVEATS, INTRODUCTION, Introduction
mm	2. Materials and Methods, 2. Methods, EXPERIMENTAL, EXPERIMENTAL PROCEDURES, Experimental Procedures, Experimental procedures, MATERIAL AND METHODS, MATERIALS AND METHODS, METHODS, Material and Methods, Material and methods, Materials And Methods, Materials and Methods, Materials and methods, Methods, Methods and Materials, Methods and materials, RESEARCH DESIGN AND METHODS
res	Results, RESULTS, Results
res_dis	Results and Discussion, RESULTS AND DISCUSSION, Results And Discussion, Results and Discussion, Results and discussion, Results/Discussion
dis	Discussion, DISCUSSION, Discussion, Discussions
concl	Conclusions, Conclusion, Conclusions
dis_concl	Discussion and Conclusions, Discussion and conclusion
addf	Additional Files, Additional data files, SUPPLEMENTARY DATA, SUPPLEMENTARY MATERIAL, Supplemental Data, Supplemental Information, Supplementary Material, Supplementary material, Supporting Information
other	Abbreviations, Acknowledgements, Appendix, Author summary, Author's contributions, Authors contributions, Authors' Contributions, Authors' contribution, Authors' contributions, Authors' information, Availability & requirements, Competing interests, Conflict of Interest Statement, Dissection, FUNDING, Funding, Immunohistochemistry, Lack of Whirlin Affects Actin Filament Elongation and Packing, Lack of Whirlin Affects the Stereocilia of the Inner and Outer Hair Cells in Different Ways, List of Abbreviations, List of abbreviations, List of abbreviations used, Pre-publication history, The Whirler Phenotype is Different From That of Shaker2 URL, mRNA In situ hybridyzation

D

PUBLICATIONS

D.1 IN PREPARATION

- Mouse-specific disease profiles as a guide to human disease

D.2 SUBMITTED

- EQ-liser: Automated decomposition of pre-composed phenotype ontologies. **Anika Oellrich**, Christoph Grabmueller, and Dietrich Rebholz-Schuhmann. 4th Ontologies in Biomedicine and Life sciences (OBML) workshop.

D.3 ACCEPTED

- *Exploring the scientist's literature to find biomedical assertions, facts and knowledge*. Dietrich Rebholz-Schuhmann, **Anika Oellrich**, and Robert Hoehndorf. Nature Genetics.

- *Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology* **Anika Oellrich**, Robert Hoehndorf, Georgious V. Gkoutos and Dietrich Rebholz-Schuhmann. JBMS – Special Issue.

D.4 JOURNAL PUBLICATIONS

2012 *Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases* **Anika Oellrich**, Robert Hoehndorf, Georgious V. Gkoutos and Dietrich Rebholz-Schuhmann. JBMS.

2011 *Ontology design patterns to disambiguate relations between genes and gene products in GENIA.* Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo, Sampo Pyysalo, Tomoko Ohta, **Anika Oellrich** and Dietrich Rebholz-Schuhmann. Special Issue, Journal of Biomedical Semantics.

2011 *Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning.* Robert Hoehndorf, Michel Dumontier, **Anika Oellrich**, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. PLoS ONE.

2011 *A common layer of interoperability for biomedical ontologies based on OWL EL.* Robert Hoehndorf, Michel Dumontier, **Anika Oellrich**, Sarala Wimalarante, Dietrich Rebholz-Schuhmann, Paul N. Schofield, and Georgios V. Gkoutos. BioInformatics.

2010 *Interoperability between phenotype and anatomy ontologies.* Robert Hoehndorf, **Anika Oellrich** and Dietrich Rebholz-Schuhmann. BioInformatics.

2010 *Relations as patterns: bridging the gap between OBO and OWL.* Robert Hoehndorf, **Anika Oellrich**, Michel Dumontier, Janet Kelso, Heinrich Herre and Dietrich Rebholz-Schuhmann. BMC Bioinformatics.

D.5 CONFERENCE AND WORKSHOP CONTRIBUTIONS

2011 *Quantitative comparison of mapping methods between Human and Mammalian Phenotype Ontology* **Anika Oellrich**, Robert Hoehndorf, Georgious V. Gkoutos and Dietrich Rebholz-Schuhmann. 3rd Ontologies in Biomedicine and Life sciences (OBML) workshop, Berlin.

2010 *Applying ontology design patterns to the implementation of relations in GENIA.* Robert Hoehndorf, Axel-Cyrille Ngonga Ngomo, Sampo Pyysalo, Tomoko Ohta, **Anika Oellrich** and Dietrich Rebholz-Schuhmann. 4th International Symposium on Semantic Mining in Biomedicine (SMBM).

2010 *A classification of existing phenotypical representations and methods for improvement.* **Anika Oellrich** and Dietrich Rebholz-Schuhmann. 2nd Ontologies in Biomedicine and Life sciences (OBML) workshop, Mannheim.

2010 *The Ontology of Primary Immunodeficiency Diseases (PIDs) – Using PIDs to Rethink the Ontology of Phenotypes* Nico Adams, Christian Hennig, Robert Hoehndorf, **Anika Oellrich**, Dietrich Rebholz-Schuhmann and Gesine Hansen. 2nd Ontologies in Biomedicine and Life sciences (OBML) workshop, Mannheim.

2010 *Relational patterns in OWL and their application to OBO.* Robert Hoehndorf, **Anika Oellrich**, Michel Dumontier, Janet Kelso, Heinrich Herre and Dietrich Rebholz-Schuhmann. OWL: Experiences and Directions (OWLED).

2010 *OWLDEF: Integrating OBO and OWL.* Robert Hoehndorf, **Anika Oellrich**, Michel Dumontier, Janet Kelso, Heinrich Herre and Dietrich Rebholz-Schuhmann. Bio-Ontologies.

BIBLIOGRAPHY

A. Abbott. Mouse project to find each gene's role. *Nature*, 465(7297): 410, May 2010. (Cited on pages 9, 35, and 72.)

T. Adamusiak, T. Burdett, N. Kurbatova, K. J. van der Velde, N. Abeygunawardena, D. Antonakaki, M. Kapushesky, H. Parkinson, and M. A. Swertz. Ontocat–simple ontology search and integration in java, r and rest/javascript. *BMC Bioinformatics*, 12:218, Jan 2011. (Cited on page 69.)

R. Agrawal, A. Ailamaki, P. A. Bernstein, E. A. Brewer, M. J. Carey, S. Chaudhuri, A. Doan, D. Florescu, M. J. Franklin, H. Garcia-Molina, J. Gehrke, L. Gruenwald, L. M. Haas, A. Y. Halevy, J. M. Hellerstein, Y. E. Ioannidis, H. F. Korth, D. Kossmann, S. Madden, R. Magoulas, B. C. Ooi, T. O'Reilly, R. Ramakrishnan, S. Sarawagi, M. Stonebraker, A. S. Szalay, and G. Weikum. The claremont report on database research. *SIGMOD Record*, 2008. (Cited on page 15.)

B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, and X. Wang. Assisted curation: does text mining really help? *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 556–67, Jan 2008. (Cited on pages 90 and 91.)

J. Amberger, C. Bocchini, and A. Hamosh. A new face and new challenges for online mendelian inheritance in man (omim®). *Human mutation*, 32(5):564–7, May 2011. (Cited on pages 1, 3, 4, 5, 36, 43, 45, 71, and 89.)

M. E. Aranguren, C. Wroe, C. Goble, and R. Stevens. In situ migration of handcrafted ontologies to reason-able forms. *Data & Knowledge ...*, 66:147–162, 2008. (Cited on page 118.)

A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, 17(3):229–36, May 2010. (Cited on pages 21 and 28.)

M. Asai-Coakwell, C. R. French, K. M. Berry, M. Ye, R. Koss, M. Somerville, R. Mueller, V. van Heyningen, A. J. Waskiewicz,

and O. J. Lehmann. Gdf6, a novel locus for a spectrum of ocular developmental anomalies. *Am J Hum Genet*, 80(2):306–15, Feb 2007. (Cited on page 64.)

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. 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. the gene ontology consortium. *Nat Genet*, 25(1):25–9, May 2000. (Cited on pages 11, 39, 52, 92, and 119.)

C. P. Austin, J. F. Battey, A. Bradley, M. Bucan, M. Capecchi, F. S. Collins, W. F. Dove, G. Duyk, S. Dymecki, J. T. Eppig, F. B. Grieder, N. Heintz, G. Hicks, T. R. Insel, A. Joyner, B. H. Koller, K. C. K. Lloyd, T. Magnuson, M. W. Moore, A. Nagy, J. D. Pollock, A. D. Roses, A. T. Sands, B. Seed, W. C. Skarnes, J. Snoddy, P. Soriano, D. J. Stewart, F. Stewart, B. Stillman, H. Varmus, L. Varticovski, I. M. Verma, T. F. Vogt, H. von Melchner, J. Witkowski, R. P. Woychik, W. Wurst, G. D. Yancopoulos, S. G. Young, and B. Zambrowicz. The knockout mouse project. *Nat Genet*, 36(9):921–4, Sep 2004. (Cited on page 64.)

O. T. Avery. A further study on the biologic classification of pneumococci. *J Exp Med*, 22(6):804–19, Dec 1915. (Cited on page 7.)

S. Aymé. Orphanet, an information site on rare diseases. *Soins*, (672): 46–7, Jan 2003. (Cited on page 4.)

S. Aymé and J. Schmidtke. Networking for rare diseases: a necessity for europe. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 50(12):1477–83, Dec 2007. (Cited on pages 4 and 36.)

A. Bairoch. The future of annotation/biocuration. *Nature Precedings*, 2009. (Cited on page 14.)

J. P. Balhoff, W. M. Dahdul, C. R. Kothari, H. Lapp, J. G. Lundberg, P. M. Mabee, P. E. Midford, M. Westerfield, and T. J. Vision. Phenex: Ontological annotation of phenotypic diversity. *PLoS ONE*, 2010. (Cited on pages 115 and 135.)

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 Res*, 37(Database issue):D396–403, Jan 2009. (Cited on page 13.)

J. Beckers, W. Wurst, and M. H. de Angelis. Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling. *Nat Rev Genet*, 10(6):371–80, Jun 2009. (Cited on page 7.)

K. Bharti, M. Gasper, S. Bertuzzi, and H. Arnheiter. Lack of the ventral anterior homeodomain transcription factor vax1 leads to induction of a second pituitary. *Development*, 138(5):873–8, Mar 2011. (Cited on page 64.)

N. Bhatia, N. H. Shah, D. Rubin, A. P. Chiang, and M. A. Musen. Comparing concept recognizers for ontology-based indexing: Mgrep vs. metamap. *AMIA Proceedings 2009*, 2009. (Cited on page 29.)

A. Bies, M. Ferguson, K. Katz, R. MacIntyre, V. Tredinnick, G. Kim, M. A. Marcinkiewicz, and B. Schasberger. Bracketing guidelines for treebank ii style penn treebank project. 1995. (Cited on page 25.)

R. M. Bilder, F. W. Sabb, T. D. Cannon, E. D. London, J. D. Jentsch, D. S. Parker, R. A. Poldrack, C. Evans, and N. B. Freimer. Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience*, 164(1):30–42, Nov 2009. (Cited on page 3.)

J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, J. T. Eppig, and M. G. D. Group. The mouse genome database (mgd): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*, 39(Database issue):D842–8, Jan 2011. (Cited on pages 3, 9, 35, 43, 45, 46, 72, and 89.)

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue): D267–70, Jan 2004. (Cited on pages 10, 15, and 21.)

O. Bodenreider and R. Stevens. Bio-ontologies: current trends and future directions. *Brief Bioinformatics*, 7(3):256–74, Sep 2006. (Cited on page 12.)

M. Bogue. Mouse phenome project: understanding human biology through mouse genetics and genomics. *J Appl Physiol*, 95(4):1335–7, Oct 2003. (Cited on page 6.)

M. A. Bogue and S. C. Grubb. The mouse phenome project. *Genetica*, 122(1):71–4, Sep 2004. (Cited on page 6.)

M. Brameier and C. Wiuf. Co-clustering and visualization of gene expression data and gene ontology terms for saccharomyces cerevisiae

using self-organizing maps. *Journal of biomedical informatics*, 40(2): 160–73, Apr 2007. (Cited on page 19.)

H. G. Brunner and M. A. van Driel. From syndrome families to functional genomics. *Nat Rev Genet*, 5(7):545–51, Jul 2004. (Cited on pages 5 and 7.)

S. Buono, F. Scannella, and M. B. Palmigiano. Self-injurious behavior: A comparison between prader-willi syndrome, down syndrome and autism. *Life Span and Disability XIII*, 2010. (Cited on page 85.)

M. N. Cantor and Y. A. Lussier. Mining omim for insight into complex diseases. *Stud Health Technol Inform*, 107(Pt 2):753–7, Jan 2004. (Cited on page 20.)

B. Carpenter. Lingpipe for 99.99 *Proceedings of the 2nd BioCreative workshop*, 2007. (Cited on pages 29, 97, and 123.)

W. W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34 (5):301–310, Oct. 2001. (Cited on page 114.)

W. W. Chapman, M. Fiszman, J. N. Dowling, B. E. Chapman, and T. C. Rindflesch. Identifying respiratory findings in emergency department reports for biosurveillance using metamap. *Stud Health Technol Inform*, 107(Pt 1):487–91, Jan 2004. (Cited on page 29.)

C.-K. Chen, C. J. Mungall, G. V. Gkoutos, S. C. Doelken, S. Köhler, B. J. Ruef, C. Smith, M. Westerfield, P. N. Robinson, S. E. Lewis, P. N. Schofield, and D. Smedley. Mousefinder: candidate disease genes from mouse phenotype data. *Human mutation*, Feb 2012. (Cited on pages 22, 46, 59, 66, 71, 89, 117, and 131.)

K. Cohen, H. Johnson, and K. Verspoor. . . . The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC . . .*, Jan 2010. (Cited on pages 91 and 113.)

R. Cohen, A. Gefen, M. Elhadad, and O. S. Birk. Csi-omim–clinical synopsis search in omim. *BMC Bioinformatics*, 12:65, Jan 2011. (Cited on page 28.)

S. Cohrs. Sleep disturbances in patients with schizophrenia : impact and effect of antipsychotics. *CNS Drugs*, 22(11):939–62, Jan 2008. (Cited on page 85.)

I. Colgiu. An assessment of feature sets for the identification of gene-disease associations from the biomedical literature. *Master Thesis*, 2012. (Cited on page 96.)

F. S. Collins, R. H. Finnell, J. Rossant, and W. Wurst. A new partner for the international knockout mouse consortium. *Cell*, 129(2):235, Apr 2007. (Cited on pages 6 and 35.)

. G. P. Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, Oct 2010a. (Cited on page 8.)

U. Consortium. The universal protein resource (uniprot) in 2010. *Nucleic Acids Res*, 38(Database issue):D142–8, Jan 2010b. (Cited on pages 11 and 13.)

M. Corpas, E. Bragin, S. Clayton, P. Bevan, and H. V. Firth. Interpretation of genomic copy number variants using decipher. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]*, Chapter 8:Unit8.14, Jan 2012. (Cited on page 39.)

I. Cruz, F. P. Antonelli, and C. Stroe. Agreementmaker: Efficient matching for large real-world schemas and ontologies. *Journal Proceedings of the VLDB Endowment 2009*, pages 1–4, Jun 2009. (Cited on pages 16 and 69.)

M. Dai, N. H. Shah, W. Xuan, M. A. Musen, Stanley, J. Watson, B. Athey, and F. Meng. An efficient solution for mapping free text to ontology terms. *Proceedings AMIA Summits on Translational Bioinformatics*, 2008. (Cited on page 28.)

M. T. Dattani, J. P. Martinez-Barbera, P. Q. Thomas, J. M. Brickman, R. Gupta, I. L. Mårtensson, H. Toresson, M. Fox, J. K. Wales, P. C. Hindmarsh, S. Krauss, R. S. Beddington, and I. C. Robinson. Mutations in the homeobox gene hesx1/hesx1 associated with septo-optic dysplasia in human and mouse. *Nat Genet*, 19(2):125–33, Jun 1998. (Cited on page 64.)

J. Day-Richter, M. A. Harris, M. Haendel, G. O. O.-E. W. Group, and S. Lewis. Obo-edit–an ontology editor for biologists. *Bioinformatics*, 23(16):2198–200, Aug 2007. (Cited on page 10.)

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological

interest. *Nucleic Acids Res*, 36(Database issue):D344–50, Jan 2008. (Cited on page 53.)

K. M. Dipple and E. R. McCabe. Phenotypes of patients with "simple" mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *Am J Hum Genet*, 66(6):1729–35, Jun 2000. (Cited on page 8.)

E. Dolgin. Mouse library set to be knockout. *Nature*, 474(7351):262–3, Jun 2011. (Cited on page 1.)

A. Doms and M. Schroeder. Gopubmed: exploring pubmed with the gene ontology. *Nucleic Acids Res*, 33(Web Server issue):W783–6, Jul 2005. (Cited on page 23.)

K. G. Dowell, M. S. McAndrews-Hill, D. P. Hill, H. J. Drabkin, and J. A. Blake. Integrating text mining into the mgi biocuration workflow. *Database (Oxford)*, 2009:bap019, Jan 2009. (Cited on pages 90, 100, and 107.)

R. Drysdale and F. Consortium. Flybase : a database for the drosophila research community. *Methods Mol Biol*, 420:45–59, Jan 2008. (Cited on page 9.)

O. Espinosa and J. M. Hancock. A gene-phenotype network for the laboratory mouse and its implications for systematic phenotyping. *PLoS ONE*, 6(5):e19693, Jan 2011. (Cited on pages 2 and 68.)

J. Euzenat, A. Ferrara, W. R. van Hague, and L. Hollink. Final results of the Ontology Alignment Evaluation Initiative 2011. *Proceedings 6th ISWC workshop on ontology matching (OM)*, 2011. (Cited on page 15.)

H. V. Firth, S. M. Richards, A. P. Bevan, S. Clayton, M. Corpas, D. Rajan, S. V. Vooren, Y. Moreau, R. M. Pettett, and N. P. Carter. Decipher: Database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am J Hum Genet*, 84(4):524–33, Apr 2009. (Cited on pages 37 and 39.)

P. Flicek, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A. K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier,

M. Schuster, D. Sobral, Y. A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Harrow, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S. M. J. Searle. Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–90, Jan 2012. (Cited on page 21.)

N. Freimer and C. Sabatti. The human phenome project. *Nat Genet*, 34(1):15–21, May 2003. (Cited on pages 5 and 6.)

K. W. Fung, O. Bodenreider, A. R. Aronson, W. T. Hole, and S. Srinivasan. Combining lexical and semantic methods of inter-terminology mapping using the umls. *Stud Health Technol Inform*, 129(Pt 1):605–9, Jan 2007. (Cited on page 54.)

S. Gaudan, A. J. Yepes, V. Lee, and D. Rebholz-Schuhmann. Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. *EURASIP journal on bioinformatics & systems biology*, page 342746, Jan 2008. (Cited on pages xi, 90, 93, 94, 95, 121, and 123.)

A. Ghazvinian, N. F. Noy, and M. A. Musen. Creating mappings for ontologies in biomedicine: simple methods work. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, 2009: 198–202, Jan 2009. (Cited on pages 15 and 50.)

J. Gillis and P. Pavlidis. The impact of multifunctional genes on "guilt by association" analysis. *PLoS ONE*, 2011. (Cited on page 18.)

G. V. Gkoutos, C. Mungall, S. Dolken, M. Ashburner, S. Lewis, J. Hancock, P. Schofield, S. Kohler, and P. N. Robinson. Entity/quality-based logical definitions for the human skeletal phenome using pato. *Conf Proc IEEE Eng Med Biol Soc*, 1:7069–72, Jan 2009. (Cited on pages 41, 50, 53, and 92.)

G. V. Gkoutos, P. N. Schofield, and R. Hoehndorf. Computational tools for comparative phenomics: the role and promise of ontologies. *Mammalian genome : official journal of the International Mammalian Genome Society*, Jul 2012. (Cited on pages 2, 20, and 117.)

C. Golbreich, S. Zhang, and O. Bodenreider. The foundational model of anatomy in OWL: Experience and perspectives. *Web semantics (Online)*, 4(3):181–195, 2006. (Cited on page 53.)

B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler. Owl 2: The next step for owl. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2008. (Cited on page 10.)

P. Groth and B. Weiss. Phenotype data: a neglected resource in biomedical research. *Current Bioinformatics*, Jan 2006. (Cited on page 2.)

P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H.-D. Pohlenz, and B. Weiss. Phenomicdb: a new cross-species genotype/phenotype resource. *Nucleic Acids Res*, 35(Database issue):D696–9, Jan 2007. (Cited on page 2.)

P. Groth, U. Leser, and B. Weiss. Phenotype mining for functional genomics and gene discovery. *Methods Mol Biol*, 760:159–73, Jan 2011. (Cited on page 38.)

T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 1993. (Cited on page 10.)

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 1995. (Cited on page 9.)

S. S. Guest, C. D. Evans, and R. M. Winter. The online london dysmorphology database. *Genet Med*, 1(5):207–12, Jan 1999. (Cited on page 38.)

X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 22(8):967–73, Apr 2006. (Cited on page 20.)

J. B. Hagen. The origins of bioinformatics. *Nat Rev Genet*, pages 1–6, Nov 2000. (Cited on page 1.)

A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33(Database issue):D514–7, Jan 2005. (Cited on page 43.)

J. M. Hancock and A.-M. Mallon. Phenobabelomics–mouse phenotype data resources. *Brief Funct Genomic Proteomic*, 6(4):292–301, Dec 2007. (Cited on pages 43 and 99.)

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, Apr 1982. (Cited on pages 28, 59, and 63.)

T. F. Hayamizu, M. Mangan, J. P. Corradi, J. A. Kadin, and M. Ringwald. The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Genome Biol*, 6(3):R29, Jan 2005. (Cited on pages 41, 52, 92, and 119.)

L. Hirschman, G. A. P. C. Burns, M. Krallinger, C. Arighi, K. B. Cohen, A. Valencia, C. H. Wu, A. Chatr-Aryamontri, K. G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey, T. Wiegers, and A. G. Winter. Text mining for the biocuration workflow. *Database (Oxford)*, 2012: bas020, Jan 2012. (Cited on pages 3 and 90.)

R. Hoehndorf, A. Oellrich, M. Dumontier, J. Kelso, D. Rebholz-Schuhmann, and H. Herre. Relations as patterns: bridging the gap between obo and owl. *BMC Bioinformatics*, 11:441, Jan 2010a. (Cited on page 11.)

R. Hoehndorf, A. Oellrich, and D. Rebholz-Schuhmann. Interoperability between phenotype and anatomy ontologies. *Bioinformatics*, 26 (24):3112–8, Dec 2010b. (Cited on pages 36, 41, 42, 120, 132, and 135.)

R. Hoehndorf, M. Dumontier, A. Oellrich, S. Wimalaratne, D. Rebholz-Schuhmann, P. Schofield, and G. V. Gkoutos. A common layer of interoperability for biomedical ontologies based on owl el. *Bioinformatics*, Feb 2011a. (Cited on pages 53 and 135.)

R. Hoehndorf, A.-C. N. Ngomo, S. Pyysalo, T. Ohta, A. Oellrich, and D. Rebholz-Schuhmann. Ontology design patterns to disambiguate relations between genes and gene products in genia. *Journal of Biomedical Semantics*, 2 Suppl 5:S1, Jan 2011b. (Cited on page 25.)

R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res*, 39(18):e119, Oct 2011c. (Cited on pages 2, 22, 46, 51, 59, 66, 71, 89, 117, 131, and 135.)

P. Hogeweg. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*, 7(3):e1002021, Mar 2011. (Cited on pages 1 and 2.)

I. Horrocks. Obo flat file format syntax and semantics and mapping to owl web ontology language. *Technical Report*, 2007. (Cited on page 10.)

W. Hu and Y. Qu. Falcon-ao: A practical ontology matching system. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2008. (Cited on page 16.)

V. M. INGRAM. Gene evolution and the haemoglobins. *Nature*, 189: 704–8, Mar 1961. (Cited on page 1.)

R. Jacob, A. Chowdhury, S. Kamath, and J. Ganesan. Psychiatric comorbidity in prader-willi syndrome a case series. *Reprinted from the German Journal of Psychiatry*, 2009. (Cited on page 85.)

S. Jaeger, S. Gaudan, U. Leser, and D. Rebholz-Schuhmann. Integrating protein-protein interactions and text mining for protein function prediction. *BMC Bioinformatics*, 9 Suppl 8:S2, Jan 2008. (Cited on pages 13, 14, and 89.)

L. J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7 (2):119–29, Feb 2006. (Cited on pages 3 and 22.)

A. Jimeno, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9 Suppl 3:S3, Jan 2008. (Cited on pages 28 and 90.)

C. Jonquet, N. H. Shah, and M. A. Musen. The open biomedical annotator. *Summit on Translat Bioinforma*, 2009:56–60, Jan 2009. (Cited on pages 3, 28, 29, 90, and 100.)

Y. Kazakov. Consequence-driven reasoning for horn shiq ontologies. *Proceedings of the 21st International Conference on Artificial Intelligence*, 2009. (Cited on page 53.)

Y. Kazakov, M. Krötzsch, and F. Simančík. Elk: a reasoner for owl el ontologies. *System Description*, 2012. (Cited on page 79.)

J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpus–semantically annotated corpus for bio-textmining. *Bioinformatics*, 19 Suppl 1: i180–2, Jan 2003. (Cited on page 25.)

J.-D. Kim, T. Ohta, and J. Tsujii. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9:10, Jan 2008. (Cited on page 25.)

R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E.

Whelan, and A. Clare. The automation of science. *Science*, 324(5923): 85–9, Apr 2009. (Cited on page 24.)

S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85(4):457–64, Oct 2009. (Cited on page 43.)

J. O. Korbel, T. Doerks, L. J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S. D. Hooper, M. A. Andrade, and P. Bork. Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol*, 3(5):e134, Jan 2005. (Cited on pages 3, 22, 28, 71, and 90.)

M. Krallinger, A. Valencia, and L. Hirschman. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*, 9 Suppl 2:S8, Jan 2008. (Cited on pages 14 and 89.)

K. Lage, E. O. Karlberg, Z. M. Størling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–16, Mar 2007. (Cited on pages 2 and 21.)

S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, and L. Hunter. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol*, 5(3):e1000215, Mar 2009. (Cited on page 2.)

R. Leaman and G. Gonzalez. Banner: an executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 652–63, Jan 2008. (Cited on page 24.)

I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res*, May 2011. (Cited on page 45.)

S. Leonelli and R. A. Ankeny. Re-thinking organisms: The impact of databases on model organism biology. *Stud Hist Philos Biol Biomed Sci*, 43(1):29–36, Mar 2012. (Cited on pages 9 and 35.)

J. Li, J. Tang, Y. Li, and Q. Luo. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and Data Engineering*, 2009. (Cited on page 16.)

M. Liakata, S. Saha, S. Dobnik, C. Batchelor, and D. Rebholz-Schuhmann. Automatic recognition of conceptualisation zones in scientific articles and two life science applications. *Bioinformatics*, Feb 2012. (Cited on page 103.)

T. S. Lisse, F. Thiele, H. Fuchs, W. Hans, G. K. H. Przemeck, K. Abe, B. Rathkolb, L. Quintanilla-Martinez, G. Hoelzlwimmer, M. Helfrich, E. Wolf, S. H. Ralston, and M. H. de Angelis. Er stress-mediated apoptosis in a new mouse model of osteogenesis imperfecta. *PLoS Genet*, 4(2):e7, Feb 2008. (Cited on pages 5 and 72.)

I. Lobo. Pleiotropy: One gene can affect multiple traits. *Nature Education*, 2008. (Cited on page 7.)

P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10): 1275–83, Jul 2003. (Cited on page 19.)

K. L. Lunetta. Genetic association studies. *Circulation*, 2008. (Cited on page 16.)

Y. A. Lussier and Y. Liu. Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc*, 4(1):18–25, Jan 2007. (Cited on page 2.)

P. M. Mabee, M. Ashburner, Q. Cronk, G. V. Gkoutos, M. Haendel, E. Segerdell, C. Mungall, and M. Westerfield. Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol (Amst)*, 22(7):345–50, Jul 2007. (Cited on page 119.)

A. Magi, L. Tattini, M. Benelli, B. Giusti, R. Abbate, and S. Ruffo. Wnp: A novel algorithm for gene products annotation from weighted functional networks. *PLoS ONE*, 7(6):e38767, Jan 2012. (Cited on page 14.)

D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq. Gotoolbox: functional analysis of gene datasets based on gene ontology. *Genome Biol*, 5(12):R101, Jan 2004. (Cited on page 11.)

K. L. McGary, T. J. Park, J. O. Woods, H. J. Cha, J. B. Wallingford, and E. M. Marcotte. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci USA*, 107(14):6544–9, Apr 2010. (Cited on page 2.)

S. Meystre and P. J. Haug. Evaluation of medical problem extraction from electronic clinical documents using metamap transfer (mmtx). *Stud Health Technol Inform*, 116:823–8, Jan 2005. (Cited on page 29.)

M. Mistry and P. Pavlidis. Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327, Jan 2008. (Cited on pages 20 and 21.)

R. Morante and E. Blanco. *SEM 2012 shared task: resolving the scope and focus of negation. *Proceedings of the First Joint Conference on . . .*, pages 265–274, June 2012. (Cited on page 114.)

D. A. Moreira and M. A. Musen. Obo to owl: a protege owl tab to read/save obo ontologies. *Bioinformatics*, 23(14):1868–70, Jul 2007. (Cited on page 11.)

H. Morgan, T. Beck, A. Blake, H. Gates, N. Adams, G. Debouzy, S. Leblanc, C. Lengger, H. Maier, D. Melvin, H. Meziane, D. Richardson, S. Wells, J. White, J. Wood, T. E. Consortium, M. de Angelis, S. Brown, J. Hancock, and A. Mallon. Europhenome: a repository for high-throughput mouse phenotyping data. *Nucleic Acids Res*, Nov 2009. (Cited on pages 1 and 45.)

C. J. Mungall. Obol: integrating language and meaning in bio-ontologies. *Comparative and functional genomics*, 5(6-7):509–20, Jan 2004. (Cited on page 118.)

C. J. Mungall, G. V. Gkoutos, C. L. Smith, M. A. Haendel, S. E. Lewis, and M. Ashburner. Integrating phenotype ontologies across multiple species. *Genome Biol*, 11(1):R2, Jan 2010. (Cited on pages 21, 36, 40, 50, 53, 96, 117, 119, and 135.)

C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome Biol*, 13(1):R5, Jan 2012. (Cited on pages 52 and 92.)

J. H. Nadeau. Modifier genes in mice and humans. *Nat Rev Genet*, 2 (3):165–74, Mar 2001. (Cited on page 8.)

N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. A. Musen. Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*, 37(Web Server issue):W170–3, Jul 2009. (Cited on pages 11, 29, and 100.)

A. Oellrich and D. Rebholz-Schuhman. A classification of existing phenotypical representations and methods for improvement. *Proceedings 2nd Ontologies in Biomedicine and Life Sciences workshop.*, 2010. (Cited on pages 6, 36, and 117.)

A. Oellrich, R. Hoehndorf, G. V. Gkoutos, and D. Rebholz-Schuhmann. Improving disease gene prioritization by comparing the semantic similarity of phenotypes in mice with those of human diseases. *PLoS ONE*, 7(6):e38937, Jan 2012. (Cited on page 46.)

P. Ogren, K. B. Cohen, G. K. Acquaah-Mensah, J. Eberlein, and L. Hunter. THE COMPOSITIONAL STRUCTURE OF GENE ONTOLOGY TERMS. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 214, 2004. (Cited on page 118.)

P. V. Ogren, K. B. Cohen, and L. Hunter. *Implications of Compositionality in the Gene Ontology for Its Curation and Usage*. *Proceedings of Pacific Symposium on Biocumputing 2005*, 2005. (Cited on page 118.)

M. Oti, M. A. Huynen, and H. G. Brunner. The biological coherence of human phenome databases. *Am J Hum Genet*, 85(6):801–8, Dec 2009. (Cited on pages 3, 22, 71, 76, 88, and 89.)

C. A. Ouzounis and A. Valencia. Early bioinformatics: the birth of a discipline–a personal view. *Bioinformatics*, 19(17):2176–90, Nov 2003. (Cited on page 1.)

K. Paigen and J. T. Eppig. A mouse phenome project. *Mammalian genome : official journal of the International Mammalian Genome Society*, 11(9):715–7, Sep 2000. (Cited on page 6.)

H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag, I. Emam, A. Farne, E. Hastings, E. Holloway, N. Kurbatova, M. Lukk, J. Malone, R. Mani, E. Pilicheva, G. Rustici, A. Sharma, E. Williams, T. Adamusiak, M. Brandizi, N. Sklyar, and A. Brazma. Arrayexpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, 39(Database issue):D1002–4, Jan 2011. (Cited on page 13.)

B. Percha, Y. Garten, and R. B. Altman. Discovery and explanation of drug-drug interactions via text mining. *Pacific Symposium on Biocomputing*, 2012. (Cited on page 24.)

C. Pesquita, D. Faria, H. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9 Suppl 5:S4, Jan 2008. (Cited on page 11.)

C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7):e1000443, Jul 2009. (Cited on pages 14, 15, 19, 20, 21, and 69.)

L. Philips. The double metaphone search algorithm. *C/C++ Users Journal*, 2000. (Cited on page 16.)

M. F. Porter. An algorithm for suffix stripping. *Program*, 1980. (Cited on pages 94, 95, and 122.)

W. Pratt and M. Yetisgen-Yildiz. A study of biomedical concept identification: Metamap vs. people. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, pages 529–33, Jan 2003. (Cited on page 29.)

J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*, 11(20):2417–23, Oct 2002. (Cited on page 4.)

A. Rath, A. Olry, F. Dhombres, M. M. Brandt, B. Urbero, and S. Ayme. Representation of rare diseases in health information systems: the orphanet approach to serve a wide range of end users. *Human mutation*, 33(5):803–8, May 2012. (Cited on pages 4 and 71.)

S. Raychaudhuri. Mapping rare and common causal alleles for complex human diseases. *Cell*, 147(1):57–69, Sep 2011. (Cited on page 1.)

D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–8, Jan 2008. (Cited on page 90.)

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence(IJCAI)*, 1995. (Cited on pages 19 and 21.)

E. Roberts, D. J. Hampshire, L. Pattison, K. Springell, H. Jafri, P. Corry, J. Mannon, Y. Rashid, Y. Crow, J. Bond, and C. G. Woods. Autosomal

recessive primary microcephaly: an analysis of locus heterogeneity and phenotypic variation. *J Med Genet*, 39(10):718–21, Oct 2002. (Cited on page 7.)

P. N. Robinson and S. Mundlos. The human phenotype ontology. *Clin Genet*, 77(6):525–34, Jun 2010. (Cited on pages 21 and 29.)

P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5):610–5, Nov 2008. (Cited on pages 39, 43, 117, and 133.)

N. Rosenthal and S. Brown. The mouse ascending: perspectives for human-disease models. *Nat Cell Biol*, 9(9):993–9, Sep 2007. (Cited on pages 8 and 46.)

C. Rosse and J. L. V. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of biomedical informatics*, 36(6):478–500, Dec 2003. (Cited on pages 52 and 119.)

D. L. Rubin, N. H. Shah, and N. F. Noy. Biomedical ontologies: a functional perspective. *Brief Bioinformatics*, 9(1):75–90, Jan 2008. (Cited on page 9.)

Y. Saito, S. Hanaoka, M. Fukumizu, H. Morita, T. Ogawa, K. Takahashi, M. Ito, and T. Hashimoto. Polysomnographic studies of lesch-nyhan syndrome. *Brain Dev*, 20(8):579–85, Dec 1998. (Cited on page 85.)

E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, S. Federhen, M. Feolo, I. M. Fingerman, L. Y. Geer, W. Helmberg, Y. Kapustin, S. Krasnov, D. Landsman, D. J. Lipman, Z. Lu, T. L. Madden, T. Madej, D. R. Maglott, A. Marchler-Bauer, V. Miller, I. Karsch-Mizrachi, J. Ostell, A. Panchenko, L. Phan, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T. A. Tatusova, L. Wagner, Y. Wang, W. J. Wilbur, E. Yaschenko, and J. Ye. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 40(Database issue): D13–25, Jan 2012. (Cited on page 23.)

G. Schindelman, J. Fernandes, C. Bastiani, K. Yook, and P. Sternberg. Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*, 12(1):32, Jan 2011. (Cited on pages 130 and 131.)

A. Schlicker and M. Albrecht. Funsimmat update: new features for exploring functional similarity. *Nucleic Acids Res*, 38(Database issue): D244–8, Jan 2010. (Cited on page 17.)

A. Schlicker, T. Lengauer, and M. Albrecht. Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*, 26(18):i561–7, Sep 2010. (Cited on pages 11 and 17.)

P. N. Schofield, J. B. L. Bard, J. Boniver, V. Covelli, P. Delvenne, M. Ellender, W. Engstrom, W. Goessner, M. Gruenberger, H. Hoefler, J. W. Hopewell, M. Mancuso, C. Mothersill, L. Quintanilla-Martinez, B. Rozell, H. Sariola, J. P. Sundberg, and A. Ward. Pathbase: a new reference resource and database for laboratory mouse pathology. *Radiat Prot Dosimetry*, 112(4):525–8, Jan 2004. (Cited on page 53.)

P. N. Schofield, J. P. Sundberg, R. Hoehndorf, and G. V. Gkoutos. New approaches to the representation and analysis of phenotype knowledge in human diseases and their animal models. *Brief Funct Genomics*, 2011. (Cited on pages 5 and 71.)

P. N. Schofield, R. Hoehndorf, and G. V. Gkoutos. Mouse genetic and phenotypic resources for human genetics. *Human mutation*, 33(5): 826–36, May 2012. (Cited on pages 18, 45, and 117.)

S. Schulz, B. Suntisrivaraporn, F. Baader, and M. Boeker. SNOMED reaching its adolescence: ontologists' and logicians' health check. *International journal of medical informatics*, 78 Suppl 1:S86–94, Apr. 2009. (Cited on page 53.)

S. H. Settle, R. B. Rountree, A. Sinha, A. Thacker, K. Higgins, and D. M. Kingsley. Multiple joint and skeletal patterning defects caused by single and double mutations in the mouse gdf6 and gdf5 genes. *Dev Biol*, 254(1):116–30, Feb 2003. (Cited on page 64.)

N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10 Suppl 9:S14, Jan 2009. (Cited on pages 13, 28, and 29.)

H. Shatkay, A. Höglund, S. Brady, T. Blum, P. Dönnes, and O. Kohlbacher. Sherloc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics*, 23(11):1410–7, Jun 2007. (Cited on page 14.)

P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2011. (Cited on page 15.)

W. C. Skarnes, B. Rosen, A. P. West, M. Koutsourakis, W. Bushell, V. Iyer, A. O. Mujica, M. Thomas, J. Harrow, T. Cox, D. Jackson, J. Severin, P. Biggs, J. Fu, M. Nefedov, P. J. de Jong, A. F. Stewart, and A. Bradley. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature*, 474(7351):337–42, Jun 2011. (Cited on page 8.)

D. Smedley, M. A. Swertz, K. Wolstencroft, G. Proctor, M. Zouberakis, J. Bard, J. M. Hancock, and P. Schofield. Solutions for data integration in functional genomics: a critical assessment and case study. *Brief Bioinformatics*, 9(6):532–44, Nov 2008. (Cited on page 46.)

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, O. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–5, Nov 2007. (Cited on pages 11, 29, 39, 49, 52, 91, 119, and 129.)

C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 6(1):R7, Jan 2005. (Cited on pages 29, 39, 44, 72, 89, and 133.)

G. M. Spudich and X. M. Fernández-Suárez. Disease and phenotype data at ensembl. *Current protocols in human genetics / editorial board, Jonathan L Haines [et al]*, Chapter 6:Unit 6.11, Apr 2011. (Cited on page 21.)

P. Stenetorp, S. Pyysalo, G. Topić, and T. Ohta. BRAT. In *Proceedings of the . . .*, 2012. (Cited on page 115.)

D. J. Stumpo, C. B. Bock, J. S. Tuttle, and P. J. Blackshear. Marcks deficiency in mice leads to abnormal brain development and perinatal death. *Proc Natl Acad Sci USA*, 92(4):944–8, Feb 1995. (Cited on page 66.)

D. R. Swanson. Medical literature as a potential source of new knowledge. *Bull Med Libr Assoc*, 78(1):29–37, Jan 1990. (Cited on page 24.)

Y. Tateisi, A. Yakushiji, T. Ohta, and J. Tsujii. Syntax annotation for the genia corpus. *Proceedings of the IJCNLP Companion volume*, 2005. (Cited on page 25.)

E. W. Taylor, J. Xu, E. W. Jabs, and D. A. Meyers. Linkage analysis of genetic disorders. *Methods Mol Biol*, 68:11–25, Jan 1997. (Cited on page 16.)

P. Thompson, J. McNaught, S. Montemagni, N. Calzolari, R. del Gratta, V. Lee, S. Marchi, M. Monachini, P. Pezik, V. Quochi, C. J. Rupp, Y. Sasaki, G. Venturi, D. Rebholz-Schuhmann, and S. Ananiadou. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12:397, Jan 2011. (Cited on page 23.)

G. A. Thorisson, J. Muilu, and A. J. Brookes. Genotype-phenotype databases: challenges and solutions for the post-genomic era. *Nat Rev Genet*, 10(1):9–18, Jan 2009. (Cited on page 43.)

N. Tiffin, J. F. Kelso, A. R. Powell, H. Pan, V. B. Bajic, and W. A. Hide. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, 33(5):1544–52, Jan 2005. (Cited on page 17.)

S. H. Tirmizi, S. Aitken, D. A. Moreira, C. Mungall, J. Sequeda, N. H. Shah, and D. P. Miranker. Mapping between the obo and owl ontology languages. *Journal of Biomedical Semantics*, 2 Suppl 1:S3, Jan 2011. (Cited on page 11.)

L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. D. Moor, P. D. Causmaecker, and Y. Moreau. A guide to web tools to prioritize candidate genes. *Brief Bioinformatics*, 12(1):22–32, Jan 2010. (Cited on pages 17, 45, and 69.)

R. S. Travillian, T. Adamusiak, T. Burdett, M. Gruenberger, J. Hancock, A.-M. Mallon, J. Malone, P. Schofield, and H. Parkinson. Anatomy ontologies and potential users: bridging the gap. *Journal of Biomedical Semantics*, 2 Suppl 4:S3, Aug 2011. (Cited on page 113.)

M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–42, May 2006. (Cited on pages 3, 20, 21, 45, 71, and 90.)

R. A. Vos, J. P. Balhoff, J. A. Caravas, M. T. Holder, H. Lapp, W. P. Maddison, P. E. Midford, A. Priyam, J. Sukumaran, X. Xia, and

A. Stoltzfus. Nexml: rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, Feb 2012. (Cited on page 136.)

H. Wang, F. Azuaje, and O. Bodenreider. An ontology-driven clustering method for supporting gene expression analysis. *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems.*, 2005. (Cited on page 19.)

S. T. Warren and D. L. Nelson. Trinucleotide repeat expansions in neurological disease. *Current Opinion in Neurobiology*, 1993. (Cited on pages 5 and 71.)

N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol*, 7(11):e1000247, Nov 2009. (Cited on pages 2, 20, 21, 46, 66, 71, and 117.)

P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic Acids Res*, 39(Web Server issue):W541–5, Jul 2011. (Cited on pages 11 and 29.)

R. M. Winter and M. Baraitser. The london dysmorphology database. *J Med Genet*, 24(8):509–10, Aug 1987. (Cited on page 38.)

C. J. Wroe, R. Stevens, C. A. Goble, and M. Ashburner. A methodology to migrate the gene ontology to a description logic environment using DAML+OIL. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 624–635, 2003. (Cited on page 118.)

X. Wu, L. Zhu, J. Guo, D.-Y. Zhang, and K. Lin. Prediction of yeast protein-protein interaction network: insights from the gene ontology and annotations. *Nucleic Acids Res*, 34(7):2137–50, Jan 2006. (Cited on page 20.)

T. Xu, L. Du, and Y. Zhou. Evaluation of go-based functional similarity measures using s. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9:472, Jan 2008. (Cited on page 18.)

K. Yook, T. W. Harris, T. Bieri, A. Cabunoc, J. Chan, W. J. Chen, P. Davis, N. de la Cruz, A. Duong, R. Fang, U. Ganesan, C. Grove, K. Howe,

S. Kadam, R. Kishore, R. Lee, Y. Li, H.-M. Muller, C. Nakamura, B. Nash, P. Ozersky, M. Paulini, D. Raciti, A. Rangarajan, G. Schindelman, X. Shi, E. M. Schwarz, M. A. Tuli, K. V. Auken, D. Wang, X. Wang, G. Williams, J. Hodgkin, M. Berriman, R. Durbin, P. Kersey, J. Spieth, L. Stein, and P. W. Sternberg. Wormbase 2012: more genomes, more data, new website. *Nucleic Acids Res*, 40(Database issue):D735–41, Jan 2012. (Cited on page 46.)

S. Zhang and O. Bodenreider. Aligning representations of anatomy using lexical and structural methods. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, 2003. (Cited on page 16.)

S. Zhang, Z. Chang, Z. Li, H. Duanmu, Z. Li, K. Li, Y. Liu, F. Qiu, and Y. Xu. Calculating phenotypic similarity between genes using hierarchical structure data based on semantic similarity. *Gene*, Jan 2012. (Cited on pages 21 and 69.)

E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *Journal of theoretical biology*, 8(2):357–66, Mar 1965. (Cited on page 1.)

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ATS Alport syndrome, X-linked

AVATOL Assembling, Visualizing and Analyzing the Tree Of Life

AUC area under curve

BTHS Barth Syndrome

BWS Beckwith-Wiedemann syndrome

ChEBI Chemical Entities of Biological Interest

CoreSC Core Scientific Concept

DECIPHER Database of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources

EQ entity–quality

FMA Foundational Model of Anatomy

GO Gene Ontology

GOA Gene Ontology annotation

GWAS genome wide association studies

HPO Human Phenotype Ontology

HGNC Human Genome Organisation (HUGO) Gene Nomenclature Committee

IC information content

IMPC International Mouse Phenotyping Consortium

KEGG Kyoto Encyclopedia of Genes and Genomes

LDDB London Dysmorphology Database

LOOM Lexical OWL Ontology Matcher

MA Mouse adult gross Anatomy ontology

MeSH Medical Subject Heading

MGD Mouse Genome Informatics database

MOD model organism database

MP Mammalian Phenotype Ontology

MPATH Mouse PATHology ontology

NER named entity recognition

NLP natural language processing

OAEI Ontology Alignment Evaluation Initiative

OBA Open Biomedical Annotator

OBML Ontologies in Biomedicine and Life Sciences

OMIM Online Mendelian Inheritance in Man

OWL Web Ontology Language

PATO Phenotype And Trait Ontology

PheWAS phenome wide association studies

PKU phenylketonuria

PMC PubMed Central

PMID PubMed Identifier

PWS Prader-Willi syndrome

ROC receiver operating characteristic

SCD-EDS Spondylocheirodysplasia, Ehlers-Danlos syndrome-like

SNOMED CT SNOMED Clinical Terms

SOD Septo-Optic Dysplasia

TM text mining

UMLS Unified Medical Language System

UniProt Universal Protein resource

W3C World Wide Web consortium

WNP Weighted Network Predictor

COLOPHON

This thesis was typeset with LaTeX 2ε on the basis of André Miede's style classicthesis, which is available for LaTeX via CTAN as "classicthesis".

Final Version as of October 29, 2013 at 11:53.