Freie Universität Berlin

Department of Mathematics and Computer Science

Phenotype informatics:

Network approaches towards understanding the diseasome

Sebastian Köhler

Submitted on: 12th September 2012

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

1. Gutachter	Prof. Dr. Martin Vingron
2. Gutachter:	Prof. Dr. Peter N. Robinson
3. Gutachter:	Christopher J. Mungall, Ph.D.
Tag der Disputation:	16.05.2013

Preface

This thesis presents research work on novel computational approaches to investigate and characterise the association between genes and phenotypic abnormalities. It demonstrates methods for organisation, integration, and mining of phenotype data in the field of genetics, with special application to human genetics. Here I will describe the parts of this thesis that have been published in peer-reviewed journals. Often in modern science different people from different institutions contribute to research projects. The same is true for this thesis, and thus I will itemise who was responsible for specific sub-projects.

In chapter 2, a new method for associating genes to phenotypes by means of protein-protein-interaction networks is described. I present a strategy to organise disease data and show how this can be used to link diseases to the corresponding genes. I show that global network distance measure in interaction networks of proteins is well suited for investigating genotype-phenotype associations. This work has been published in 2008 in the *American Journal of Human Genetics*. My contribution here was to plan the project, implement the software, and finally test and evaluate the method on human genetics data; the implementation part was done in close collaboration with Sebastian Bauer. The people who manually compiled the list of “disease gene families” are Denise Horn and Peter Robinson.

The often-discussed problem of missing structured and computer interpretable representation of human phenotypic abnormalities is the focus of chapter 3. Replacing the free-text description of phenotypic abnormalities, especially for human hereditary diseases, has long been demanded by the research community. I was involved in developing the Human Phenotype Ontology (HPO), which has demonstrated the ability to overcome these problems (*American Journal of Human Genetics*, 2008). I contributed to this by, first implementing several computer programs for automatically constructing parts of this ontology. Secondly, I developed software for the exploration, maintenance, and quality assessment of the HPO. The output of my software was mostly manually curated by medical experts in the field of human genetics, especially Peter Robinson and Sandra Dölken.

After the HPO had been established, I showed that this ontology-based technique for representing and encoding phenotype data could help in the process of classifying hereditary diseases (*Human Mutation* 2012). Furthermore, I was involved in the development of novel strategies for using the HPO for differential diagnostics in clinical genetics, and implemented these in a web-based application (Phenomizer, *American Journal of Human Genetics*, 2009). This project was planned by Peter Robinson and myself. Furthermore, I implemented the web-based software, developed the statistical model to assign P-values to semantic similarity scores, and evaluated the performance of the search algorithms. Subsequently, a novel algorithm for exact determination of the P-values was developed mainly by Marcel Schulz and Sebastian Bauer. I was involved in this project, together with Peter Robinson. Here, the implementation and evaluation of the novel algorithm was my main contribution. This work was published in the *Proceedings of WABI* (2009) and in *BMC Bioinformatics* (2011). I coauthored a review on these topics in the journal *Medizinische Genetik* (2010).

Finally, I participate in a research collaboration with the University of Cambridge (Michael Ashburner, Paul Schofield, George Gkoutos), the Lawrence Berkeley National Laboratory, California (Suzanna Lewis, Chris Mungall), and the University of Oregon (Monte Westerfield, Barbara Ruef). We develop and investigate novel ways of semantically integrating different biomedical ontologies with application to genotype-phenotype association. This is the topic of chapter 4. The main focus is the use of a semantic representation of phenotype information from different species to build a large-scale, integrated ontological resource. This enables researchers to transfer phenotypic information from model organisms to human, and allows novel biomedical hypotheses to be generated. The approach of introducing logical definitions for human phenotypic abnormalities was presented in a paper for *IEEE Engineering in Medicine and Biology Society Proceedings* (2009), and describes strategies to create logical definitions. The idea of representing phenotypes in OWL and using reasoning was developed by several other people, such as Chris Mungall, Paul Schofield, Robert Hoehndorf, and Suzi Lewis. My projects build upon these works, such as a paper in *BMC Bioinformatics* (2011), which describes how the semantics can be used for quality control of ontologies by automatically detecting incomplete data representations. In this work, my contribution included the planning, implementation and evaluation. The semantic approach was also used in a project that investigated whole-phenome comparison between mouse and human in order to predict novel candidate genes for human diseases (*Human Mutation* (2012)). Finally, I contributed to a project (manuscript in preparation) where I created an integrated phenotype ontology across several species and implemented novel ways for analysing human chromosomal aberrations (CNVs). Here, my contribution was the creation of the Uberpheno ontology and the implementation of the analyses and visualisations.

Below is a summary of relevant publications for this thesis to which I contributed:

- Sebastian Köhler*, Sebastian Bauer*, Denise Horn, and Peter N Robinson.
  Walking the interactome for prioritization of candidate disease genes.
  The American Journal of Human Genetics, 82(4):949–58, Apr 2008.
  **equal contribution*

- Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos.
  The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease.
  The American Journal of Human Genetics, 83(5):610–5, Nov 2008.

- Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson.
  Clinical diagnostics in human genetics with semantic similarity searches in ontologies.
  The American Journal of Human Genetics, 85(4):457–64, Oct 2009.

- Marcel H Schulz, Sebastian Köhler, Sebastian Bauer, Martin Vingron, and Peter N Robinson.
  Exact score distribution computation for similarity searches in ontologies.
  WABI - Springer LNCS (Algorithms in Bioinformatics), 5724:298–309, 2009.

- George V Gkoutos, Chris J Mungall, Sandra C Dölken, Michael Ashburner, Suzanna E Lewis, John Hancock, Paul Schofield, Sebastian Köhler, and Peter N Robinson.
  Entity/Quality-Based Logical Definitions for the Human Skeletal Phenome using PATO.
  Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

- Sandra C Dölken, Sebastian Köhler, Sebastian Bauer, Claus E Ott, Peter Krawitz, Denise Horn, Stefan Mundlos, and Peter N Robinson
  Computational methods for the study of human disease manifestations. The Human Phenotype Ontology.
  Medizinische Genetik, 22(2):221–8, June 2010

- Sebastian Köhler, Sebastian Bauer, Chris J Mungall, Gabriele Carletti, Cynthia L Smith, Paul Schofield, George V Gkoutos, and Peter N Robinson
  Improving ontologies by automatically reasoning and evaluation of logical definitions.
  BMC Bioinformatics, 12:418, October 2011

- Marcel H Schulz, Sebastian Köhler, Sebastian Bauer, and Peter N Robinson
  Exact Score Distribution Computation for Ontological Similarity Searches.
  BMC Bioinformatics, 12:441, November 2011

- Chao-Kung Chen, Chris J Mungall, George V Gkoutos, Sandra C Dölken, Sebastian Köhler, Barbara J Ruef, Cynthia Smith, Monte Westerfield, Peter N Robinson, Suzanna E Lewis, Paul N Schofield, Damian Smedley
  MouseFinder: candidate disease genes from mouse phenotype data.
  Human Mutation, 33(5):858–66, April 2012

- Sebastian Köhler, Sandra C Dölken, Ana Rath, Ségolène Aymé, and Peter N Robinson
  Ontological Phenotype Standards for Neurogenetics.
  Human Mutation, 33(5):1333–9, August 2012

- Sebastian Bauer, Sebastian Köhler, Marcel H Schulz, and Peter N Robinson
  Bayesian Ontology Querying for Accurate and Noise-Tolerant Semantic Searches.
  Bioinformatics, 2012

- Sandra C Dölken*, Sebastian Köhler*, Chris J Mungall*, George V Gkoutos, Barbara J Ruef, Cynthia Smith, Damian Smedley, Sebastian Bauer, Eva Klopocki, Paul N Schofield, Monte Westerfield, Peter N Robinson, Suzanna E Lewis
  Phenome-wide interspecies semantic mapping reveals phenotypic overlap in the contribution of individual genes to CNV pathogenicity
  Under review, 2012
  **equal contribution*

I also contributed to other projects, that are not part of this thesis. One of these projects dealt with the investigation of ultra-conserved regions in the human genome with a special focus on promoter regions [Rödelsperger et al., 2009]. In another project a new method for long-range prediction of target genes for enhancer elements was developed [Rödelsperger et al., 2011]. Finally I contributed to a work on next-generation sequencing for disease gene discovery [Krawitz et al., 2010].

Thesis Contribution

Finding the molecular basis for specific phenotypes is an important topic in genetics, cell biology, molecular biology, and developmental biology, among others.

This thesis adds value to this topic by presenting a novel method for correlating information on phenotypic variation with high-throughput biological data. It also describes how phenotypic information can be represented in systematic computer interpretable structures, i.e. ontologies. This work shows how such representations pave the way for novel instruments for clinical genetics.

Utilising precise and comparable phenotypic information across different species is of major interest to the scientific community, since this is critical for gaining a detailed understanding of the connections between diseases and genes. Here, I present ways for semantic integration of phenotype data across species and the successful linkage of genotype information to phenotype data, which is fundamental for novel hypothesis generation in biomedical sciences, especially human genetics.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Phenotypes

The word phenotype is used to specify an organism's measurable or observable traits, such as human body height, blood pH values, or a mouse's reaction to a loud handclap. Phenotypes are used as criteria to differentiate between individuals or species. Phenotypic abnormalities are important criteria for disease diagnosis and treatment.

Starting with the ancient Greeks, scientists have been interested in the systematic investigation, characterisation and understanding of phenotypes, especially in abnormal phenotypes associated with disease. Finding a detailed explanation on the origin of phenotypes, and assessing the influence of environmental conditions, is even a hard task today.

During the scientific revolution in the 16th and 17th century, mankind was enriched with numerous major achievements. For example, microscopy enabled the discovery that organisms are composed of cells – the functional basic unit of life. Since the 19th century, one of the central questions for biologists has been finding cellular and molecular correlates of observable traits or phenotypes. In the beginning, scientists focused on exploring the patterns of how peculiar characteristics and qualities are "inherited" from parents or more remote ancestors. The seminal work of Gregor Mendel described the law of inheritance [Mendel, 1866], which he derived by hybridising garden peas and systematically investigating specific traits across different generations. The first work that separated the concept of hereditary material of cells from the actual appearance was done by Johannsen [1909]. He made clear that there exists an organism's heredity on one hand and the product of this heredity on the other hand.

In medical context the word phenotype is often used in reference to some deviation from normal morphology, physiology, or behavior [Robinson, 2012]. In this present thesis, the definition of Strickberger [1985] applies, in which he states that phenotypes are *the manifold biological appearances, including chemical, structural and behavioral attributes, that we can observe about an organism, but excludes its genetic constitution*. In biology, this or similar definitions are commonly used [Mahner and Kary, 1997].

1.1.1 DNA, Genes and Genotype

All eukaryotes are living organisms that are composed of one or multiple cells. Eukaryotic cells have a nucleus that contains the carrier of the hereditary information in the form of deoxyribonucleic acid (DNA) molecules. DNA consists of two long polymers and is organised in the form of a double-helix [Watson and Crick, 1953]. It can be thought of as a biological blueprint, since it contains the information needed to construct other components of cells, such as proteins. The complete set of DNA for an organism is also called the genome. A union of genomic DNA-sequences that encode a coherent set of potentially overlapping functional products is called a gene [Gerstein et al., 2007]. The complete constitution, or makeup, of the genetic material belonging to a cell, or an individual, is called the genotype. Aside from genes, the DNA contains several regulatory regions, for example, promoter- and enhancer regions, which control the efficiency of gene transcription.

The completion of the sequencing of the human genome in 2001 [Venter et al., 2001, Lander et al., 2001] was only one step in a long process to discover and examine all the variations in the DNA (e.g. in genes), and to understand how those variations influence diseases. Subsequent research projects have revealed new and largely unexpected levels of genetic complexity in the three billion DNA bases of the human genome. It is estimated that at least 15 million places along the human genome can differ from one person, or population, to the next [Pennisi, 2007]. DNA alterations that represent only a variation of a single nucleotide are called single-nucleotide polymorphisms (SNP). SNPs are assumed to play a major role in causing phenotypic differences between individuals.

1.1.2 Copy Number Variations

Besides SNPs, larger structural variations, called *copy number variations* (CNV), have been identified as an important source of human DNA polymorphism [O'Donovan et al., 2008]. In CNVs, longer stretches of DNA can get lost, duplicated, or rearranged in the genome of an individual. These CNVs may comprise only a few base pairs, but they may also span regions containing millions of bases. The detection of CNVs, and the finding that these changes alter a genome in just a few generations, revolutionised the perception of the human genome as a stable entity.

Microdeletions and microduplications are defined as CNVs that are too small to be seen through a microscope and typically span less than five million base pairs. They may contain one or more contiguous genes. The loss or duplication of these genes causes an alteration in the gene dosage, and may cause lower or higher levels of the gene product. The altered gene dosages often cause various phenotypic abnormalities.

Understanding the enormous complexity of genomic variations and, in particular, explaining the phenotypic consequences and contributions of genomic variations are important challenges for current biomedical sciences.

1.1.3 Protein, Proteome and Function

The central dogma of molecular biology [Crick, 1958] assumes that, in a process called RNA synthesis, the coded information of genes is transferred to RNA. In the following step, the RNA is processed and translated according to the genetic code into an amino-acid chain in which the amino-acids are connected by peptide bonds.

The amino-acid chain subsequently folds into a functional shape: a three-dimensional structure, called protein. Proteins are the key players in biological activities such as enzymatic reactions or structural organisation. Virtually every cellular process, from DNA-replication to signal transduction, depends on the coordinated interplay of different proteins. The entirety of proteins that are expressed by a genome is called the proteome.

1.1.4 Hereditary Diseases

Diseases or illnesses are defined and identified by their combination of symptoms, i.e. phenotypic abnormalities. A phenotypic abnormality is a phenotypic feature, whose value deviates from a commonly accepted "average" nominal value.

In 1941, Beadle and Tatum [1941] showed, for the first time, that an alteration in a gene's DNA-sequence can result in an altered phenotype and, in doing, came up with the so-called one-gene-one-enzyme hypothesis.

Hereditary diseases are diseases in which the phenotypic abnormalities are caused, in a large fraction, by variations in the genome, i.e., genes or chromosomes. They are also called genetic diseases or Mendelian diseases. There are several major classes of genetic diseases, with the most important being monogenic diseases, polygenic/multifactorial diseases, chromosomal diseases, and mitochondrial diseases. It is important to note that, in a lot of genetic diseases, environmental factors influence the spectrum of phenotypic abnormalities.

Of course, not every alteration or mutation in a gene causes altered phenotypes, but those variations are well-distributed across the genome. Variations that are known to predispose to or cause a specific disease are called *disease-causing mutations*. An example of how a gene mutation translates to a phenotype is given in Figure 1.1.

The resolution of measurements in biological sciences has improved enormously during the last centuries for both genotype and phenotype. New sequencing techniques enable scientists to sequence and characterise a complete genome within a few days. New measurement methods in

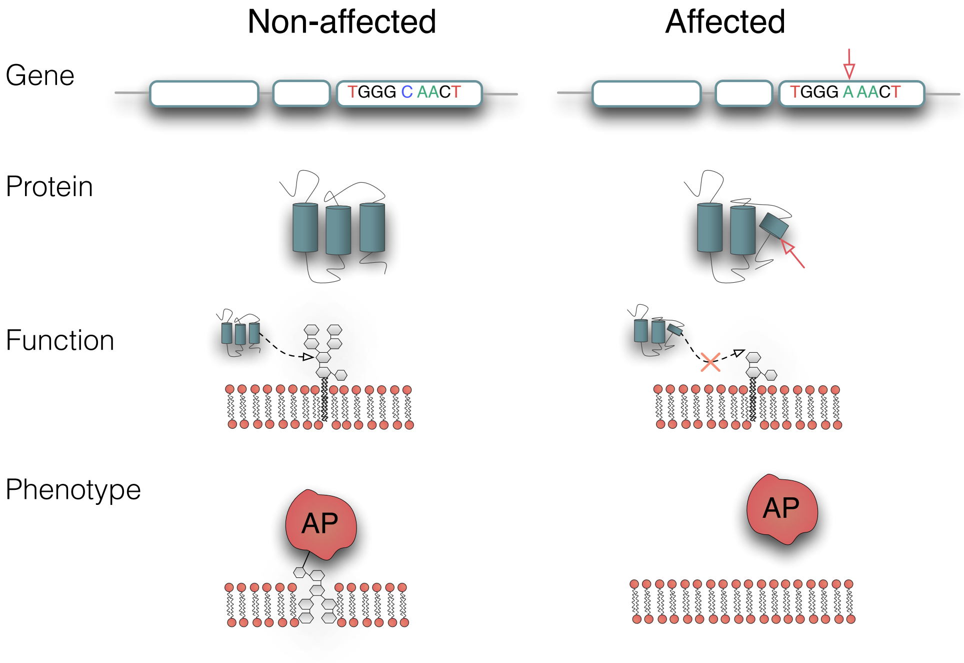


Figure 1.1: This figure illustrates an example of the mechanism by which a genomic variation in the hereditary *Hyperphosphatasia-mental retardation syndrome (HPMR)* transmits to an altered phenotype. Most affected individuals have a mutation in the gene *PIGV* [Krawitz et al., 2010, Horn et al., 2011]. In individuals without this mutation, this gene encodes a protein that is part of a cellular machinery for synthesising the GPI-anchor. This GPI-anchor is used to attach, for example, the alkaline phosphatase (AP) protein to the cell membrane. In affected individuals, the synthesis of the anchor fails, due to the malfunction of *PIGV*. Subsequently, the AP cannot be anchored to the cell and is thus circulating in the serum. The elevated serum levels of AP is one of the major phenotypic abnormalities in the HPMR syndrome.

medicine allow for fast and accurate determination of specific metabolites, e.g. in blood or urine. Despite these advancements, various problems still arise. Even to distinguish between SNPs and disease-causing mutations is a hard problem. Finding the accurate mechanism for how an individual's disease-causing mutation propagates to varied phenotypes still remains an even more challenging problem.

1.2 Networks

Complex networks of objects describe a wide range of real-world systems, such as physical, biological, and social systems. There exist networks of natural phenomena, such as food webs (who eats who) or sociological networks where individuals are connected by social relationships. Also, the Internet can be seen as a network of routers and computers connected by physical links [Albert and Barabási, 2002]. In the last decades, a lot of these systems were better understood through investigating the mechanisms underlying the topology of those complex networks. The focus here are networks of objects in molecular biology that represent the cell's functional organisation.

1.2.1 Protein-Protein Interaction Networks

Proteins are almost always performing their functions in cooperation with other proteins and rarely act alone. Often, they team up into "molecular machines", forming physicochemical dynamic connections to undertake biological functions at both cellular and systems levels [Rivas and Fontanillo, 2010]. For example, the enzyme *liver alcohol dehydrogenase* is a protein complex composed of several different subunits that, in combination, perform the task of breaking down alcohols.

In recent years, it became clear that almost all proteins are part of a huge, connected system that makes up a cell and defines its properties and behavior. The set of protein interactions in a cell is called the interactome. There are now several different methods to detect the possible physicochemical interactions of a particular protein. The most important ones are the yeast two-hybrid system and the tandem affinity purification followed by mass spectroscopy [Shoemaker and Panchenko, 2007]. In recent years, high-throughput screening for protein-protein interaction was successfully performed in cells from many different organisms, even for higher eukaryotes such as human [Stelzl et al., 2005]. This lead to the creation of publicly accessible databases that contain the interactomes for several species. Given that it is now becoming more and more evident that the interactions between proteins determine the outcome of almost every cellular process, analyses based on protein interaction networks will pave the way for a systems-level understanding of cellular processes. From the medical genetics point of view, those datasets gain major importance under the assumption that the mutations in a gene's DNA sequence propagate through the translation-step, cause a distortion of protein interfaces, and finally disturb the cellular network which, in turn, may lead to the development of many phenotypic abnormalities [Shoemaker and Panchenko, 2007].

1.2.2 Graphs

In computer science, a typical approach to represent and encode relationships of interest between objects of the same domain are graphs. In this work, graphs are used in the context of protein interactions and ontologies. In this thesis, a graph $G = (V, E)$ is defined, in accordance to Gross and Yellen [2006], as a mathematical structure that consists of two finite sets: V and E. The elements of V are referred to as vertices or nodes and represent the objects of interest. The elements of E are called edges or arcs and express relationships between the objects. Each edge has a pair of two (possibly non-unique) vertices associated with it, which are called the endpoints of the edge. If the pair of nodes is unordered, the edge is *undirected*; if it is ordered, the edge is *directed*. If G contains only undirected edges, it is referred to as *undirected graph*; on the contrary, if G only contains directed edges, it is referred to as *directed graph* (see Figure 1.2). For example, the nodes of a protein-protein interaction graph represent proteins, or the genes that encode for the protein. An edge exists in the graph if there is

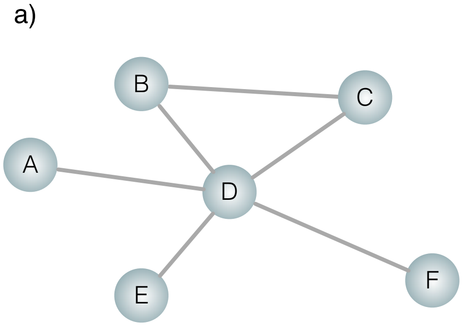


Figure 1.2: This figure shows examples of undirected and directed graphs. In *a)* an undirected graph with six nodes and six edges is shown. Part *b)* depicts a directed version of the graph, where the edges are defined as ordered pairs of nodes.

evidence for a biophysical interaction. Protein-protein interactions graphs are undirected in this thesis, whereas graphs representing ontologies are directed graphs.

The representation of biological interaction networks as graphs gives biologists and bioinformaticians several considerable advantages. First of all, graphs can be visually presented as drawings, which is essential for both data analysis and interpretation. Second of all, several other research fields have worked on the study of network representations. This lead to the development of several novel algorithms for analysis, and new theories about the structure and organisation of networks were hypothesized. For example, in social science, the theory of networks could explain social phenomena in a wide variety of disciplines, from psychology to economics [Borgatti et al., 2009].

1.2.3 Graph Theory

Graphs and their components (i.e., nodes and edges) can be mathematically characterised by several measures. For this thesis, only a small subset of these attributes are required.

The *degree* of a node v is defined as the number of edges that involve v as endpoint. As a means to characterise a complete graph, the average degree of its nodes and the degree distribution are often used.

The *clustering coefficient* of a node v ($\mathrm{CC}(v)$) measures the extent of clustering in a graph around that node, whereby a cluster is said to be a subset of nodes where there is a high chance that the edges of these nodes connect only nodes from this subset. The clustering coefficient of a node v is defined as

$$\mathrm{CC}(v) = \frac{2 * e}{k * (k-1)}\,, \tag{1.1}$$

where e is the number of edges that connect two neighbors of v, and k is the number of neighbors of v. A node has a high clustering coefficient if a lot of its neighbors are adjacent to each other as well.

A path in a graph is defined as a sequence of nodes of the graph in which the successive nodes have an edge between them. A very important measure in graph theory is the *shortest path* (SP) between two nodes. Intuitively, it can be seen as an estimate for how long information transmission between those two nodes takes.

1.3 Knowledge Representation

In the biomedical setting, computer science has focussed largely on efficient analysis of huge amounts of molecular data, such as DNA sequences or protein structures. In recent years, biomedicine has been confronted with additional challenges caused by an unprecedented increase in the size of the data sets due to modern high-throughput technologies. For example, having sequenced the human genome, now even on a personalised level, more and more research projects aim at systematically unravelling the genomic complexity behind specific phenotypic abnormalities. Thus, the problem of efficient and reliable retrieval and analysis of data has become the most important scientific bottleneck [Soldatova and King, 2005]. This is because different databases use inconsistent naming conventions, different intentions of users in their writing, varying definitions, and so forth. The differences may be syntactical and semantical [Schulze-Kremer, 1997].

These problems induced the need for sophisticated approaches for storing the generated knowledge in knowledge bases (KB) so that it is preserved for the future, and processable by machines in an automated way. This brings up new problems regarding the integration, management and interpretation of the generated information. The goal is that scientists should, at best, be provided with methods for consistent annotation of data and simultaneously be provided with tools for greater interoperability among people and machines [Mabee et al., 2007]. Of course, the data should be made available to computer-based search and to algorithmic processing.

The study of knowledge representation (KR) to enable automated processing by computers has been a research field since the 1970's with the emergence of complex, artificial intelligence systems. Representing knowledge by means of symbols, and in a way that enables accurate and effective reasoning is one of the major hallmarks of KR research. It aims to design and apply systems for storing facts and rules about subjects.

In contrast to KR based approaches, describing knowledge in the form of free text allows for maximal expressivity, but impedes reliable mining of information by means of automated methods. This becomes obvious when trying to extract information from text-based resources. Firstly, a search in this data inventory leads to too many search results, e.g., querying the internet for *ventricle* will return results related to both the brain ventricles and the heart ventricles. Here, the underlying problem is the missing context that would distinguish homonymic words. Another typical problem is the inability to capture synonymic terms, e.g., querying a free-text resource for *acrocephaly* will not yield results that only match the synonym *oxycephaly*. Another problem is that inconsistencies cannot be discovered systematically, so that one can write that, e.g., the liver has the function to store glycogen or the negation of that statement without consequences. Thus text-based systems are not able to recognise inconsistencies in the data.

1.3.1 Semantic Web and Ontologies

The problems and ideas mentioned above induced the step to move biomedical KR in the direction of Semantic Web techniques, which is a proposal for the next-generation of the World-Wide Web (WWW). Originally, the WWW was created as a web of hypertext documents using hyperlinks to allow users to browse between documents located on different web servers. Semantics is the study of deriving the sense and meaning of complex concepts from simpler concepts, and draws upon syntax. Semantics is dependent on context and pragmatics. In general, free-text descriptions, be they phenotype descriptions or the content of the world wide web, contain only implicit semantics and lack an explicit semantic representation. The motivation to put the Semantic Web techniques forward was that, although Web browsers can easily parse Web pages in order to create a layout and link out to other Web pages, there is no reliable way for a computer to recognise the meaning of the content of the Web pages they display. The inventor of the Semantic Web, Tim Berners-Lee, defined it as "a web of data that can be processed directly and indirectly by machines." Thus, the Semantic Web should function as a framework in which computers are capable of analyzing the meaning – the semantic content – of the data on the Web in order to act as "intelligent agents."

The ideas used in this thesis were largely influenced and developed by the Semantic Web community. One of the initial steps towards the semantic web was to represent knowledge as a list of statements, whereby each statement is a triple that consists of a *subject*, a *predicate*, and an *object*. For representing and addressing the parts of these triples, unique Uniform Resource Identifiers (URI)[1] are used. Standardised coding and structuring of the statements is done using the Resource Description Framework (RDF), which is has an XML serialisation. This paves the way towards interoperability, but contains only a minimal amount of semantics. Basically, RDF is used for representing assertional knowledge (ABox), i.e. facts and knowledge about instances, which corresponds to the description of actual data.

In order to introduce semantics, standardised knowledge representa-

[1] The *object* may also be a literal, which is not a URI.

tion (i.e. ontologies) is introduced, which can explicitly formalise the semantics. The word ontology was originally used to describe a specific branch of metaphysics that examined the nature and relations of being. This branch tries to answer which entities do exist, in which way there exist groupings among them, and if there exist hierarchical relationships. An ontology can be seen as a semantic model for a domain of reality. Probably, the best known modern day definition of an ontology in computer science from Gruber [1995] states:

> An ontology is an *explicit*, *formal* specification of a *shared conceptualization*.

This definition contains several important aspects of an ontology. First, *explicit* denotes that the meaning of every concept is well defined. The word *formal* refers to the fact that the information in an ontology must be machine-readable and interpretable, and *shared* implies that there is some kind of consensus regarding the information. Finally, *conceptualisation* refers to an abstract model of the phenomenon of interest that is represented by the ontology and that the ontology contains the relevant concepts, which are named and defined entities of the domain of interest. Note that, in this thesis, the words *concept* and *term* are interchangeable. Later on, the word *class* is also used in the context of OWL. The main component of ontologies is a set of *axioms*. Axioms are herein used to refer to statements that say what is true in the domain.[2] There are two kinds of axioms: asserted axioms and inferred axioms.

Ontologies formally describe concepts, both by their meaning and by their relationship to each other [Bard and Rhee, 2004]. The most important semantic *roles* or *relationships* in this thesis are `part_of` and `is_a` (also `subClassOf`). For this thesis it is important that ontologies enable that the knowledge is machine interpretable, integratable across heterogenous datasources, and applicable to automated reasoners to infer implicit knowledge. Figure 1.3 shows an excerpt of what may be the most important biomedical ontology, the *Gene Ontology* (GO). In this example, the protein *PIGV* is annotated to the GO term *GPI anchor biosynthetic process*, which

[2] `http://www.w3.org/TR/owl2-syntax/#Axioms`

means that there is evidence that this protein plays a role in the named process. In the ontology, there are several more general concepts that are semantically linked to the GO term.

The extension of RDF with RDF schema (RDFS) enables the formulation of statements on classes, their hierarchical organisation, and inheritance. RDFS is used for TBox axioms, which describe terminological knowledge, i.e. the knowledge about classes. Its axioms describe the structure of the domain that is to be modelled in the form of ontologies. RDF(S)[3] allows only for simple reasoning tasks by utilising the transitivity property of the `subClassOf` and `subPropertyOf` relations. Also, the domain and range restrictions of relations can be utilised. However, this kind of KR does not allow for testing the truth-value of statements. Additionally, it is not possible to express negations or define exclusion criteria. For this, the combination with logics and formal knowledge modelling is required. A logical foundation is used to supply a well-defined, formal semantics and to enable automated systems to operate on the KB. This is required for the important task of testing the logical entailment between statements and drawing valid conclusions.

Typical candidates of logics include propositional logic (PL) or first order logic (FOL). Although PL would be perfect in terms of decidability, etc., it is not well-suited for KR because several important statements, such as assertions about groups, cannot be expressed. FOL on the other hand, has the advantage of being highly expressive, but the disadvantage of being semidecidable. This means that reasoning algorithms are not guaranteed to terminate.[4] This resulted in the idea of using restricted subsets of FOL for modeling. One popular subset is the Description Logic (DL) family of languages, which has a reduced expressivity but the advantage that most DLs are decidable. Most notably, the existential and universal quantifications are limited by role restrictions in most DLs. The language of choice for authoring DL KBs is the W3C[5] standardised Web Ontology

[3] Denotes the combination of RDF and RDFS.

[4] Answering if something is contained will terminate if the containedness is given, but may not terminate if it is not contained.

[5] World Wide Web Consortium

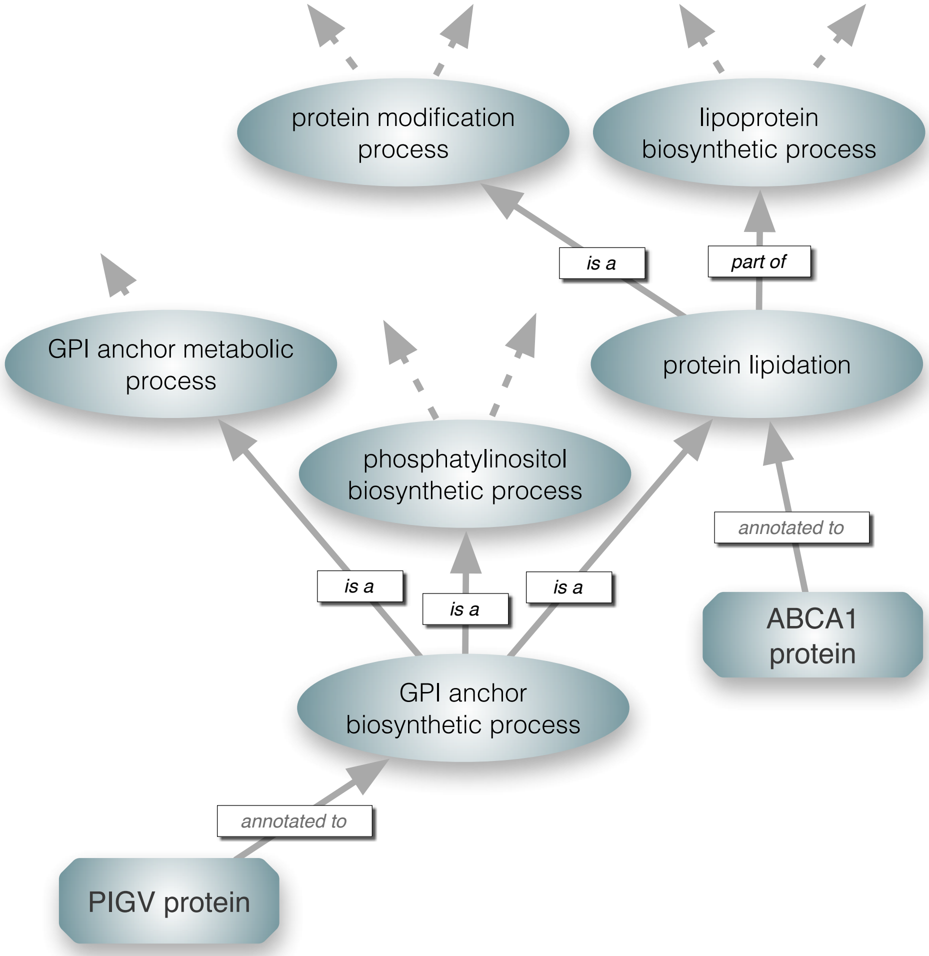

protein modification process
lipoprotein biosynthetic process
is a
part of
GPI anchor metabolic process
protein lipidation
phosphatylinositol biosynthetic process
annotated to
is a
is a
is a
ABCA1 protein
GPI anchor biosynthetic process
annotated to
PIGV protein

Figure 1.3: This figure shows an excerpt of the graphical representation of the *Gene Ontology* (GO). Here, the *PIGV* protein is annotated to the GO-concept or term *GPI anchor biosynthetic process*. Two kinds of relationships are shown: the `is_a` and `part_of` relationships.

Language (OWL DL[6]), for which a serialisation based on RDF/XML exists. Using OWL DL, complex classes can be described from simpler descriptions using, for example, intersections, unions, and complements. Different DLs are characterised by the sets of allowed constructors. For example, in this work, the OWL 2 EL profile is used to enable efficient reasoning using the ELK reasoner [Kazakov et al., 2011].

For the interpretation of expressions the commonly accepted approach by Tarski is based on set-theoretical constructions and formalises logic with model-theoretic semantics [Tarski and Corcoran, 1983]. When assessing the truth of statements, this approach allows proofs that are not dependent on testing all possible interpretations, but rather apply purely syntactic rules in order to prove statements in a formally correct way. This is needed to test for entailment, which is a relationship between two expressions: A and B[7]. A model of an expression A is an interpretation for which the expression is true. A entails B if, under any interpretation for which A is true, B is also true. This is equivalent to stating that all the models of A are also models of B.

The basis for efficient, automated proving for entailments is the *reduction to unsatisfiability*. For a specific statement or formula, one assumes that it is a valid formula, i.e. it is true under every possible interpretation. To show this, one simply reduces the question of validity to a question of satisfiability by assuming that the negation of the statement is unsatisfiable. If it is then possible to show a contradiction in the negated formula (e.g. it contains $a \wedge \neg a$), this proves that the negated formula is indeed unsatisfiable, which in turn shows that the original formula is valid. Algorithms such as the Tableau (with blocking) allow for efficient reasoning, even for larger KBs. In the context of this thesis, it especially important that proofs of satisfiability and subsumption are decidable. One of the simplest examples of inference is illustrated in Figure 1.4, where four terms of the GO are linked by `is_a` relationships. Furthermore, one *inferred* relationship is shown. Here, the basis of the inference is the transitivity of the `is_a` relationships, i.e., if class C `is_a` D, and D `is_a` E, then it is also true that C

[6] The successor of OWL DL is OWL 2 DL.

[7] Note that A and B must be truth-valued expressions.

`is_a` *E*. Thus, one can infer that every *GPI anchor biosynthetic process* `is_a` *macromolecule modification*. These kinds of inferences can be used to detect hidden or new knowledge that has not yet been asserted in the ontology.

For this thesis, a more exhaustive introduction to *description logic* and reasoning, such as decidability and complexity, is out of scope. Also, introduction of the different DL flavours, such as $\mathcal{ALC}$, $\mathcal{SHOIN}(\mathcal{D})$, and $\mathcal{SHROIQ}(\mathcal{D})$, is skipped here and can be found in the books by Baader et al. [2003] and Hitzler et al. [2009].

Several syntax representations exist for OWL ontologies, but in this thesis, only the Manchester syntax [Horridge et al., 2006] and OBO syntax[8] are shown. In the actual implementation and serialisation, RDF/XML is used. The OBO syntax is used for OBO Foundry ontologies and has several favourable characteristics, such as human readability and minimal redundancy. For dealing with OWL ontologies programmatically, the OWL-API [Horridge and Bechhofer, 2009] and OWL-Tools [owl] are used. Building and editing of ontologies can be done with OboEdit [Day-Richter et al., 2007] or Protégé [pro].

Tools like the oboformat library[9] for the interconversion between the OBO format and OWL 2 have been developed and are used in this work. In OWL, an ontology is a collection of axioms. Having created OWL ontologies, one can apply automatic reasoners, which are ready-to-use systems for computing the logical consequences that can be inferred from a set of asserted axioms. As for the OBO to OWL conversion, tools for automated reasoning in ontologies represented in OWL DL exist. Commonly used software includes FaCT++ [Tsarkov and Horrocks, 2006], HermiT [Motik et al., 2007], Pellet [Sirin et al., 2007], or ELK [Kazakov et al., 2011]. All of the mentioned reasoners can be invoked via OWL-API.

1.3.2 OBO Foundry

The GO (see Figure 1.3) has probably been the most successful ontology in the domain of biology. This success has lead to a "golden age" of ontolo-

[8] `http://www.geneontology.org/GO.format.obo-1_2.shtml`

[9] Available at `http://code.google.com/p/oboformat/`.

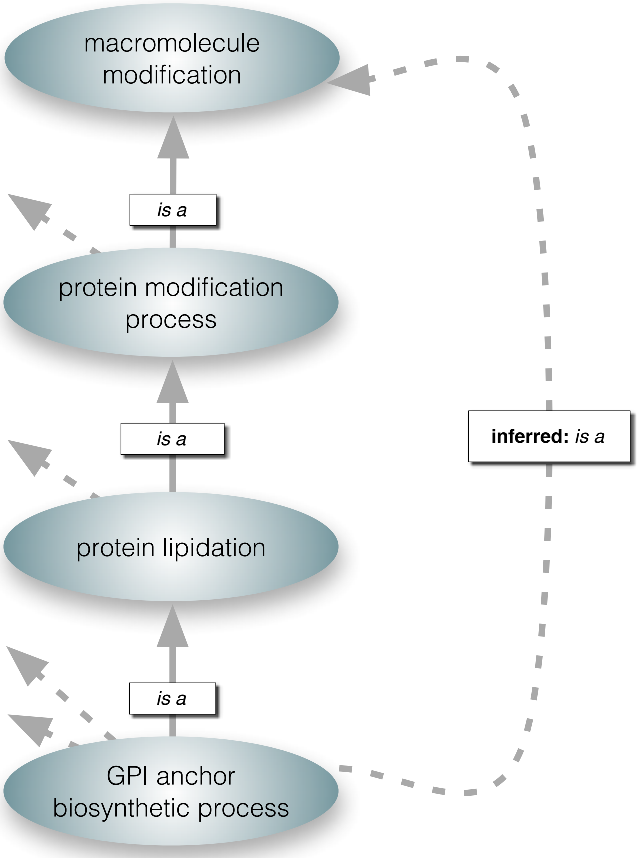

macromolecule modification
is a
protein modification process
is a
inferred: is a
protein lipidation
is a
GPI anchor biosynthetic process

Figure 1.4: Example of a simple inference in ontologies. Using the transitivity property of the *subclass* relationship, a reasoner infers that, if a term T is asserted to be a subclass of a parent term P, then the term T is also a subclass of the parents of P. Intuitively, this is the same as stating that the ancestors of my father (which is my ancestor) are also my ancestors, since the ancestor relationship is also transitive.

gies in biomedical research. The fact that most of these domain-ontologies were developed in isolation, created novel obstacles to interoperability across these domains [Smith et al., 2007].

Thus, the OBO Foundry (The Open Biological and Biomedical Ontologies Foundry) has the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain family of ontologies [Smith et al., 2007]. There exist several different types of relationships that are used in ontologies, such as `is_a`, `part_of`, `contained_in`, `instance_of`, and `has_participant`. In this work, the `is_a` relationship is of major importance. This relationship is equivalent to the subsumption or subclass relationship and means that, if a term t_1 `is a` term t_2 then t_1 represents a subclass of the more general parent term t_2. The `part of` relationship expresses a part-whole relationship.

The *annotation propagation rule* states that the `annotated_to` relation is propagated along the `is_a` and `part_of` relations to parent terms, and thus to all ancestral terms. This implies that the annotation of an object O to the term t results in the annotation of O to all the ancestor terms of t. For example, if a gene product is annotated to a specific GO term, it is implicitly annotated to all of its ancestors. This means that the *PIGV* protein in Figure 1.3 is annotated not only to *GPI anchor biosynthetic process* but also to all the shown ancestors such as *protein modification process*. More details and theoretical foundations of this rule are given in the book of Robinson and Bauer [2011, chap.6] and are not included in this thesis due to the space constraints.

1.3.3 Semantic Similarity

Ontologies can be separated by their purpose, e.g. there exist ontologies that are used to provide a controlled vocabulary for the objects of a domain (*domain ontologies*). In this thesis, a more specific type of domain ontologies is introduced, which are ontologies that are used to describe the attributes of the domain objects (*attribute ontologies*). Typical *domain ontologies* are the Foundational Model of Anatomy (FMA) [Rosse and Mejino, 2003], which provides a controlled vocabulary for anatomical entities or the Chemical Entities of Biological Interest (ChEBI) ontology [de Matos et al., 2010] which describes objects from the domain of biologically relevant chemicals. The GO is a typical *attribute ontology*, i.e., its terms are used to create *annotation relationships* between terms and genes or proteins. Thus, annotation is used to assign biological functions, characteristics, or attributes to the genes.

Domain ontologies such as the FMA allow subsumption searches. This means that, for example, a search for objects referring to *skeleton of hand* will also return entries on *set of carpal bones*, because it is a subclass of *skeleton of hand*. On the other hand attribute ontologies have the advantage that they additionally allow searches for objects that are annotated to terms of the ontology based on *semantic similarity*.

Information Content

Resnik [1995] presented a novel approach to calculate the semantic similarity between two concepts in an ontology with `is_a` relationships. The first step of the method is to associate an *information content* (IC) value with each of the terms of the ontology. The definition of the IC follows the standard argumentation of information theory, and the IC of a term t is calculated as

$$\mathrm{IC}(t) = -\ln(p(t)) \, , \tag{1.2}$$

where $p(t)$ is the probability of encountering t in a corpus, such as all the gene products in a database. This is normally estimated by the fraction of objects being annotated to t among all objects in the database. The *annotation propagation rule* implies that p is a monotonically increasing function when following the links of an ontology from bottom to top, whereby the root r of the ontology has no information content ($p(r) = 0$), since all objects are annotated to the root. In general, if t_1 `is_a` t_2, then $p(t_1) \leq p(t_2)$.

From this follows that the IC of the terms decreases as one moves up the hierarchy along the links of the ontology. Taking the example of Figure 1.3 and assuming that there are 10.000 genes in the database the term *GPI anchor biosynthetic process* would have an IC of $-\ln(1/10{,}000) = 9.2$ because only *PIGV* is annotated to this term. The IC of *protein lipidation* would then be $-\ln(2/10,000) = 8.5$ because two genes (*PIGV* and *ABCA1*) are annotated to this term. In summary, very specific terms are only associated with very few genes and, therefore, have a high IC, whereas nonspecific terms associated with many genes have a low IC.

Similarity Between Terms

In recent years, several methods have been published that can be used to determine the similarity of two terms t_1 and t_2 in an ontology. Again a comprehensive introduction to this topic would be out of scope for this thesis and can, instead, be found in the work of Pesquita et al. [2009]. Here, only the information-content-based measure developed by Resnik [1995] is of particular interest. Let $\mathrm{anc}(t)$ be the function that returns the set of all the nodes that are found on all paths from term t to the root, including t. When determining the similarity between t_1 and t_2 the intersection between $\mathrm{anc}(t_1)$ and $\mathrm{anc}(t_2)$ is created and the highest IC of these terms is taken to be the similarity value. At first, let $\mathrm{CA}(t_1, t_2) = \mathrm{anc}(t_1) \cap \mathrm{anc}(t_2)$ set of common ancestors of two nodes. Then, the similarity between two terms is defined as

$$\mathrm{sim}(t_1, t_2) = \max\{\mathrm{IC}(a) | a \in \mathrm{CA}(t_1, t_2)\} \ . \tag{1.3}$$

This means that the information content of the most-specific common ancestor (measured by IC) is defined as the similarity between two terms.

Similarity Between Objects

Objects like genes or diseases are in most cases annotated to multiple terms of attribute ontologies. For example the *PIGV* protein from Figure 1.3 is annotated to ten different terms of the GO; one of them is shown in the Figure. Thus, in order to assess the similarity between two objects based on the attributes, it is necessary to define a measure between two sets of terms. Similar to the term-wise similarity, there exist several different methods to calculate this similarity, but a detailed discussion of these is beyond the scope of this work. A broader treatise on this subject has presented by Pesquita et al. [2009].

Thus, the focus in this work lies on a measure that is referred to as best-match average (BMA) [Pesquita et al., 2009]. Let O be an object, and let $\text{annot}(O) = \mathcal{T}_O = \{t_1, \ldots, t_n\}$ be the function that returns all the terms to which the object O is annotated. Then $|\mathcal{T}_O| = |\,\text{annot}(O)|$ is the number of terms to which O is annotated. The asymmetric BMA method calculates the similarity between two objects as

$$\text{BMA}_{asym}(O_1, O_2) = \frac{1}{|\mathcal{T}_{O_1}|} \sum_{t_1 \in \mathcal{T}_{O_1}} \max_{t_2 \in \mathcal{T}_{O_2}} \text{sim}(t_1, t_2) \; . \tag{1.4}$$

Note that Equation 1.4 is not symmetric [Couto et al., 2007], i.e., it is not necessarily true that $\text{BMA}_{asym}(O_1, O_2) = \text{BMA}_{asym}(O_2, O_1)$. Sometimes, the symmetric version of this measure is mentioned, which is then defined as

$$\text{BMA}_{sym}(O_1, O_2) = \frac{1}{2}\text{BMA}_{asym}(O_1, O_2) + \frac{1}{2}\text{BMA}_{asym}(O_2, O_1) \; . \tag{1.5}$$

1.4 Thesis Organisation

In the past decade, our understanding of biological systems has undergone a revolution, mainly due to the emergence of novel, high-throughput techniques, such as microarrays, and of genomic and systems-biological algorithms. The growing understanding of signal transduction cascades, metabolic pathways, and protein-protein interactions has led to the realisation that networks between different biological entities exist, and that those networks pervade all aspects of biology and of human health and disease. However, the mathematical and computational tools that have been so successful in the analysis of biological systems have only very recently begun to be applied to understanding human phenotypic abnormalities and hereditary syndromes.

This thesis involves the development of novel, bioinformatic tools designed to enable the integrated analysis of the "diseasome", which refers to the network comprised of genes, proteins, phenotypic abnormalities, and hereditary syndromes. It is aimed at showing new ways of analyzing the relationships between biological networks, networks of phenotypic abnormalities and syndromes using ontologies and graph-theoretic algorithms. An illustration of these topics and their interrelation is shown in Figure 1.5. The goal is to exploit these networks for gaining novel insights into the mechanism and structures underlying pathobiological processes on a systems level. This thesis shows that these approaches can give meaningful and helpful results.

Phenotypes in Cellular Networks

Following an introduction on the current approaches towards the identification of disease genes, this chapter presents a discussion of the inherent problems and challenges. Previous computational approaches developed for fostering disease gene detection by utilising protein interaction networks are presented. A novel approach for the prioritisation of genes being related to a particular phenotype or set of phenotypes is presented, which enables analyzing the relationships between biological networks and disease using graph-theoretic algorithms. Finally, the novel method is compared with other approaches, and its advantages are discussed.

Phenotypes in Ontologies

This chapter describes problems associated with the current solutions for storing and integrating information about human phenotypic abnormalities. It starts with a presentation of the *Human Phenotype Ontology* (HPO) as a means to overcome these problems by embedding the knowledge on phenotypic features into an ontology. This ontology specifies abnormal phenotypes and the semantic relationships between them. Two applications of the HPO are presented. First, the ontology, which can be used to automatically generate syndrome networks based on the phenotypic spectrum of the syndromes, is described. Afterwards, a novel, HPO-based approach for clinical diagnostics is presented. Finally, a statistically motivated measure that expresses the significance of semantic similarity search is also presented. This measure is applied to the clinical diagnostics testing, and its performance is evaluated.

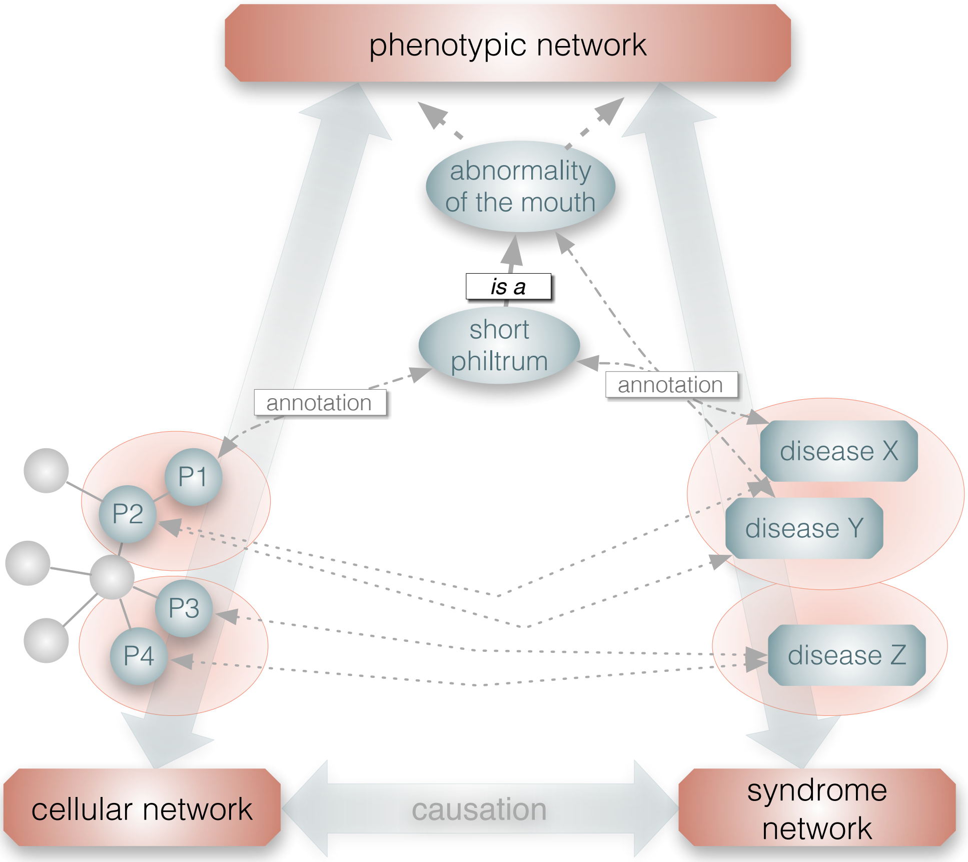


Figure 1.5: A visualisation of the topics covered throughout this thesis and the interrelations between them. In the beginning the main focus lies on mining *cellular networks* in order to identify novel disease genes that underlie human phenotypic abnormalities. Thus, there exists a *causation* relation in the figure. The next chapters of the thesis focus on semantic representation of phenotypic abnormalities associated with diseases, which are almost always defined by a broader spectrum of phenotypic features. This figure shows how genes and disease are *annotated* to these phenotypes. The semantics can be exploited for construction of *disease networks*. Finally, a combination of disease-, phenotype- and network-based approaches are presented in the last chapter.

Semantic Web Techniques for Genotype to Phenotype Discovery

The focus of this chapter is the integration of description logic (OWL) into phenotypic descriptions, in order to build semantic resources across ontologies, across domains, and across species. At first, the approach of bringing semantics to phenotype information is described. The subsequent section depicts an approach to ontological interoperability to create a huge, cross-species phenotype ontology. Finally, the application of this ontology to the analysis of copy number variations, which makes use of the technologies presented in preceding chapters, is presented.

The generally superior performance of the random-walk methods suggests that the clustering and neighborhood methods are too restrictive when defining their locality.

NAVLAKHA AND KINGSFORD [2010]

Chapter 2

Phenotypes in Cellular Networks

2.1 Identification of Gene - Phenotype Associations

At the time of this writing, there are more than 3,000 well-defined monogenic syndromes in humans. Understanding clinical phenotypes and their genetic origin is one of the principal objectives of genetic research. Furthermore, it is widely accepted that almost all medical conditions are directly or indirectly influenced by human genetic variation. Although this is a very important field of genetics and plenty of novel techniques have been developed in this field, the Online Mendelian Inheritance in Man (OMIM) database [Amberger et al., 2009] still lists over 1,500 Mendelian phenotypes and nearly 2,000 phenotypes with suspected Mendelian basis where the underlying molecular mechanism remain unclear (see Table 2.1).

However, the identification of those genes that influence or are even responsible for human phenotypes is a critical step towards gaining a deeper understanding of the pathological and biochemical disease mechanisms. Botstein and Risch [2003] indicate that the genes involved in inherited disease discovered so far were more apparent and thus easier to discover. It could be possible that the remaining tasks are much more complicated, due to the rarity of the phenotypes or the heterogeneity of the underlying genes. Also, the investigation of complex diseases, where several genes

	Total Number
Phenotype description, *molecular basis known*	3,231
Mendelian phenotype or locus, *molecular basis unknown*	1,774
Other, mainly phenotypes with suspected Mendelian basis	1,944

Table 2.1: Statistics from the OMIM database (August 17, 2011)

and modifiers contribute to a phenotype (e.g. in hypertension), will be even more challenging. But still, explicit phenotype-to-gene descriptions are essential components in order to both improve medical care and better understand gene functions, interactions, and pathways [Brunner and van Driel, 2004]. The identification of genes associated with phenotypic abnormalities is thus a goal of numerous research groups.

One starting point to investigate this link is called genetic mapping (linkage analysis or association studies), whereby disease phenotypes are correlated with specific markers on the genomic axis to narrow down the area of a chromosome that is associated with a specific phenotype abnormality [Altshuler et al., 2008]. This approach comes up with a genomic interval that may have a large number of candidate genes. Another approach is the functional candidate gene approach. The key ingredient is biological knowledge about the disease phenotypes under investigation and, if possible, also related molecular processes. This knowledge can, in turn, be used for the generation of hypotheses about the involvement of particular genes in the phenotype.

At any rate, most of the current efforts at disease-gene identification result in large sets of candidate genes. E.g., approaches involving linkage analysis or association studies result in a genomic interval of 0.5 - 10 centiMorgan containing up to 300 genes [Botstein and Risch, 2003, Glazier et al., 2002]. Sequencing large numbers of candidate genes remains a time-consuming and expensive task, and it is often not possible to identify the correct disease gene by inspection of the list of genes within the interval. Several approaches have been put forward in order to help scientists that are confronted with huge lists of candidate genes. These tools and methods provide a means for processing a set of genes in order to produce a ranked or truncated list with which the wetlab scientist has a high chance of finding the disease-causing gene as quickly as possible. The approaches use different data sources and different mathematical ideas and methods.

2.1.1 Disease Gene Families

The term *syndrome family* (or *disease gene family*) was proposed by Spranger [1985] as a concept for grouping clinically distinct diseases into groups of diseases that share important phenotypic features. These groupings are motivated by the presumed pathogenetic similarities that underlie the phenotypic commonalities. An example of this is the *Stickler-Kniest* family of skeletal dysplasias, where several years after the family had been characterised phenotypically, researchers showed that the two syndromes were caused by mutations in physically interacting proteins [Brunner and van Driel, 2004].

The reason for shared phenotypes in disease gene families is assumed to be based on shared molecular events. Different mutations of grouped diseases may affect different functional domains of a single protein, or they may disturb the function of proteins that in normal cells form physical complexes. Other reasons may be that proteins of one cellular pathway or a ligand-receptor crosstalk are impaired. These disease-phenotype groups are thus a key concept in research on the connection between phenotypic manifestations and their molecular pathological modifications.

For this work, 110 disease-gene families have been defined based on entries in the Online Mendelian Inheritance in Man (OMIM) database [Amberger et al., 2009]. This was done for

- Genetically heterogeneous disorders in which mutations in distinct genes are associated with similar or even indistinguishable phenotypes (*Monogenic*)

- Cancer syndromes comprising genes associated either with hereditary cancer, increased risk, or somatic mutation in a given cancer type (*Cancer*)

- Complex disorders that are known to be influenced by variation in multiple genes (*Polygenic*)

Additionally, domain knowledge and literature or database searches were used to select all genes clearly associated with the disorder at hand. The 110 families contain a total of 783 genes. These represent 665 distinct genes, because some genes are members of more than one disease family. The largest family (*Nonsyndromic hearing loss*) contains 41 genes. The smallest families (e.g. Achromatopsia) contain only 3 genes. Of the 110 disease gene families, 86 comprise monogenic disorders, 12 contain cancer syndromes, and 12 are polygenic diseases. A complete listing of the disease-gene families, with links to the corresponding entries in the OMIM database, can be found online.[1]

2.1.2 Guilt-by-Association

Guilt-by-association is a forbidden principle in the law, but due to often fuzzy and/or incomplete data, biomedicine can be seen as a rather 'uncertain' science. Even though biomedical technologies made major progress in both efficiency and resolution, scientists are often still unable to extract facts with a 100 % certainty. But in order to be able to expand knowledge, and to hopefully generate novel interesting hypotheses, scientists often use associations, although they may be weak and may represent an unreliable way for gathering facts [Altshuler et al., 2000]. When using this principle in biology, it is assumed that “associations” in the data (e.g., co-expression or protein interaction partners) of a gene are necessary in establishing “guilt”. The procedure mostly begins by identifying some genes that are known to be involved in a specific function that the researcher is interested in. Subsequently, one tries to identify other genes that also participate in the same process, pathway, or even physical complex (i.e.

[1] `http://compbio.charite.de/genewanderer/diseaseGeneFamilies.html`

the data-associations), and infers from those associations that those genes have the same or at least a similar function [Oliver, 2000]. It can be seen as an approach similar to the old proverb *"Show me your friends, and I'll know who you are."*.

2.2 PPINs for Associating Genes to Disease Phenotype

The focus in this part of the thesis is *guilt-by-association* methods that make use of protein-protein interaction networks (PPIN) in order to rank a list of candidate disease genes (C)[2], with the genes on top of the list most-likely being responsible for the disease phenotype. The working hypothesis of this approach is that networks form a hierarchical network of subnetworks that cooperatively determine the cellular behavior. The list of the candidate genes may have been identified by linkage analysis, association studies, or other approaches. Those methods require a non-empty set K of genes already known to be related to the biological question. A schematic illustration of the guilt-by-association approaches towards identification of genes related to specific phenotypes can be found in Figure 2.1. A list of symbols that are used throughout this chapter is given in Table 2.2.

It is assumed that genes linked to diseases with common (patho-) biochemical mechanism are interacting at the protein level in subnetworks, or at specific pathway steps. This is motivated by the fact that proteins are often part of larger protein machineries or complexes [Gandhi et al., 2006]. The corollary of this hypothesis is that interacting proteins often lead to related or similar phenotypes, and that those groups of genes, when related to a particular phenotypic spectrum, show a tendency to cluster in the interaction network [Barabasi et al., 2011]. Thus, most methods that use PPIN for disease gene identification or related tasks (e.g. protein function prediction) make use of these hypotheses and assume that the genes of C

[2] Note that in the context used here genes and proteins can be used synonymously. Many PPIN do not distinguish isoforms of genes, so that one can say that two genes interact with each other, although actually the gene's products interact with each other.

Symbol	*Definition*
K	A set of genes known to be associated with a particular phenomenon, e.g., phenotype, cellular process, disease.
C	A set of candidate genes from which one (or multiple) genes should be extracted or listed first when the set C is ordered by some method.
$G = (V, E)$	A graph with $m = \|V\|$ nodes and the edge set E.
g_i	A node of graph G, e.g. referring to a gene. The index i denotes which row or column corresponds to this node in the matrices $\mathbf{A}$, $\mathbf{A}'$, or $\mathbf{R}$.
$\mathbf{A}$	Adjacency matrix of a graph G.
$\mathbf{A}'$	The column normalised adjacency matrix of a graph G.
$\mathbf{I}$	The identity matrix.
$\vec{p}_0$	$m \times 1$ start vector with non-zero values for the nodes where the random walk starts. For n start nodes the values at the corresponding indices are set to $1/n$ for equal start probabilities for each of the nodes. The values for all other nodes are set to zero.
$\vec{p}_\infty$	$m \times 1$ proximity vector. The value $\vec{p}_\infty[i]$ is the probability of the random walker being in node i after an infinite number of time points t.
$\mathbf{R}$	The precomputed random walk matrix, where each entry $\mathbf{R}[i, j]$ contains the probability of being in node i after an infinite number of steps, given that the random walk started in node j.

Table 2.2: A summary of the symbols and notations used throughout chapter 2. Short definitions are listed as well.

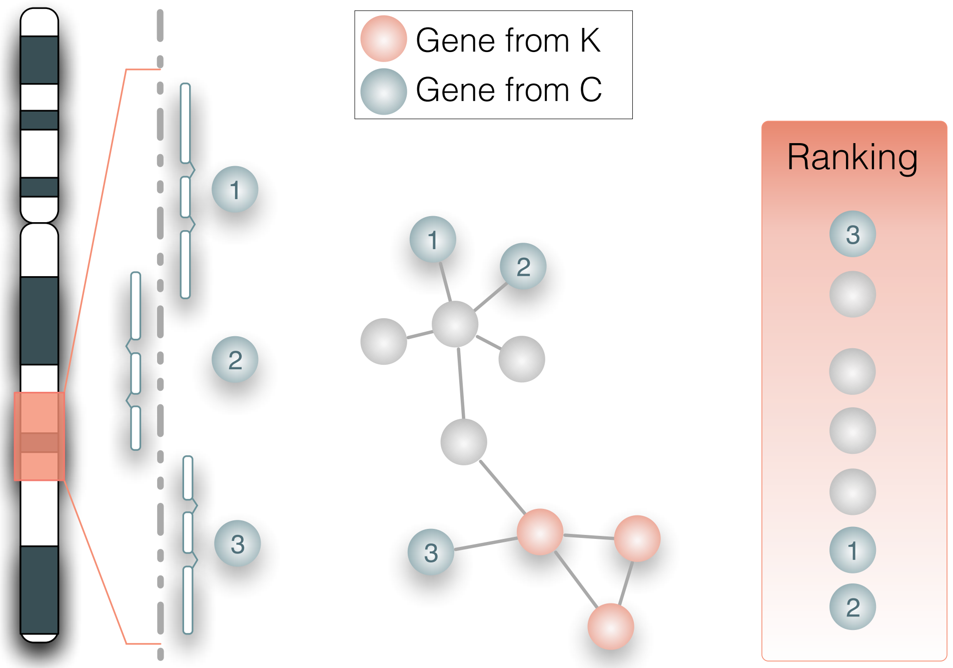


Figure 2.1: The general idea of PPIN-based *guilt-by-association* methods, which aim at prioritising genes for their involvement in a particular phenotype. A set of candidate genes C is identified, e.g., by linkage analysis and afterwards investigated for association with genes already known (K) to be involved. The candidates ranked first are then selected for further investigation in wet labs.

are likely candidates if they are located in the PPIN *vicinity* of the genes from K. During the last few years, several methods to measure this *vicinity* have been used and evaluated for their applicability in human disease gene identification.

The hypothesis of the method developed by Oti et al. [2006] is that disease genes tend to interact directly with other known disease genes, an observation previously described by Gandhi et al. [2006]. If a gene from C directly interacts with a gene from the set K, it is considered to be the gene bearing the disease-causing mutation. Note that, in their work, the set K is the set of all genes known to be associated with any known hereditary disease. This means that every gene suggested by this method interacts with a gene causing any disease, but this disease may be totally unrelated to the disease investigated by the user.

An extension of this approach, developed by Xu and Li [2006], identifies patterns of topological graph measures of the genes of K, in order to set up associations to the genes of C. Genes from C which show similar patterns as the genes from K are considered to be likely disease relevant. Their method makes use of several local topological features of disease genes, such as the degree, or the frequency of other disease genes in the set of direct neighbors and in the set genes having a shortest-path distance of two. Also, they calculated the average shortest path distance of disease genes to all other disease genes. They chose the *k-nearest neighbours* machine learning algorithm to perform the classification, and performance measures showed again that PPIN can be a good means for identifying genes related to disease phenotypes.

A very similar method was developed by George et al. [2006]. One aspect of their work was to identify novel disease genes by finding proteins that are linked with the product of a known disease gene in the same pathway, or by interaction in the PPIN. In contrast to the previously mentioned methods, they restricted K to the set of genes known to be involved in the disease phenotype. The list K was compiled by incorporating all disease phenotypes where at least three disease genes have been identified.

A gene interaction network was compiled by Franke et al. [2006], for which they integrated interactions derived from protein-protein interactions, as well as regulatory, functional, and metabolic data. Their method assumes that genes involved in a particular disease are involved in only a few different biological pathways. Thus, it is expected that the causing genes from different susceptibility loci are clustered in the network. Using this hypothesis, the system analyses the user-defined (at least three) genomic intervals and prioritises the candidate genes found to be closely related to each other in their network.

All of the previous methods use only local properties of PPIN to define similarity or association between the network's nodes. The hypothesis here is that incorporating the global structure of the network will improve the applicability of PPIN for linking genes to phenotypes. This hypothesis can be explained, for example, by looking at Figure 2.2 and asking if the proteins C and A are closer to each other than C and G are. The answer of previous approaches that used the shortest-path measure would be that both pairs of proteins are equally similar; whereas, by visual inspection, one would say that C and G are more closely associated with each because of the multiple paths that connect those two.

Thus, it may in reality not really be true that *"Show me your friends, and I'll know who you are"* is enough in order to reliably generate hypotheses using the underlying networks. More sophisticated graph and network algorithms offer tools for understanding cellular physiology on a systems level and also for understanding cellular responses to disease.

2.3 Random Walks on PPIN

The cell's numerous tasks and life-sustaining responses to internal and external signals are mostly accomplished through interactions of cellular components. These components that represent the nodes of the network may correspond to proteins, genes, metabolites, non-coding RNA's or others. The resulting networks can be used to gain deeper insight on how complex molecular tasks are accomplished. Furthermore, they can be used to understand what mechanisms and patterns apply when small disturbances of single nodes spread along the network's links and affect the function in other areas of the system [Barabasi et al., 2011]. Thus, the work-

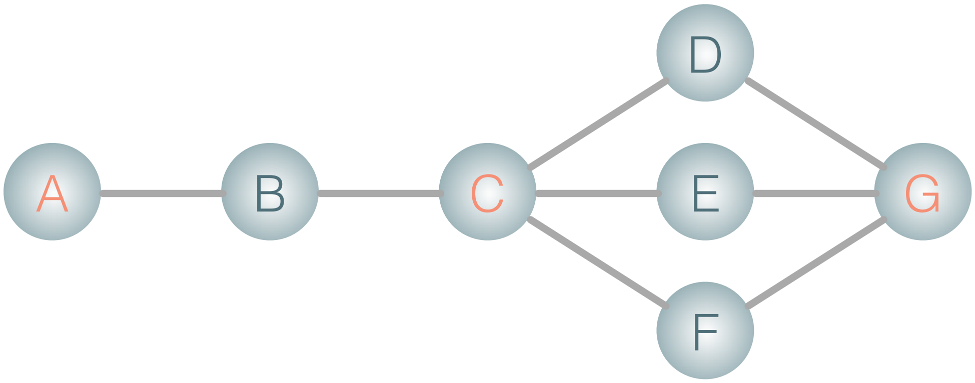


Figure 2.2: Small exemplary protein-protein interaction network (PPIN) consisting of seven proteins (A,B,C,D,E,F,G).

ing hypothesis here is that behind each cellular function there is a highly interconnected area of the cellular network, and that clinical abnormalities found in genetic and other diseases are the result of breakdown of one or multiple of such functional modules. Such modules are perhaps best described as subnetworks of a highly complex network connecting many cellular components and functions. A major problem in using PPIN for this task is the enormous network-complexity with the human gene network having around 25,000 nodes and several orders of magnitude more edges.

Protein-protein interaction networks can be represented as graph structures, which in turn can be visualised, as e.g. in Figure 2.2. In this Figure, a small toy protein-protein interaction network consisting of seven proteins (A-G) is shown. The problem is that, due to the enormous size of real cellular networks, a systems-level understanding of the inherent properties requires more sophisticated mathematical approaches. In order to enable this, graphs are often presented as adjacency matrices, and methods from linear algebra are used.

An adjacency matrix $\mathbf{A}$ for a given graph $G = (V, E)$ is defined as the $m \times m$ matrix, where m is the number of nodes. Each entry at row i and column j is defined as

$$\mathbf{A}[i,j] = \begin{cases} 1 & \text{if the node } i \text{ and the node } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

The adjacency matrix $\mathbf{A}$ for the graph shown in Figure 2.2 (with the nodes $V = \{A, B, C, D, E, F, G\}$) thus is:

$$\mathbf{A} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \end{array} \begin{array}{c} \begin{array}{ccccccc} A & B & C & D & E & F & G \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array} . \tag{2.2}$$

As the graph G is undirected, the adjacency matrix is symmetric.

From the matrix $\mathbf{A}$, the column normalised adjacency matrix ($\mathbf{A}'$) can easily be computed by normalising every column so that the entries of each column sum up to 1, i.e.

$$\mathbf{A}'[i,j] = \frac{\mathbf{A}[i,j]}{\sum_{k=1}^{m} \mathbf{A}[k,j]} . \tag{2.3}$$

The matrix $\mathbf{A}'$ can be used for sophisticated graph analysis, such as the random walk analysis. The idea of the random walk computation is to exploit the global structure of a network by simulating and tracking the behaviour of an artificial, random (or 'drunk') walker. The random walker starts at a defined node (or a set of nodes) at timepoint t and randomly visits adjacent nodes. These walking steps are performed in every tick of time and this process is repeated for a specific fraction of time.

For each tick of time ($t \rightarrow t+1$), the computation determines the probability of the random walker being located at a node g_i at timepoint t and being located at node g_j at timepoint $t+1$. This is done in terms of matrix operations by defining a vector $\vec{p}_t$ of size m with each entry $\vec{p}_t[i]$ giving the probability of the random walker being at node i at timepoint t. To compute the probabilities for the timepoint $t+1$, a simple matrix operation is performed:

$$\vec{p}_{t+1} = \mathbf{A}' \times \vec{p}_t \ . \tag{2.4}$$

Looking at the example from Figure 2.2 and assuming the random walker is with a probability of one located at node C ($i = 3$), the computation becomes:

$$\begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.333 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.333 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.333 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0.25 \\ 0 \\ 0.25 \\ 0.25 \\ 0.25 \\ 0 \end{pmatrix} . \tag{2.5}$$

In this example, it is easy to see that if the walker starts at node C and is randomly going to one of the four adjacent nodes, then there is a chance of 1/4 for the walker to be in one of these nodes after taking this step.

A totally random exploration of the walker is not useful, since the association to the start nodes would be lost after a few time points. Instead, a *restart probability* (r, $0 \leq r \leq 1$) is introduced that determines the extend to which the random walk is reset to its initial configuration. This means the walk starts again at the defined start nodes. For a vector of starting probabilities $\vec{p}_0$, the state probabilities $\vec{p}_{t+1}$ can be computed iteratively [Can et al., 2005]:

$$\vec{p}_{t+1} = (1-r)\mathbf{A}' \times \vec{p}_t + r \times \vec{p}_0 \ . \tag{2.6}$$

For $t \rightarrow \infty$ the state probabilities converge to a stationary distribution $\vec{p}_\infty$. In the example of Figure 2.2, the goal was to have a measure that can discriminate the distance between node C and A compared to the distance between node C and G. The intuitive feeling of C and G being closer, is well reflected using the random walk method, i.e. the probability of being in node A when starting in node C is 0.006, whereas the probability of being in G is approximately three times higher (0.019).

In several projects and applications, the steady-state probability vector is determined at the time of the query by performing the iteration from equation 2.6 until the difference between the subsequent vectors $\vec{p}_t$ and $\vec{p}_{t+1}$ is negligible, so that the computation can be assumed to be converged. Normally this is measured by the $L1$ (or $L2$) norm, and changes of the $L1$ norm below a value of 10^{-12} are taken as an adequate threshold [Can et al., 2005]. Of course, computation time is directly influenced by this threshold, since lower thresholds take a higher number of iterations until the convergence is reached. But the additional computation time is rewarded in an increase in precision.

The data underlying PPINs is updated regularly, but only at longer intervals. An often used resource is the *Search Tool for the Retrieval of Interacting Genes/Proteins* (STRING) [Szklarczyk et al., 2011]. The STRING database (Version 4.0 to 9.0) has been subject to updates on average every 250 days.[3] Given that PPINs are rather stable entities, a major speedup query time can be achieved by preprocessing the network. Furthermore, it can be shown that it is possible to provide an exact solution rather than the iterative approximation given above. Approaches that optimise both the preprocessing and the query time exist, but implicate a decrease of performance in terms of precision [Tong et al., 2006]. Precision is of high importance for the applications presented here, and the cellular network with less than 20,000 nodes can be precomputed in an adequate time.

The strategy, thus, is to precompute a random walk matrix, that can be used at query time. The first step is to set $\vec{p}_{t+1}$ identical to $\vec{p}_t$, which is the definitional criterion of convergence. Applying this rule to equation 2.6 gives:

$$\begin{aligned}
\vec{p}_t &= (1-r)\mathbf{A}' \times \vec{p}_t + r \times \vec{p}_0 && (2.7)\\
\vec{p}_t - (1-r)\mathbf{A}' \times \vec{p}_t &= r \times \vec{p}_0 && (2.8)\\
(\mathbf{I} - (1-r)\mathbf{A}') \times \vec{p}_t &= r \times \vec{p}_0 && (2.9)\\
\vec{p}_t &= (\mathbf{I} - (1-r)\mathbf{A}')^{-1} \times r \times \vec{p}_0\,. && (2.10)
\end{aligned}$$

Note that the matrix $\mathbf{I}$ denotes the identity matrix. Equation 2.10 allows

[3] Data taken from `http://string-db.org/server_versions.html`.

one to determine the exact solution of the random walk by computing:

$$\vec{p}_\infty = \underbrace{(\mathbf{I} - ((1-r)\mathbf{A}'))^{-1} \times r}_{\text{Precomputation as } \mathbf{R}} \times \vec{p}_0 \, . \tag{2.11}$$

As stated above, the strategy is to precompute a matrix **R** which can be used at query time. Thus, the precomputation becomes:

$$\mathbf{R} = (\mathbf{I} - ((1-r)\mathbf{A}'))^{-1} \times r \, . \tag{2.12}$$

The matrix **R**, thus, stores in each column j the random walk similarity of all nodes from node j, so that each entry $\mathbf{R}[i, j]$ contains the probability of a random walker starting at node j and being at node i after an infinite number of steps. Note that this matrix is not symmetric, i.e. it must not be true that $\mathbf{R}[i, j] = \mathbf{R}[j, i]$. This can easily be seen from the example in Figure 2.2. There are only two outgoing edges from node B but four edges leaving node C. Calculating **R** with $r = 0.75$ for this graph will give $\mathbf{R}[3, 2] = 0.1$ and $\mathbf{R}[2, 3] = 0.05$.

In order to compute the matrix **R**, one can make use of highly efficient libraries such as jblas (`http://jblas.org`), which is a light-weight wrapper for state-of-the-art, highly optimised linear algebra implementations, namely BLAS, LAPACK, and ATLAS. Given a PPIN of around 11,500 ($m = 11,621$), the precomputation requires less than 7 GB RAM and takes less than 15 minutes to calculate.

Once this matrix is computed, it can be written to the computer's hard disk for later use. Due to the huge amount of nodes and the strong connectivity of many real world networks [Watts and Strogatz, 1998], this matrix is very dense and contains a large fraction of very small values. Writing this matrix to disk may require a lot of space. Thus, it is advisable to store the matrix in binary format or take the logarithm of each matrix element beforehand and round this value to, e.g., four digits after the decimal point.

In applications, the matrix **R** can then easily be read from hard disk and kept in memory for fast access times. For a given set of start nodes, a linear combination of the corresponding columns of **R** can then easily be computed. To do so, the corresponding vector $\vec{p}_0$, that contains the start node probabilities, is then multiplied with **R**:

$$\vec{p}_\infty = \mathbf{R} \times \vec{p}_0 \ . \tag{2.13}$$

In the simplest setting, the probabilities of $\vec{p}_0$ are set to $1/n$ for the n start nodes and to 0 for the others. The resulting vector $\vec{p}_\infty$ can be used to rank all nodes by network similarity to the set of start nodes.

Note that other measures exploring the global structure of networks exist. For example, the PageRank-algorithm [Page et al., 1998] is used for information propagation across the edges of a network. This method is different from the approach presented here, in that it represents no measure of *vicinity* between nodes; rather, the PageRank of a node gives an approximation of importance or centrality.

2.4 Protein-Protein Interaction Data

For disease gene identification, a major problem is the incomplete knowledge about the human genes from C, because for every gene from C that has no interaction data, no PPIN-based method can make any statement. This motivates an attempt to increase the coverage by means of mapping interaction knowledge from other organisms. Also, there exist methods that expand interaction data by predicting interactions by, e.g., text mining. As mentioned, completeness of interaction data is desirable, but it is important as well to investigate the dependence of the results from those procedures. Thus, several different PPINs were compiled. To construct the PPINs, protein-protein interaction datasets from five organisms were used:

- *Homo Sapiens*
- *Mus musculus*
- *Drosophila melanogaster*
- *Caenorhabditis elegans*

- *Saccharomyces cerevisiae*

The datasets were downloaded from the Entrez Gene website (July 2007). These datasets comprise interactions extracted from HPRD [Peri et al., 2004], BIND [Alfarano et al., 2005], and BioGrid [Stark et al., 2011]. Additional interactions were extracted from IntACT [Kerrien et al., 2007], and DIP [Salwinski et al., 2004]. The protein interactions were mapped to the genes coding for the involved proteins, and redundant interactions stemming from multiple data sources were removed. The interaction relation was modelled as a binary relation, i.e. all interactions are treated equally as being present or not. Furthermore, self-interactions are not considered in this analysis.

Interactions stemming from the four nonhuman species were mapped to homologous human genes identified by Inparanoid [Ostlund et al., 2010]. Inparanoid provides a score cutoff that can be raised to exclude borderline inparalog cases [O'Brien et al., 2005]. In order to increase the reliability of the used ortholog predictions, an Inparanoid score threshold of 0.8 was used to filter. If both interaction partners could still be mapped to human proteins, the interaction was used.

This work also included data from the STRING database. STRING is a comprehensive dataset containing functional links between proteins on the basis of both direct experimental evidence for protein-protein interactions as well as interactions predicted by comparative genomics and text mining. STRING uses a sophisticated scoring system that is intended to reflect the reliability of predicted interactions. In this work, interactions with a score of at least 0.4 are included. According to Szklarczyk et al. [2011] this corresponds to a medium-confidence network. For the analysis, interactions solely identified through text mining are excluded, since those interactions are likely introducing a bias. This bias is probable because a given gene is likely to be intensively studied in the years following its identification as a human disease gene. In Table 2.3, the different networks are summarised.

Network	*Number of Interactors*	*Number of Interactions*
Human	9,169	35,910
Mapped:		
Worm	684 (146)	831 (768)
Mouse	1,412 (78)	1,972 (853)
Fruitfly	2,176 (590)	4,930 (4,613)
Yeast	1,557 (441)	33,396 (32,855)
Total human and mapped	10,231	74,885
All Data Sources excluding text mining data	11,673	133,612

Table 2.3: Networks tested in this work. *Mapped* indicates protein-protein interaction data mapped to homologous human proteins. The number of new interactors/interactions that were added to the interaction network by mapping is shown in parentheses. *All Data Sources* denotes the STRING data, human and mapped interactions.

2.5 Comparison of Disease Gene Prediction Methods

As mentioned above, PPINs have been used for the prioritisation or identification of the phenotype related genes from C. These use the prior knowledge that the genes from K are associated to the disease (or any disease). The methods assign scores s to each of the genes of C and rank them according to s in ascending order. Given the matrices $\mathbf{A}$ and $\mathbf{R}$, as well as the gene sets C and K, three approaches can be compared.

Direct interactions (DI)

As defined by Oti et al. [2006], all genes from C that *directly* interact with a gene from K are considered a candidate, i.e. ranked first place. Note that K comprises all genes associated to any disease and not only the disease under investigation. For each gene g_i from C the score is thus defined as:

$$s_{DI}(g_i) = \begin{cases} 1 & \text{if } \exists g_k (g_k \in K \wedge \mathbf{A}[i,k] = 1) \\ \infty & \text{otherwise} \end{cases} . \tag{2.14}$$

Shortest Path (SP)

The genes from C are ranked by their single *shortest path* distance to a gene from K. Let $SP(g_i, g_j)$ be the function that returns the shortest path distance of two genes then for each gene g_i from C the similarity is calculated as

$$s_{SP}(g_i) = \min_{g_k \in K} \left(SP(g_i, g_k) \right) . \tag{2.15}$$

Random Walk with Restart (RWR)

Intuitively, the random walk starts with equal probabilities from each of the known disease-gene family members (C) in order to search for an additional disease gene family member (K). Each gene from C is ranked by the random walk with restart probability. Therefore, the start vector is created so that for each g_k from K the corresponding entry $\vec{p}_0[k]$ is set to $1/n$, where $n = |K|$. The steady state probabilities ($\vec{p}_\infty$) are then calculated according to equation 2.13. For each gene g_i from C, the similarity to K is then

$$s_{RWR}(g_i) = \vec{p}_\infty[i] . \tag{2.16}$$

2.5.1 Comparison Strategy

The disease-gene families described above were used for testing. For each family, a leave one out cross validation is performed. This means that one gene g_i of the disease gene-family is left out, whereas the others were chosen as the set K. An artificial linkage interval is constructed as the set of genes containing the first 99 genes located nearest to g_i according to their genomic distance on the same chromosome. Note that g_i was added as well. The goal was to rank g_i, which is the *true* disease gene, as high as possible.

2.5.2 Comparison Evaluation

Enrichment Score (ES)

For a set C with n genes, the enrichment score is calculated as

$$ES(a, n) = \frac{n/2}{a} \,, \tag{2.17}$$

where a is the actual rank of the *true* disease gene that has been left out of the disease-gene family. The reason for calculating $n/2$ is that, with random ordering, a user would expect to find the sought-after gene in the middle of the list, which would correspond to an enrichment score of 1. If a ranking method gives the *true* disease gene the highest ranking and it is sequenced first, there is an enrichment of 50-fold, since the size of C is 100 in this thesis. Disease genes for which no interaction data are available are given a rank of 100 (and therefore an enrichment score of 0.5). The fact that some genes of C have no interactions at the protein level was not corrected. Note that two ways of determining the rank a in cases of ties are used. In the first case, each gene is given the mean rank of all tied genes. The second option is the worst case scenario, in which it is assumed that the *true* disease gene is the last to be sequenced from the set.

In Figure 2.3 the comparison of the three methods in terms of enrichment is shown. The results are split according to disease gene family classes (monogenic, polygenic, cancer). A clear advantage of global versus local network search algorithms can be seen. This is especially true for the disease-gene families classified as cancer or polygenic diseases. The RWR method ranked all genes of 43 disease gene families first. This means, in almost 40 % of the disease gene families this method always achieves a perfect, 50-fold, enrichment. On average, the random walk approach achieved a 44-fold enrichment for all 783 disease genes when using all data sources, including the text-mining data of STRING. When excluding interactions found by text mining, this enrichment value drops to 27-fold (shown in Figure 2.3). This is a hint towards a systematic bias towards disease gene interactions (interactions between disease genes) introduced by text mining data.

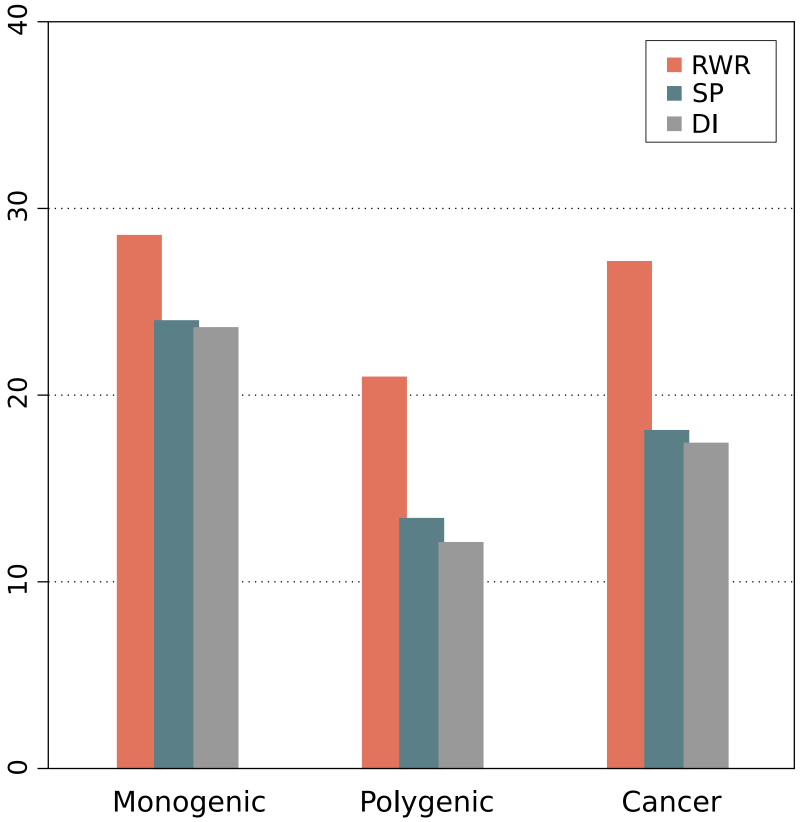


Figure 2.3: Comparison of enrichment scores for the tested methods. Here the network containing all-interactions but excluding the STRING text-mining data is used. The disease gene families are classified into three groups. Note that in the case of multiple genes receiving the same score, each gene is given the mean rank of all tied genes. A clear performance improvement of global (RWR) versus local (SP,DI) measures can be seen.

The advantage of global network measures is achieved through an increased resolution of the rankings. In the text mining filtered network, genes have a high average degree of 22.9, which means they have almost 23 direct interaction partners in the network. In addition, there is a mean path length of only 3.7 between randomly chosen pairs of genes. This means, for all nodes one can expect a high number of direct interaction partners, and pairs of nodes are rarely far apart in the interactome. A direct consequence of this is that local similarity methods, such as DI and SP, are observing a high proportion interactions that have a high chance of being unrelated to the problem under investigation. Here, in 61 % of the cases in which the DI method correctly identifies the true disease gene, it additionally identifies other unrelated genes with a direct interaction to a known disease gene. On the other hand, in only 1.4 % of the cases in which the true disease gene is ranked in first place by the RWR method, another unrelated gene was also given the same score. Obviously, the RWR method is better able to discriminate among genes within a dense network of interactions by incorporating the global network structure into the calculations. This is true independently of how the rank of the of the *true* disease gene is determined in cases of tied rankings (see Figures 2.3 and 2.4).

Receiver-Operating Characteristic (ROC)

Another often used measure of the performance of the methods is the receiver-operating characteristic (ROC) analysis. It plots the true-positive rate (TPR) versus the false-positive rate (FPR) subject to the decision threshold separating the prediction classes. During ROC analysis every possible decision threshold is tested. The TPR/FPR is the rate of correctly/incorrectly classified samples of all samples classified to class +1. For evaluating rankings of disease-gene predictions, ROC values can be interpreted as a plot of the frequency of the disease genes above the threshold versus the frequency of disease genes below the threshold, where the threshold is a specific position in the ranking [Aerts et al., 2006]. In order to compare different curves obtained by ROC analysis, the area under the ROC curve (AUROC) is calculated for each curve.

In order to compare the different prioritisation methods (RWR, SP, DI), the ROC curves were plotted in Figure 2.4. For this plot, the network that contains all interactions but excludes the STRING text-mining component is used. The curve labeled "random order" displays the results obtained by the sequencing of genes within the linkage interval at random, i.e., without use of any prioritisation method. As for the enrichment evaluation, a clear advantage of the global network similarity measure can be seen. The random walk approach achieves an AUROC value of 91.2 %, which is more than 7 % higher than the AUROC of the SP approach. SP and DI achieve an AUROC of 84.1 % and 73.2 %.

These results are independently confirmed in the work of Navlakha and Kingsford [2010]. A comparison of seven PPIN-based methods for determining gene-to-phenotype associations is performed. It was again found that the RWR approach individually outperforms clustering and neighborhood approaches.

In recent years, it became more and more evident that networks pervade almost all aspects of human health. Instead of only looking at lists of disease genes, it is important to examine subnetworks of the cellular system and apply network approaches to the analysis of cellular functions. This is necessary for a detailed understanding of complex disease mechanisms [Barabasi, 2007]. Lim et al. [2006] showed an example for the common observation that proteins mutated in phenotypically similar diseases form highly interlinked subnetworks within the larger protein interaction network. Compared to analyses that measure only direct interactions and shortest-path distances, random walk analyses in PPINs enable one to take the global structure of the interactome into account. In the test case of identifying phenotype-gene associations, these have a clear performance advantage.

In summary, the analysis suggests that the assumption that phenotypically similar diseases are associated with disturbances of subnetworks within the protein interactome is correct. Also, it becomes obvious that the exploration of global network structures with appropriate graph-theoretic algorithms is an important approach towards understanding the biological foundations of phenotypes in PPINs.

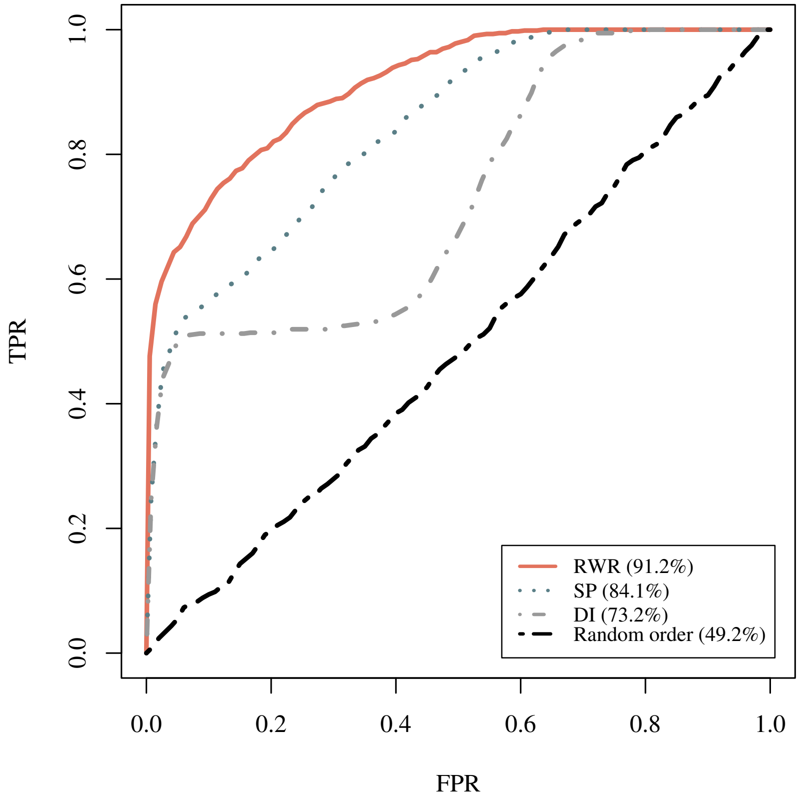


Figure 2.4: Rank ROC curves for the 110 disease-gene families and the three tested approaches. Intuitively, the area under the ROC curve (AUROC) reflects the false-positive rate needed in order to achieve a certain level of sensitivity, with a perfect classifier having an AUROC of 100 % and a random classifier having an AUROC of 50 %. Note that disease genes with no interaction data were excluded for this evaluation. Here, if a prioritisation method assigns an identical score to a set E of multiple genes and the *true* disease gene is contained in E, it is assumed that the *true* disease gene is the last to be sequenced from the set E.

2.6 KR in Molecular Biology vs. Clinical Dysmorphology

In this work, disease-gene families have been manually compiled by clinical experts. These lists are necessary, because all *guilt-by-association* methods rely on some predefined set of entities for which information, e.g. on their function, exists (here K). The properties of this set can then be used to calculate some kind of association, in order to establish a "guilt" of other entities with, e.g., unknown function. Besides that, these predefined sets are also an adequate means to test and compare the performance of different methods, e.g. by leave-one-out testing.

Setting up lists of disease-gene families was possible in this work because mainly common disorders were used, for which the molecular etiology has been determined. But in reality it is very often not possible to compile lists of genes that are known to be related to a particular disease. In the context of genotype-phenotype associations, it may well be the case that rare and orphan diseases are investigated. These diseases have a very low prevalence, i.e. they affect a small percentage of the population. It can thus be impossible to suggest candidate genes for rare and orphan diseases without a known molecular basis. All information that scientists or physicians have, in this case, is a list of phenotypic features of the patient(s).

One approach to enable *guilt-by-association* techniques for those cases would be to make use of this information in order to identify other diseases that share the phenotypic spectrum of the patient(s) and that are better characterised in terms of their molecular basis. This information of the molecular basis can then be used to measure "association" of candidate genes (C) and subsequently establish a guilt for the "most associated" genes. The bottleneck then becomes the lack of an approach to reliably determine phenotypic similarity.

In recent years, novel research techniques in biology have accumulated enormous amounts of data. In order to integrate, unify, reliably retrieve, and compute on this data, vocabularies with a well-defined and explicit semantics are needed. Knowledge that is stored in a structured way can then be linked to the molecular databases [Bard and Rhee, 2004]. In the area of cell biology, the Gene Ontology (GO) is successfully applied to solve this problem, which in turn motivated the introduction, promotion, and success of several biomedical ontologies (bio-ontologies). But the rapid changes on the molecular biology side of genotype-phenotype research are not reflected on the clinical side. This means that one of the most limiting factors for the coming era of next generation sequencing for personalised medicine will be clinical knowledge representation, i.e. the clinical analysis of affected individuals is largely hampered by the lack of standards.

Although abnormalities in phenotypes are among the most reliable manifestations of altered gene functions, research using systematic analysis of phenotype relationships to study human biology is still in its infancy [Lussier and Liu, 2007]. The current "standards" for representing phenotypic features of patients in publications are mainly free text, or in the best case, tables that summarise this information. Hence, the literature becomes a phenotypic "database". But standardised clinical descriptions and systematic storing procedures are not available, so that easy tasks, such as the identification of phenotypically overlapping syndromes, are hampered. In conclusion, one can infer that, if a decade ago the systematic storage of molecular biological knowledge was the major bottleneck, it now becomes obvious that the lack of standards for phenotypic information is currently the major bottleneck in genotype-phenotype research [Biesecker, 2005]. Systems to allow for calculating similarity between and grouping of phenotypes, patients, and syndromes are essential. This will become especially important with next generation sequencing techniques and personalised medicine being a realistic objective for the coming years.

. . . that prose is a poor medium by which developmental anomalies should be codified and recorded, and for this reason it is necessary to begin to construct a standardized nomenclature of clinical phenotyping.

BIESECKER [2005]

Chapter 3

Phenotypes in Ontologies

3.1 Bridging the Gap of Knowledge Representation

As mentioned above (see Section 2.6), there exists a huge gap between the sophisticated methods to store, integrate, and analyse the data of molecular biology and the counterpart for phenotypic knowledge. As a result, development of phenotypic databases dramatically lags behind the rapid advance in genomic databases [Lussier and Liu, 2007]. Thus, it is a major requirement to develop approaches towards standardisation of phenotype terminology, measuring and annotation techniques, and the encoding and recording of data [Biesecker, 2005]. The development of novel appropriate analytic tools should, of course, be part of this process.

3.1.1 Problems of Missing Standards

The great majority of human hereditary syndromes have been described in detail in the Online Mendelian Inheritance in Man (OMIM) database [Amberger et al., 2009]. It is often regarded as the single most valuable resource of human genetics. Recently, hierarchical systems that are based on the clinical descriptions in OMIM have been generated by automatic text mining [van Driel et al., 2006, Masseroli et al., 2005]. But those approaches have been hampered by the lack of controlled vocabularies. Also, the poor consistency of annotations and the lack of well-defined relationships between phenotypic features obstructs progress in the field of genotype-phenotype research.

To give an example, there are four phenotypic features for four different syndromes in OMIM (accessed November 17th 2011):

- *generalized amyotrophy* (OMIM:601162)
- *generalized muscle atrophy* (OMIM:613561)
- *muscular atrophy, generalized* (OMIM:609241)
- *muscle atrophy, generalized* (OMIM:258450)

Computational text mining based approaches might not easily recognise these four descriptions as synonyms.

Homonymic terms are another problem, which inherently lead to false-positive hits during text-based searches. For example, querying OMIM using the word "ventricle" will return both, entries mentioning abnormal *brain ventricles* and entries containing *heart ventricle* abnormalities, although one can assume that the user wanted only one of both classes of abnormalities to be returned.

OMIM provides a categorisation of the phenotypic features according to the affected organ system. For example, in the description of Marfan syndrome (OMIM:154700), *aortic root dilatation* is listed under the category *cardiovascular* and the subcategory *vascular*. But the fact that this hierarchy is very flat (only three levels deep) prevents more elaborate computational analyses. For example, the category *nose* lists the features *hypoplastic nasal septum*, *smooth philtrum*, and *hypoplastic philtrum*. Cases like this prevent automatic analyses from detecting that *smooth philtrum* and *hypoplastic philtrum* are more closely related to one another than to *hypoplastic nasal septum*.

3.1.2 Need for Standardisation

Figure 1.5 illustrates a subset of possibilities that are opened when such a harmonisation of data is achieved. First of all, given standardised terms of phenotypic abnormalities, a network of phenotypic features can be constructed, in which the relationships and groupings between the terms are made explicit. On the side of cellular networks, this allows one to locate modules that are correlated with specific phenotypic features. Syndrome networks can be created by identifying relationships among disorders based on phenotypic information. This can be used to establish syndrome families, from which functional genomic relationships can be inferred when phenotypic similarities within disease families are assumed to be related to dysfunction of a regulatory network, such as a signaling pathway or a biochemical module.

In summary, phenotypic analysis is of great importance for the understanding of physiology and pathophysiology of cellular networks because it can offer clues about groups of genes that, in cooperation, make up modules or pathways in which errors or dysfunction possibly lead to related or similar phenotypic consequences. There are several unresolved issues surrounding the computational description and analysis of human phenotypes [Robinson et al., 2008]. In order to gain a detailed understanding of the relationship between genotype and phenotypes, accurate, precise, and comparable phenotypic information is of major importance. Although free-text descriptions in natural language allow for the highest expressivity, this results in data that are very difficult to compute over [Mungall et al., 2010].

3.2 The Human Phenotype Ontology

There are several considerations that suggest that an ontological description of human phenotypes has distinct advantages. Therefore, an ontology to describe human phenotypic abnormalities was developed. The *Human Phenotype Ontology* (HPO) has the main goal of covering the complete set of phenotypic abnormalities that are commonly encountered in human monogenic syndromes.

This was done by using the file `omim.txt`[1] from the OMIM website.

[1] url: `ftp://ftp.ncbi.nih.gov/repository/OMIM/ARCHIVE/omim.txt.Z`

This file was parsed with a suite of Java programs and Perl scripts, and the “Clinical Synopsis” (CS) section was extracted for each syndrome. The basic hierarchy structure provided by OMIM was maintained. For every syndrome, a list of free-text phenotypic features was thus obtained. Only those features that were used at least twice in the complete `omim.txt` were kept. For example the term *aortic root dilatation* was kept, since it is used to describe a number of diseases such as Marfan syndrome and Ehlers-Danlos syndrome, type I (OMIM:130000). On the other hand, terms such as *medial rotation of the medial malleolus*, which is only used once as a feature of Marfan syndrome, were excluded.

Using the software OBO-Edit [Day-Richter et al., 2007], the HPO was initially constructed on the basis of the list of CS descriptions from the file `omim.txt`. OBO-Edit is an ontology visualisation and editing tool, which is particularly useful for the viewing, editing and creation of biomedical ontologies. An HPO-term was created for every distinct phenotypic description of the CS-list extracted from OMIM. Synonymic descriptions were merged into one term. For example, the four descriptions listed in Section 3.1.1 were merged into the HPO term *generalized amyotrophy* (HP:0003700).

In order to support manual curation, clusters of textually similar descriptions were suggested by adopting the Smith-Waterman alignment algorithm [Smith and Waterman, 1981] for this task. This approach was also used to suggest synonymous or more specific descriptions of HPO terms that occurred only once in the `omim.txt` file. In the case of more specific descriptions, a child term was created during manual curation. The domain knowledge of the manual curators was also used to define the semantic relationships to more general terms. Due to this, the HPO has a very deep hierarchy, compared to the three levels used by OMIM. For example, the term *aplasia/hypoplasia of the outer ear* was created and the related terms were linked there. The terms in the HPO are related to parent terms by `is_a` (`subClassOf` in OWL) relationships and the terms are allowed to have multiple parents. This allows for the expression of various aspects of phenotypic features, as in, for example, the term *hip dislocation* in Figure 3.1. It is possible to express that this feature is both an *abnormality of the hip joint*

'Class' of phenotype	*HPO examples*
morphological abnormality	*Wide nose* (HP:0000445), *Arachnodactyly* (HP:0001166)
abnormal process (organ)	*Epistaxis* (HP:0000421), *Ileus* (HP:0002595)
abnormal process (cellular)	*Abnormality of amino acid metabolism* (HP:0004337), *Abnormality of Krebs cycle metabolism* (HP:0000816)
abnormal laboratory finding	*Hyperlipidemia* (HP:0003077), *Glycosuria* (HP:0003076)
electrophysiological abnormality	*Decreased nerve conduction velocity* (HP:0000762), *Hypsarrhythmia* (HP:0002521)
abnormality by medical imaging	*Butterfly vertebrae* (HP:0003316), *Choroid plexus cyst* (HP:0002190)
behavioural abnormality	*Nystagmus-induced head nodding* (HP:0001361), *Self-mutilation* (HP:0000742)

Table 3.1: Different types of phenotypic abnormalities covered by the HPO.

and a *joint dislocation*.

In medical contexts, the word *phenotype* is usually used to refer to some deviation from "normal", and this is the definition taken by the HPO. Many different types of phenotypic abnormality are represented in the HPO. This includes, besides morphological signs, also cellular, physiological, and behavioral abnormalities. A summary of different types of abnormalities is shown in Table 3.1. Similar to GO, the HPO provides three sub ontologies, whereas the major fraction of terms describe phenotypic abnormalities. There are separate ontologies that describe the *mode of inheritance* and the *onset and clinical course*.

The HPO was used annotate all clinical entries of OMIM. Subsequently, annotations to entries of the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) [Firth et al., 2009] database and Orphanet [Aymé, 2003] were enabled, due to semi-automatically generated mappings from the phenotype vocabularies used by these databases to the HPO. All of the mappings were manually confirmed or improved by a medical expert (Sandra Doelken, Peter Robinson). Using an iterative process more and more entries from those databases are now integrated into the HPO repository.

All annotations are made to the most specific terms possible, and the *annotation propagation rule* applies to these annotations. More details on this rule can be found in Section 1.3.2. From this rule follows that the

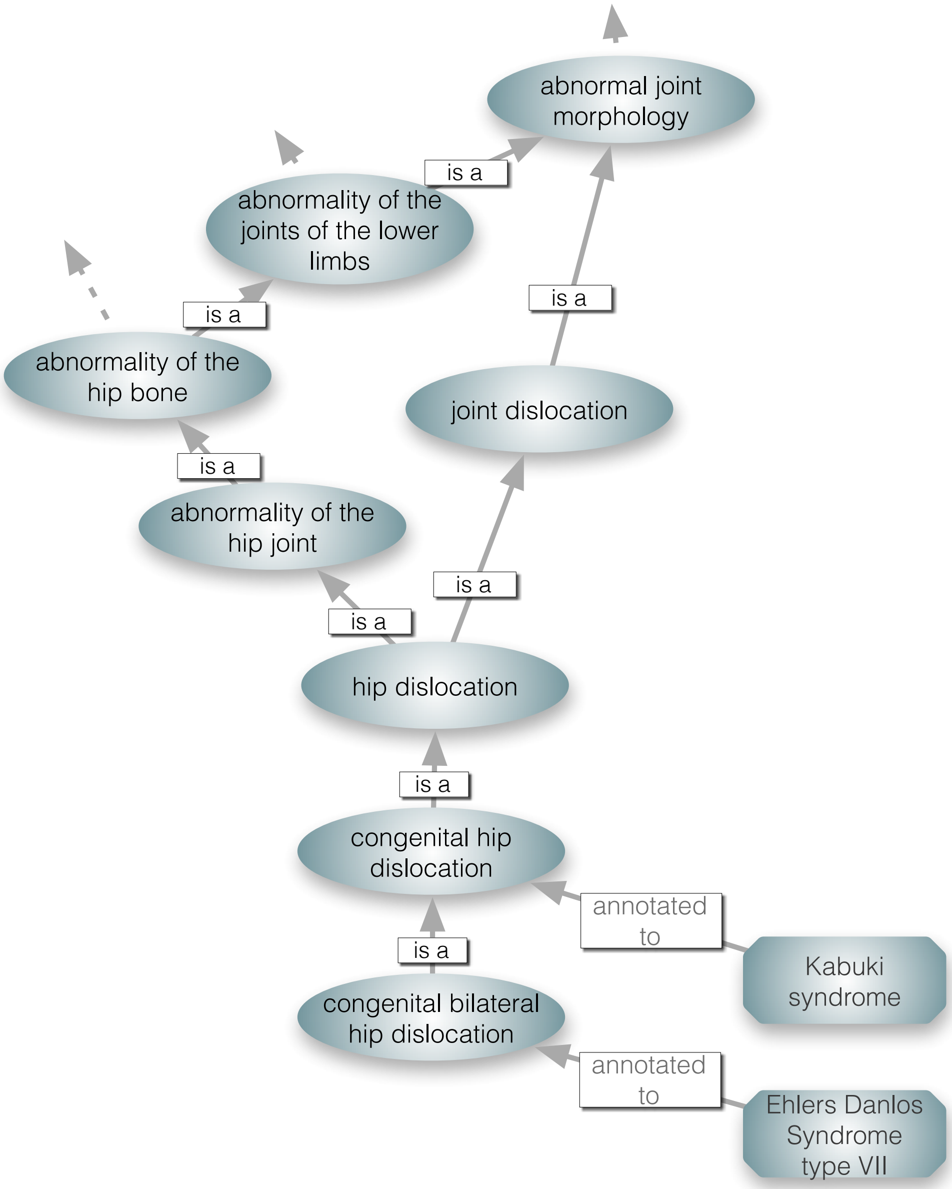

abnormal joint morphology
is a
abnormality of the joints of the lower limbs
is a
is a
abnormality of the hip bone
joint dislocation
is a
abnormality of the hip joint
is a
is a
hip dislocation
is a
congenital hip dislocation
annotated to
Kabuki syndrome
is a
congenital bilateral hip dislocation
annotated to
Ehlers Danlos Syndrome type VII

Figure 3.1: Excerpt of the Human Phenotype Ontology. A phenotypic feature of Kabuki syndrome is the *congenital hip dislocation*. Because of the "annotation propagation" the syndrome is implicitly annotated to all ancestor terms, such as *joint dislocation* and *abnormal joint morphology*.

annotated_to relation is propagated along the is_a to all ancestral terms. For example, since Kabuki syndrome is annotated to *congenital hip dislocation* it can be inferred that all of the ancestors of this term (e.g. *joint dislocation*) also apply.

HPO annotations also include a number of metadata items that allow the addition of further specifications. Each annotation is provided with a datasource, in most cases the name of the biocurator. Furthermore, the evidence code of an annotation indicates how the annotation to a particular term is supported. Initially the annotations were automatically generated from OMIM, so these were assigned the evidence code "IEA" (inferred from electronic annotation). Since 2008, many records have been revised and extended by expert biocuration. These annotations are given the code "PCS" (published clinical study), and the source of the study is indicated (usually, this is a PubMed ID). The evidence code "ICE" can be used for annotations based on individual clinical experience. This may be appropriate for disorders with a limited amount of published data when annotated by an experienced clinician. This must be accompanied by an entry in the DB:Reference field denoting the individual or center performing the annotation, together with an identifier. Additionally, the evidence code "ITM" (inferred by text-mining) can be used to mark annotations retrieved by text-mining efforts.

Another important characteristic of an annotation is the frequency with which individuals with a given disease have a certain phenotypic feature. For instance, 9 of 43 persons with the disorder *sialidosis type II* have as a phenotypic characteristic a *cherry red spot of the macula*. This can be important for differential diagnostic purposes. For instance, if 100 % of patients with some other disease *X* have *cherry red spot of the macula* then, all else being equal, a person with *cherry red spot of the macula* would be more likely to have disease *X* than *sialidosis type II*. In many cases, exact numerical information on a frequency is not available, and more or less vague terms such as "occasional" are used in medical textbooks and articles. Descriptions such as these have been used in extremely variable ways in the medical literature. The HPO team has defined a set of eight such categories to describe the frequency of features (Table 3.2).

Description	*% of patients*
very rare	1%
rare	5%
occasional	7.5%
frequent	33%
typical	50%
common	75 %
hallmark	90 %
obligate	100 %

Table 3.2: Eight frequency categories for features in HPO annotations. For instance, if we say that feature α is hallmark in disease X, then we mean that 90% of individuals with disease X have feature α. Numerical values are given for the categories to provide a rough guide.

The HPO (November 21st, 2011) contains 10,218 terms and 13,473 semantic relationships. There are 56,641 annotations to 5,035 OMIM entries. Mappings between HPO terms and the vocabulary of the LDDB and the Orphanet [Aymé, 2003] signs and symptoms vocabulary were created. This semantic structure of the HPO allows flexible searches for diseases according to phenotypic abnormalities. This has been implemented in the tool *PhenExplorer*[2]. Recently, Orphanet has made a large number of annotations for rare diseases available. These data are particularly valuable because most of the phenotypic features have been assigned to one of three frequency categories by expert annotation. This was used to explore the annotations in the setting of neurogenetic disease in the second part of the following chapter, where the Orphanet data has been used to supplement data from OMIM and biocuration by the HPO team.

A summary of symbols and notations can be found in Table 3.3.

[2] `http://compbio.charite.de/phenexplorer`

Symbol	*Definition*
D	A set of HPO terms that represent the phenotypic abnormalities associated with patients suffering from a specific disease.
Q	A query that contains a set of phenotypic abnormalities represented by HPO terms.
q	Size of the query Q, i.e. the number of HPO terms.
BMA_{asym}	Calculation of asymmetric semantic similarity between two sets of ontology terms as explained in Section 1.3.3 (Equation 1.4).
BMA_{sym}	Calculation of symmetric semantic similarity between two sets of ontology terms as explained in Section 1.3.3 (Equation 1.5).
FV	Feature Vector. Used to calculate a similarity between Q and D.
SD	Empirical score distribution of an object for a given q.
P_{est}	Estimated P-value. Calculated using an empirical score distribution SD. For a given semantic similarity score (e.g. BMA_{asym}) the statistical significance of this score can be determined.
n	The number of Monte Carlo sampling steps used to calculate the empirical score distribution used for P_{est}

Table 3.3: A summary of the symbols and notations used throughout chapter 3. Short definitions are given as well.

3.3 Classification of Syndromes

It can be shown that the HPO and the annotations of diseases are able to capture phenotypic similarity. This was done by calculating the semantic similarity between diseases that are annotated to HPO terms. In recent work by Goh et al. [2007], diseases were classified based on the physiological system affected. For the analysis here, 727 diseases belonging to one of 21 classes were included. A network was constructed in which every node represents a disease and every edge reflects the phenotypic relationships between these diseases. All pairs of diseases are linked if their semantic similarity score (BMA_{sym}, see Section 1.3.3, Equation 1.5) exceeds the threshold of 4.5.[3] This network is visualised in Figure 3.2. It can be seen that the diseases cluster according to the predefined disorder classes that were generated independently by Goh et al. [2007]. But there also exist links between clusters reflecting the phenotypic relatedness of disturbances in specific organ systems. For instance, *immunological* and *hematological* disorders are strongly connected to each other. Also, *bone* disorders are strongly connected to *skeletal* disorders, and *neurological*, *muscular*, and *psychiatric* disorders are multiply linked to one another. These interconnections were not visible in the disease map based only on shared disease genes in the work by Goh et al. [2007]. Furthermore, clusters also contain diseases from different classifications, but do in fact share important phenotypic features. This can be seen, for example, in the *muscular* cluster which contains four diseases classified as *metabolic* disorders. These four diseases, *Enolase-beta deficiency*, *MCardle disease*, *dimethylglycine dehydrogenase deficiency*, and *elevated serum creatine phosphokinase*, show important muscular symptoms.

Analysis of randomised versions of this network showed that the observed correlation between network connections and disease class is highly significant [Robinson et al., 2008]. Thus, this phenotypic network, as defined by the HPO, is made up of dense clusters of shared phenotypic features that show characteristic patterns of interconnections between selected areas of the phenotypic continuum.

[3] This threshold was chosen to decrease the number of edges in the graph visualisation.

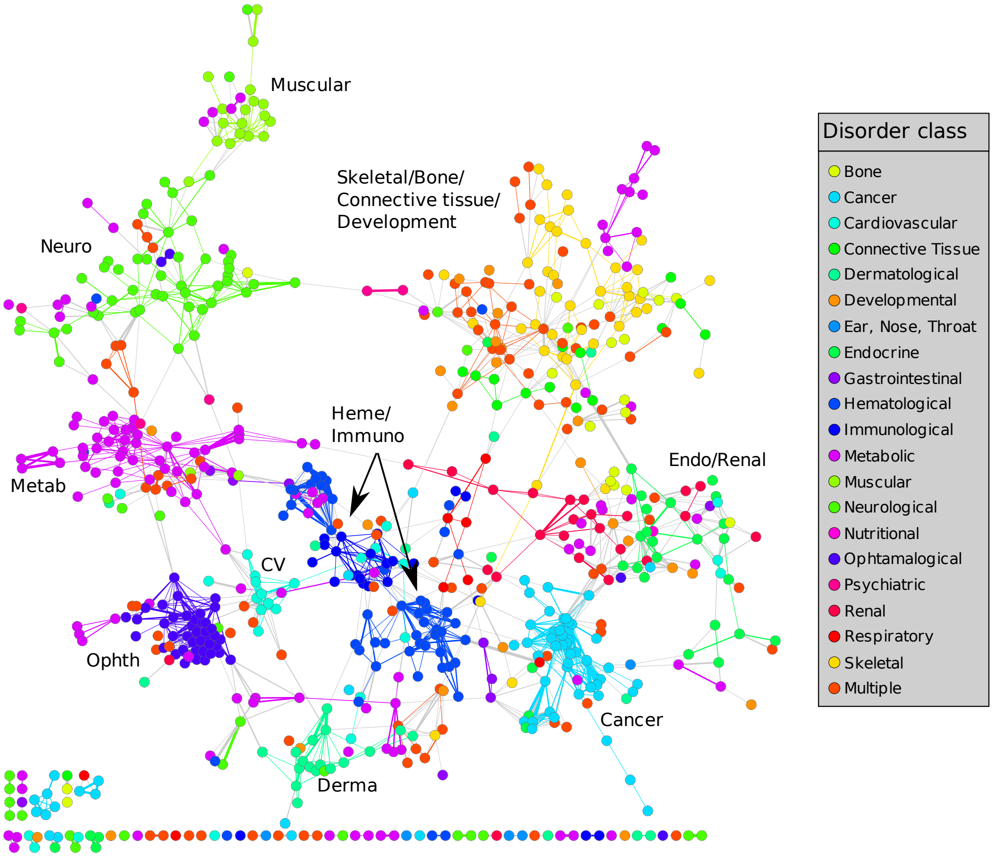


Figure 3.2: 727 diseases listed in OMIM and classified according to the affected physiological system by Goh et al. [2007]. The disease-nodes are colored according to their disorder class. The thickness of the links reflects the degree of phenotypic similarity. Abbreviations: CV, cardiovascular; derma, dermatological; endo, endocrinological; heme, hematological; immuno, immunological; metab, metabolic; neuro, neurological; ophth, ophthalmological.

In order to reconfirm the usability of the HPO for determining disease similarity, a similar analysis is performed on a different dataset. Orphanet provides, besides others, a classification of rare diseases elaborated using existing published expert classifications. The Orphanet classification of rare neurological diseases was downloaded[4] from the orphadata.org website. For every disease in this classification hierarchy, the most general disease classification(s) (i.e. level one diseases) was assigned and taken as the broad classification(s) of this disease. For example, the disease *X-linked distal spinal muscular atrophy* is classified as both *Neuromuscular disease* and *Rare peripheral neuropathy*. For each category, the number of diseases was determined, i.e. the number of descendant nodes in the Orphanet classification. The ten categories with the highest numbers of diseases were used for the analysis and can be seen in the legend of Figure 3.3. All diseases that belong to at least one but not more than four categories were included.

For each of the selected diseases, all annotations from Orphanet were transferred to HPO-terms. These annotations were completed with the HPO-annotations from the corresponding OMIM entries. To map the Orphanet diseases to one or multiple OMIM disorders, the Orphanet cross-reference file[5] was used. Diseases were excluded if they had less than three HPO annotations.

Again, the symmetric similarity measure as described before (BMA$_{sym}$, see Section 1.3.3, Equation 1.5) was used to define the phenotypic similarity between all pairs of diseases. If a pair of diseases had a semantic similarity of 3.0 or higher[6], an edge between the two diseases was added to the graph shown in Figure 3.3. The resulting network consists of 354 nodes and 1,316 edges. The nodes are colored according to their classification membership (using the MultiColoredNodes Plugin [Warsow et al., 2010] for Cytoscape [Smoot et al., 2011]). If an edge is drawn between two diseases that share the same category, the edge is colored using the color for that category. Edges connecting diseases that share multiple categories are colored yellow (see legend). Also, the edge thickness is chosen

[4] `http://www.orphadata.org/data/xml/en_product3_181.xml`, accessed June 2011

[5] `http://www.orphadata.org/data/xml/en_product1.xml`, accessed June 2011

[6] This threshold was chosen to decrease the number of edges in the graph visualisation.

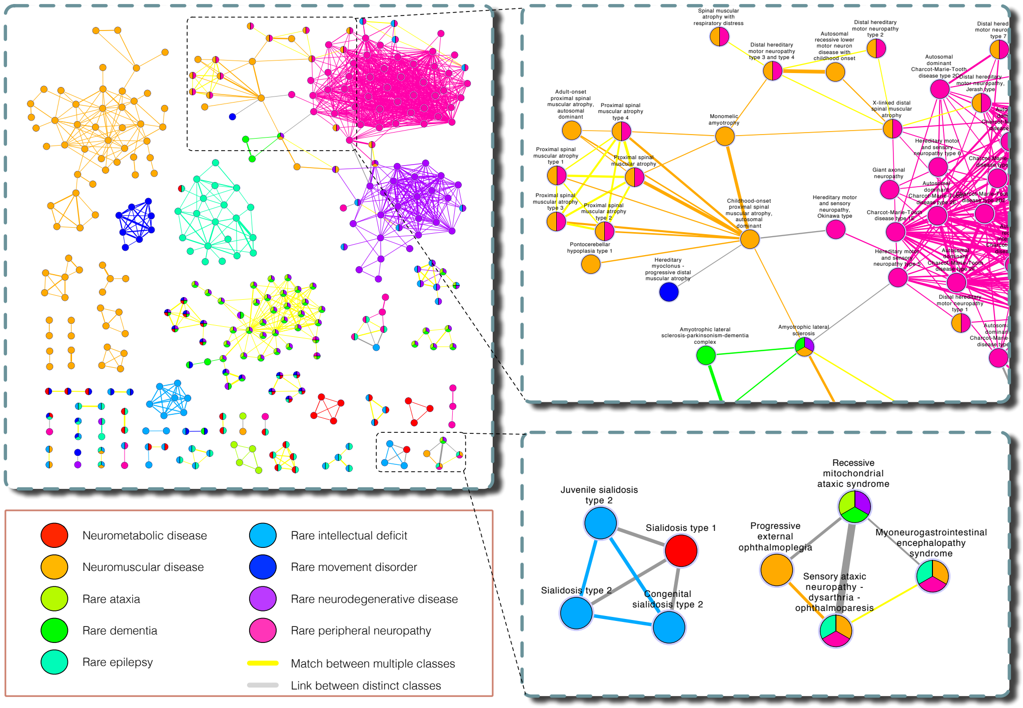

Spinal muscular atrophy with respiratory distress
Distal hereditary motor neuropathy type 3 and type 4
Autosomal recessive lower motor neuron disease with childhood onset
Distal hereditary motor neuropathy type 2
Autosomal dominant Charcot-Marie-Tooth disease type 2C
Distal hereditary motor neuropathy, Jerash type
Adult-onset proximal spinal muscular atrophy, autosomal dominant
Proximal spinal muscular atrophy type 4
Monomelic amyotrophy
X-linked distal spinal muscular atrophy
Proximal spinal muscular atrophy type 1
Proximal spinal muscular atrophy
Hereditary motor and sensory neuropathy type 6
Giant axonal neuropathy
Proximal spinal muscular atrophy type 3
Proximal spinal muscular atrophy type 2
Childhood-onset proximal spinal muscular atrophy, autosomal dominant
Hereditary motor and sensory neuropathy, Okinawa type
Autosomal dominant Charcot-Marie-Tooth disease type 2E
Charcot-Marie-Tooth disease type 2B2
Charcot-Marie-Tooth disease type 1E
Pontocerebellar hypoplasia type 1
Hereditary myoclonus - progressive distal muscular atrophy
Hereditary motor and sensory neuropathy type 5
Autosomal dominant Charcot-Marie-Tooth disease type 2F
Distal hereditary motor neuropathy type 1
Amyotrophic lateral sclerosis-parkinsonism-dementia complex
Amyotrophic lateral sclerosis
Juvenile sialidosis type 2
Sialidosis type 1
Sialidosis type 2
Congenital sialidosis type 2
Recessive mitochondrial ataxic syndrome
Progressive external ophthalmoplegia
Myoneurogastrointestinal encephalopathy syndrome
Sensory ataxic neuropathy - dysarthria - ophthalmoparesis
Neurometabolic disease
Neuromuscular disease
Rare ataxia
Rare dementia
Rare epilepsy
Rare intellectual deficit
Rare movement disorder
Rare neurodegenerative disease
Rare peripheral neuropathy
Match between multiple classes
Link between distinct classes

Figure 3.3: Network of Orphanets *rare neurological diseases* based on semantic similarity in the HPO. The "organic layout" feature of Cytoscape was used to create the graph. Afterwards some nodes were moved manually to improve readability.

to correlate with the degree of semantic similarity between the connected diseases.

The clustering shown in Figure 3.3 is a visualisation of the phenotypic relationships that exist between the ten largest classes of rare neurological diseases. Although the clustering procedure knew nothing of the classification assignment of the diseases to the ten classes, the resulting network clearly shows that the diseases cluster, in great measure, according to the assignment by the Orphanet disease classification. This is an indirect confirmation of the correctness of the disease classification, in the sense that the classification of the diseases reflects the spectrum of shared and distinct phenotypic abnormalities that characterise the diseases. In addition to this, there appear to be some interesting interconnections between some of the clusters. For instance, a cluster of *rare peripheral neuropathy diseases* shows a number of links to the diseases that are classified as *neuromuscular diseases*. This reflects the well-known similarity of muscular and neurologic phenotypes in this special field of neurogenetic diseases, and confirms the connections shown in Figure 3.2. A phenotype such as *muscle weakness* may be either caused by a primarily muscular problem or a problem of the muscular innervation, i.e. caused by a neural deficiency. The links between those two clusters reflect the situation in clinical practice and show that there is much more phenotypic similarity between peripheral neuropathies and neuromuscular diseases than between these two and, for example, ataxias. In some cases, phenotypic similarity is present even between diseases that belong to different Orphanet classifications. In Figure 3.3, this is indicated by gray edges. For instance, *sialidosis type 1* is classified as a *neurometabolic disease*, whereas the other sialidosis types are classified as *rare intellectual deficit*. Another example is *Niemann-Pick disease type A*, which is classified as a *neurometabolic disease*, *rare epilepsy*, and *rare movement disorder*. According to the results of the analyses presented here, this disease shows a high phenotypic similarity to *Niemann-Pick disease type B* which is classified as *rare peripheral neuropathy* and thus shares no classification in Orphanet with the *Niemann-Pick disease type A*.

Thus, HPO-based phenotypic analysis can be used to point out areas in the Orphanet classification that require curator attention. Therefore, such approaches can be used to systematically improve the quality of the classification. In the future, these methods might also help to place novel neurogenetic diseases correctly. Given a set of clinical features inherent to the novel disease, an algorithm could then place this syndrome according to its phenotypic similarity to already existing disease ontologies and nosologies.

3.4 Computational Phenotype Guided Clinical Diagnostics

As has been shown above, the semantic structure of the HPO can well capture phenotypic similarities between known syndromes. Clinical geneticists are often in a different situation, in which patients with different phenotypic features present and ask for a diagnosis. Note that diagnosis can have different meanings in medicine. It may denote the process of the identification of the exact causal nature of a patient's syndrome. But it may also refer to the task of assigning the patients conditions to a classification, which can then give insight into possible ways of treatment or information on prognoses [Pelz et al., 1996]. In this work, the second definition of diagnosis applies.

The task of establishing the correct diagnosis is arguably the most important role of the physician. Especially in medical genetics, this task can be very challenging, due to the huge number of Mendelian and chromosomal disorders. Each of these disorders often has numerous phenotypic features, and these features are often shared among different diseases. Also, variable expression and pleiotropy may affect the patient's phenotypic occurrence. Thus, patients may have different, partially overlapping combinations of clinical signs and symptoms, which complicates the process of identifying the correct differential diagnosis. However, a prompt and correct genetic diagnosis is essential, because thereby it is possible to avoid unnecessary diagnostic procedures, identifying appropriate therapeutic measures and clinical management strategies, and providing adequate genetic counseling. However, an etiological diagnosis can be made in only about half or fewer of the children presenting with dysmorphic signs with or without mental retardation [Köhler et al., 2009].

In order to aid clinicians in medical genetics with these tasks, several tools and databases were developed. For example, the commercial programs "Pictures of Standard Syndromes and Undiagnosed Malformations" (POSSUM) [Bankier and Keith, 1989] and London Dysmorphology Database (LDDB) [Fryns and de Ravel, 2002] were created. Alternatively, the freely accessible data repositories from the OMIM and Orphanet websites can, to some extent, be used for this purpose. In these tools and platforms, the typical workflow starts with a user entering one or more clinical phenotypes that the patient(s) presents with. From this, a list of candidate diagnoses is created, each of which is characterised by some or all of the entered phenotypes. However, these tools do not provide explicit rankings or define measures of plausibility for the potentially long lists of results [Pelz et al., 1996]. Also, none of the methods is able to exploit the semantic relationships that exist between clinical phenotypes in order to weight the candidate syndromes.

Here, it was evaluated how well ontological search routines based on the structure of the HPO can improve this process. As introduced in Section 1.3.3, semantic similarity measures can be used to define similarity scores that weight clinical features on the basis of their specificity. The strategy is that a user enters HPO terms describing the clinical abnormalities observed in a patient (query) and a ranked list of the best matching differential diagnoses (diseases) is calculated. This ranking can be calculated using raw semantic similarity, but this measure has inherent problems.

3.4.1 Significance of Semantic Similarity Searches

There are two major problems with the commonly used semantic similarity scores, e.g. the one introduced in Section 1.3.3. First of all, the raw similarity score depends on a number of factors, including the number and specificity of the terms both of the query and of the diseases represented in the database. It is thus not possible to say what score constitutes a “good match” for a general query. Especially when ranking all objects of a database by semantic similarity, one is not able to determine if the object that achieved the highest similarity represents a meaningful result.

Second of all, semantic similarity scores are highly correlated with annotation length [Wang et al., 2010]. That means the more annotations an object has, the higher will be its semantic similarity score for a given query (protein or syndrome annotations). Together with the fact that well studied diseases and genes are having more annotations, this may introduce an undesirable bias for a tool aimed at supporting clinical geneticists.

In the case of the application studied here, the search is based on a query Q consisting of q HPO-terms. Q consists of the clinical phenotypes seen in a patient. For this patient, the most likely differential diagnosis is sought. To find this diagnosis, every disease D in the database used to calculate the semantic similarity $s = \mathrm{BMA}_{asym}(Q, D)$, whereby D has annotations to HPO-terms that correspond to the signs and symptoms characterising the disease.

When applying Equation 1.4, for each of the q terms the best match to one of the terms from D is found and the average over theses scores is determined. All disease are ranked according to this score, with the top ranked diseases representing the suggested differential diagnosis. As can be seen in this calculation, the score that a disease can achieve varies with the number of annotations of D and the specificity (measured by IC) of the annotated HPO-terms [Schulz et al., 2011]. As noted by Wang et al. [2010], this and similar measures are biased towards domain objects that have more annotations. Thus, the rankings of diseases that are based on the scores alone tend to preferentially select items with higher numbers of annotations, which may lead to wrong conclusions.

Therefore, a statistical model was developed that assigns a P-value to each of the raw semantic similarity scores. This P-value reflects the probability of a randomly chosen query Q_r of same size q obtaining the same or a higher semantic similarity to D than Q. This P-value is calculated by n times generating a random query Q_r consisting of q HPO-terms and calculating the similarity $s_r = \mathrm{BMA}_{asym}(Q_r, D)$.

At first, an empirical score distribution SD of size n for a disease D and a query size q is generated. This is done by performing a Monte Carlo sampling to approximate the score distribution. For this, a set R is generated that consists of n randomly generated queries of size q. The score distribution is then

$$SD(q, D) = \{s_r | s_r = \mathrm{BMA}_{asym}(Q_r, D), Q_r \in R\} \tag{3.1}$$

From this empirical score distribution, one can easily determine an estimated P-value by calculating the fraction of cases where a random query Q_r leads to score s_r that is higher than (or equal to) the score s from the actual query Q:

$$P_{est}(s,q,D) = \frac{|\{s_r | s_r \geq s, s_r \in SD(q,D)\}|}{n} \ . \qquad (3.2)$$

This then compensates for the fact that different domain objects have a different number of annotations and subsequently leads to the hypothesis that rankings based on this P-value are better suited in the setting of medical diagnostics.

For comparison, a simple feature vector (FV) method was implemented as well, in which the exact overlap between Q and D is calculated. This method is meant to be similar to text-matching methods used by POSSUM and the London Dysmorphology Database (LDDB) [Fryns and de Ravel, 2002], as well as the search routines available with the OMIM website and Orphanet. Note that there was no attempt to perform an explicit comparison with these databases because of the different clinical vocabularies used by each of these databases and the fact that they do not provide a ranking for the results of searches. The FV-score is thus the size of intersection between the sets of terms Q and D, i.e.

$$FV(Q,D) = |\{Q \cap D\}| \ . \qquad (3.3)$$

Note that the FV method does not take the semantic inheritance structure of the HPO into account.

3.4.2 Clinical Diagnostics with Semantic Similarity Searches

It was tested how well semantic similarity searches in the HPO are suited in the setting of clinical diagnostics, when a patient's phenotypic features are used to rank all entries of a disease database. Three methods are evaluated:

- Feature Vector (FV)
- Semantic Similarity Score (BMA$_{asym}$)
- P-value of the Semantic Similarity Score (P_{est})

Strategy

It is complicated to validate such a diagnostic procedure using data from real patients. One of the major reasons is that it is difficult to obtain phenotypic information on hundreds or thousands of patients. This would be required in order to calculate statistical measures during validation. This data would have to be collected by normalised and standardised procedures and stored using controlled vocabularies for comparability. Instead, an automated procedure was used for testing, in which “simulated patients” are generated on the basis of clinical features among persons diagnosed with a certain disease. Although it is not correct, an independence between the occurrence of individual clinical features was assumed, due to the fact that sufficient data are currently not available for modeling of the interdependencies of clinical features.

Query Modification

The phenotypic features of the artificial patients were modified in order to test the performance of the methods in the presence of two typical inaccuracies that have to be expected in day-to-day situations.

The first difficulty with clinical databases is that physicians may not choose the same phrase to describe some clinical anomaly as that which is used as a feature for a syndrome in the database. This can either happen because the physician is unaware of the correct terminology or because detailed laboratory or clinical investigations have yet to be performed and a clinical anomaly can only be described on a general level. This type of inaccuracy is referred to as *imprecision*. For the simulation, if *imprecision* was applied to the query Q, every term of Q was exchanged with probability 0.5 with one of its parent terms, if possible. That means that no term was exchanged if the parents were only the root or the direct subclass of the root.

A second typical scenario in clinical practice is that patients not only have phenotypic signs and symptoms that are related to some underlying disorder but may also have several unrelated clinical problems. This will be referred to as *noise*. The simulation of noise was performed by exchanging a specific fraction of terms of Q with randomly chosen HPO terms. For $3 \leq q \leq 5$, all but 2 terms were replaced by random terms. For q in the range of 6 to 8, only 3 terms were not replaced by randomly chosen HPO features. For $q \geq 9$, only 4 terms were kept. If q was less than 3, no *noise* was introduced.

When both *noise* and *imprecision* are applied, the first step was to introduce *imprecision* and afterwards the *noise*-step was applied. This is important for cases where the first modification may lead to a reduced number of Q, which may happen if two query terms are mapped to the same parent term.

Methods Tested

Three methods (FV, BMA$_{asym}$,P_{est}) were tested for their ability in ranking the correct differential diagnosis.

Here the focus is the ranking of diseases with only a few annotations. Thus the methods were tested with 2,727 diseases with at least two but no more than ten annotations to an HPO-term from the *phenotypic abnormality* (HP:0000118) subontology of the HPO were extracted. Note that the performance advantages shown here are also valid when all diseases irrespective of their annotation size are taken into account [Bauer, 2011, Bauer et al., 2012]. For each of the diseases, an artificial patient was generated and used as query. For testing, the query was also modified by introducing *imprecision*, *noise*, or both. The query was then used to rank all 5,035 HPO-annotated diseases using one of the measures (FV, BMA$_{asym}$, P_{est} with $n = 10^3$). Afterwards, the rank of the disease from which the query was initially generated was calculated. In case of ties, the average rank was taken. For example, if four diseases rank first with the same value, all four diseases obtain rank 2.5. Note that for the rankings based on P_{est} all diseases were ranked first by P_{est} and then by the score BMA$_{asym}$.

As can be seen in Figure 3.4, the inclusion of the semantics of phenotypes defined in the HPO can improve the performance of a differential diagnosis search tool. It can be seen that the FV method has a slight advantage in the ideal situation when the query is not modified, or in the case when only *noise* is introduced. The effect of *imprecision* simulates the situation when a physician enters a HPO term to describe a phenotypic feature that is more general than the term used to describe the disease in the database. It can be seen that the performance of the FV method greatly suffers in this situation. On the contrary, the ontological methods intuitively use the semantic structure of phenotypes encoded in the ontology to recognise that the imprecise term has a meaning similar or related to that of the term used in the database. Thus, the methods BMA_{asym} and P_{est} show only a minimal decrease in the ranking performance.

The P_{est}-based ranking method, which bases the ranking on the P-value of attaining a given score for each disease in the database, is superior to the results of ranking on the basis of the raw similarity scores BMA_{asym}. This reflects the fact that the distribution of similarity scores is not the same for all diseases in the database, and suggests that search methods that in fact do take the local score distributions into account are superior [Schulz et al., 2009, 2011, Köhler et al., 2009]. In sum, it was shown that ontological approaches are especially robust in the presence of inaccuracies in the terms used to query a database.

For the results shown in Figure 3.4, the P_{est} rankings were based on quite a small number of randomisation steps, i.e. $n = 10^3$. In order to estimate the influence of the parameter n on the ranking performance, the results were compared to rankings based on a higher number of sampling steps, i.e. $n = 10^4$ and $n = 10^5$. These results are shown in Figure 3.5, where it can be seen that the performance improves with higher n. In the work by Schulz et al. [2011], a novel algorithm was developed, which for small q can compute the exact calculation of the P-values. This algorithm does not use the extensive simulation procedure and largely improves the ranking performance in the setting of HPO-based diagnosis identification. Unfortunately, this approach again becomes computationally too expensive for q larger than six.

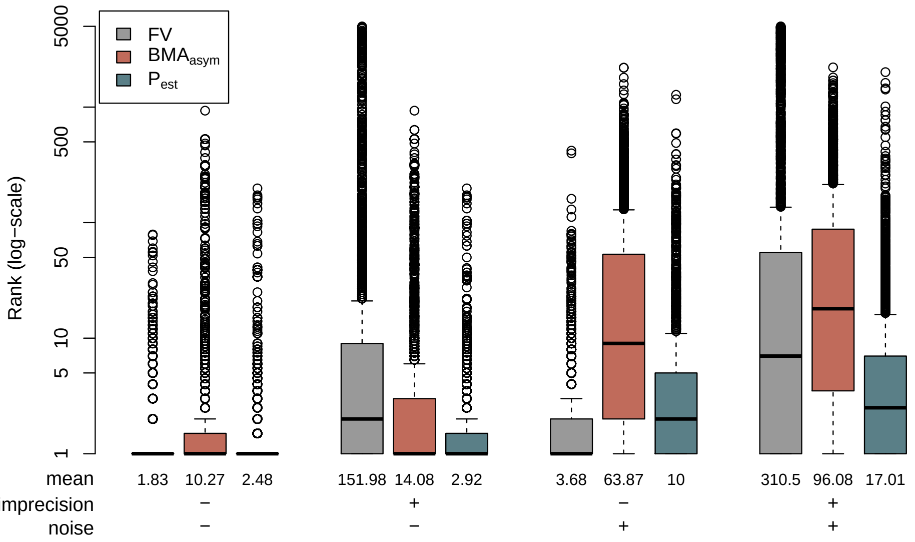


Figure 3.4: Rankings of correct differential diagnosis of the simulated patients by the feature vector method (FV), semantic similarity score (BMA$_{asym}$), and the P-value of the semantic similarity score (P_{est}). Lower ranks indicate superior performance, whereby the optimal rank is 1. A boxplot of the ranks of the correct diagnosis for each of the simulated patients is shown for the three methods tested. Also, the results for the different combinations of adding "noise" and/or "imprecision" are shown. Each boxplot shows 50 % of the data points surrounding the median in the box, where the position displays the skewness of the data. The whiskers extend to the most extreme data point that is no more than 1.5 times the length of the box away from the box. More extreme outliers are displayed as circles. Below the boxplots the mean of the ranks is shown. It can be seen that the rankings based on P_{est} method are especially robust when parent terms of syndrome annotations are used in the query, which is essential in the setting of clinical diagnostics. The rankings in presence of "noise" can be improved for the sampled P_{est} (here $n = 10^3$) method, which is shown in Figure 3.5 and the publication by Schulz et al. [2011].

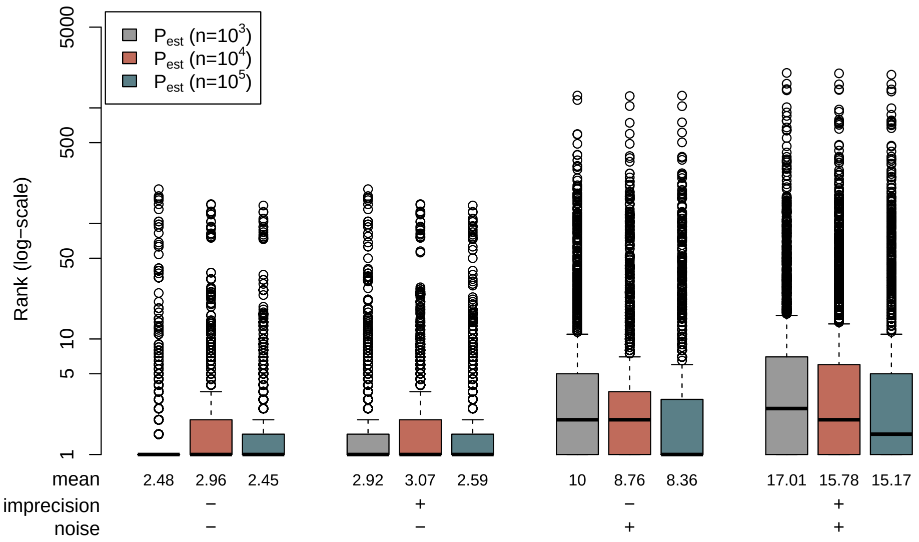


Figure 3.5: Ranking performance is improved when more sampling steps are performed when creating the empirical score distribution. In the paper by Schulz et al. [2011], it is shown that calculating the exact score distribution improves the results even more, but this computation is still infeasible for queries consisting of more than 6 terms.

3.4.3 The Phenomizer

The algorithms described before are implemented in a web application called the *Phenomizer* [Köhler et al., 2009], which is a GWT[7]-based rich internet application, meaning that it has much of the functionality of standard desktop applications, but requires no installation. To increase functionality the GWT-EXT[8] library was used. One of the major aims was to provide an easily accessible tool for physicians all over the world.

The *Phenomizer* was thus developed to be platform independent and users of this tool are only required to have an internet connection and a browser (i.e., Safari, Internet Explorer, Chrome, or Firefox). To the best of the author's knowledge, this is the only freely available tool for clinical diagnostics in human genetics with semantic similarity searches in ontologies. Three years after its initial publication, the *Phenomizer* is still accessed approximately 900 times per month by users all over the world.[9]

The Phenomizer in Clinical Practice

The usability of some of the features implemented in the *Phenomizer*[10] in clinical practice is delineated here in short. For this, a short example workflow is outlined.

It is assumed that a boy has developmental retardation as a major indication. The physician in charge finds that the boy has an *arachnodactyly* and furthermore notices an *abnormality of the sternum*. When using the corresponding HPO terms as a query, the similarity to each of the diseases in the database is calculated. Therefore, the P-value is adjusted by multiple-testing correction. The default multiple-testing correction method for the *Phenomizer* is that of Benjamini and Hochberg [Benjamini and Hochberg, 1995], but users can choose among several other multiple-testing corrections. The corrected P-values are calculated using R [R Development Core

[7] Google Web Toolkit, `http://code.google.com/webtoolkit/`

[8] `http://www.gwt-ext.com`

[9] Estimate generated evaluating one month (26th June 2012 – 26th July 2012) using the tool available at `http://www.revolvermaps.com/?target=enlarge&i=4qfhq0flvot&color=ff0000&m=3`.

[10] `http://compbio.charite.de/phenomizer`

Team, 2011] on the server side.

For the query related to the young boy above, the *Phenomizer* will return several OMIM entries, all of them having no significant *P*-value. This lack of significance reflects the fact that the clinical findings are not specific enough, per se, to allow a diagnosis. For these situations, the *Phenomizer* can be used to generate a list of clinical features that are most specific for individual syndromes in a set of previously selected candidate diagnoses. In the example used here, 13 top-scoring diseases were selected and the *Improve Differential Diagnosis* functionality was used. In the list of specific features the *arterial tortuosity, generalized* is shown, which could prompt further investigations such as magnetic resonance imaging of the vasculature. The addition of this HPO term to the list of the patient's features leads to a significant *P*-value for Loeys-Dietz syndrome (Type 1A and 1B) and Arterial tortuosity syndrome.

The clinical features listed by the *Phenomizer* can suggest further exact clinical examination (e.g., *fine, brittle hair*) or technical examinations (e.g., radiography to search for *codfish vertebrae*). In many cases, adding a feature from the list of specific terms results in a disease ranking, in which only one or a few syndromes have a significant *P*-value. Thus, the *Phenomizer* can function as a tool for planning of further clinical workup by referring patients to an appropriate specialised physician, or to initiate the appropriate genetic mutations analysis.

All models are wrong, but some models are useful.

GEORGE E. P. BOX [1979]

Chapter 4

Semantic Web Techniques for Genotype to Phenotype Discovery

4.1 Logical Definitions of Phenotypes

With the enormous increase in biomedical data and publications, the use of controlled vocabularies and ontologies for representing biomedical entities gained in importance, with the Gene Ontology (GO) being probably the most successful representative in the field of bio-ontologies. There exist several ontologies, each of which has evolved from a specific discipline in biomedicine. A major problem is the lack of interoperability between ontologies of different domains of biomedical knowledge. The elimination of this gap is one of the major aims of the OBO Foundry (see Section 1.3.2).

In principle, a good way to develop ontologies is to define complex classes in terms of other more elementary (atomic) classes (building blocks). When doing so, several ontologies would use shared building block ontologies further increasing interoperability across a larger domain.

Unfortunately, due to historical reasons, the creation of ontologies such as the GO, the Mammalian Phenotype Ontology [Smith et al., 2007] (MPO), or the HPO, predated typical building block ontologies. Examples for building block ontologies that can be used for representation of classes of phenotypic abnormalities are given in Table 4.1. OBO Foundry ontologies, such as the GO [Mungall et al., 2011], the MPO [Smith et al., 2005],

Domain	Name (Abbreviation)	Reference
biochemistry	Chemical Entities of Biological Interest (ChEBI)	De Matos et al. [2010]
proteins	Protein Ontology (PR)	Natale et al. [2011]
cell types	Cell Ontology (CL)	Bard et al. [2005]
anatomy (human)	Foundational Model of Anatomy (FMA)	Rosse and Mejino [2003]
anatomy (mouse)	Mouse adult gross anatomy (MA)	Finger et al. [2011]
anatomy (zebrafish)	Zebrafish anatomy and development (ZFA)	Sprague et al. [2008]
phenotype	Phenotype, Attribute and Trait Ontology (PATO)	Gkoutos et al. [2004]

Table 4.1: Examples of typical building block ontologies for the biomedical domain. Here the focus lies on ontologies that can be used to represent complex classes of phenotypic abnormalities.

the HPO [Robinson et al., 2008, Gkoutos et al., 2009], the Worm Phenotype Ontology [Schindelman et al., 2011], and also the CL [Meehan et al., 2011], are now developing *logical definitions* for ontology terms using terms from other building block ontologies, with PATO, an ontology of phenotypic qualities, being a key tool in this effort [Gkoutos et al., 2004, 2009].

To reach the goal of interoperability in biomedicine, the approach taken here is termed EQ, which was developed for the biomedical domain. In the EQ-approach, the main idea is that phenotypic descriptions can be abstracted into two parts. First, an entity that is affected, i.e. the thing for which observations are made. This can be entities of various types, e.g., a protein, a cellular compartment, or an anatomical structure. Second, the quality of that entity, which is described in a qualitative or quantitative way [Gkoutos et al., 2004]. In the typical setting, a phenotype is described using a class expression consisting of a PATO quality class differentiated by a bearer entity class (from an OBO ontology) using the `inheres_in` relation (from OBO Relation Ontology) [Hancock et al., 2009].

This approach has several advantages. Firstly, consider, for example, all the terms that deal in some way with the chemical entity *glucose*. E.g.

- *glucose metabolic process* from the GO
- *abnormal glucose tolerance* from the MPO
- *glucose intolerance* from the HPO

- *glucose 1,6-bisphosphate synthase* from the PRO

A major target is now, that all those terms make use of the term *glucose* (CHEBI:17234) from the Chemical Entities of Biological Interest ontology (ChEBI) by referring to it as the entity. It is easy to see, that this would make it trivial to query across different databases and domains for biomedical entities that are related to *glucose*.

Besides increasing interoperability, this would also imply that only a minimum number of definitional relations would have to be asserted, because relations would then be defined in the building block ontology (here ChEBI). These building block assertions can subsequently form the basis for inferring logical consequences in ontologies referring to it [Köhler et al., 2011]. In the glucose example, super- and subclass relations of *glucose* are defined in ChEBI, e.g. *glucose* is defined to be a subclass of *aldohexose* (see Figure 4.1).

To give an example for logical definitions, consider the HPO term *Hypoglycemia* and its E/Q definition, specified in OBO Format:

```
[Term]
id: HP:0001943 ! Hypoglycemia
intersection_of: PATO:0001163 ! decreased concentration
intersection_of: qualifier PATO:0000460 ! abnormal
intersection_of: towards CHEBI:17234 ! glucose
intersection_of: inheres_in FMA:9670 ! Portion of blood
```

The word *Hypoglycemia* refers to an abnormally decreased concentration of glucose in the blood. The logical definition uses relations and follows the pattern described in previous work on defining phenotypes by Mungall et al. [2011]. The logical semantics are made explicit when translating the definitions to the Ontology Web Language (OWL) [Motik et al., 2008]. The translation to OWL is the subject of current research [Mungall et al., 2010, Hoehndorf et al., 2011a, Loebe et al., 2012] and a more detailed discussion is out of scope of this thesis. The translation is done by using the oboformat library[1] and the chosen model leads to the desired inferences. The translation is shown in Manchester syntax. Note that for the

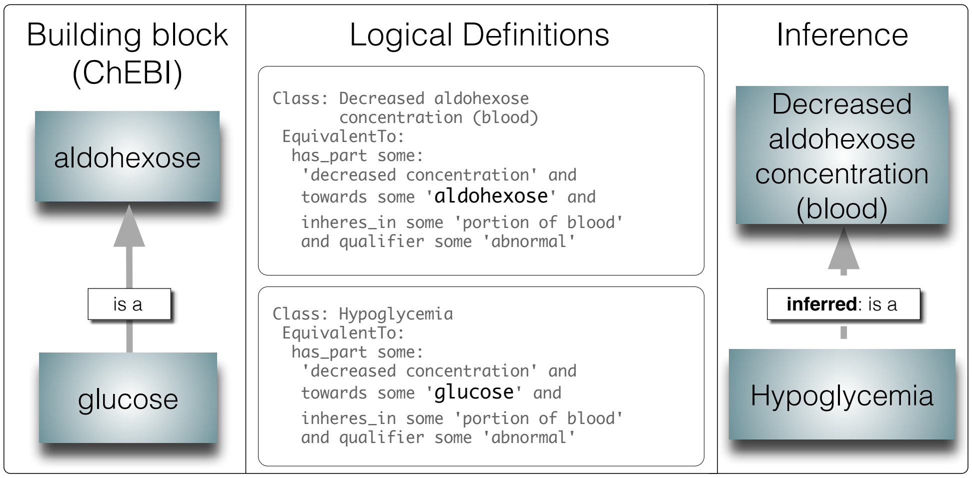
[1] http://code.google.com/p/oboformat

purpose of increased readability, only the term's labels are shown and the ontology URIs are skipped.

```
Class: Hypoglycemia
 EquivalentTo:
  has_part some:
   'decreased concentration' and
   towards some 'glucose' and
   inheres_in some 'portion of blood'
   and qualifier some 'abnormal'
```

The class *Hypoglycemia* is defined as being equivalent to the intersection of all classes of things that are "A concentration which is lower relative to the normal" (*decreased concentration*), "deviate from the normal or average" (*abnormal*), with respect to (towards) *glucose* , and inhering in "blood" (using the term *portion of blood* from the FMA). The formal `inheres_in` relation expresses the relationship between qualities and their bearers. In this case the bearer of the quality is the blood. The relation `towards` is used to connect the quality (here, *decreased concentration*) to the additional entity type on which the quality depends (here *glucose*). By applying this to the term for *glucose* in the ChEBI ontology, it is essentially stated that the concentration is a concentration "of" glucose. Thus, the term *Hypoglycemia* is defined as the intersection of these four classes. The details on the exact format specifications can be found in Mungall et al. [2010].

Given that logical definitions exist for the major fractions of classes of an ontology, one can apply automatic reasoners. As stated in Section 1.3, these are systems for computing the logical consequences that can be inferred from a set of asserted axioms. As mentioned above, *glucose* is defined to be a subclass of *aldohexose* in ChEBI. Assume that there exists another HPO-term (besides *Hypoglycemia*) called *Decreased aldohexose concentration (blood)*, and this term is defined almost equivalent as *Hypoglycemia* (only replacing the ChEBI reference to `'glucose'` by `'aldohexose'`). The reasoner is then able to automatically infer that *Hypoglycemia* is a subclass of *Decreased aldohexose concentration (blood)*. These connections are illustrated in Figure 4.1.



```
Class: Decreased aldohexose
       concentration (blood)
 EquivalentTo:
  has_part some:
   'decreased concentration' and
   towards some 'aldohexose' and
   inheres_in some 'portion of blood'
   and qualifier some 'abnormal'
```

```
Class: Hypoglycemia
 EquivalentTo:
  has_part some:
   'decreased concentration' and
   towards some 'glucose' and
   inheres_in some 'portion of blood'
   and qualifier some 'abnormal'
```

Figure 4.1: Illustration of the approach to build phenotype ontologies based on building block ontologies. Here two classes, *Hypoglycemia* and *Decreased aldohexose concentration (blood)*, are defined equivalently, except for the molecule that is affected (*glucose* vs. *aldohexose*). Given the two logical definitions, a reasoner can be used to infer that the two phenotype classes are in a subclass relationship based on the asserted link in the building block ontology.

This means that reasoners are able to use computable, logical definitions to infer the positions of classes in a subsumption hierarchy. Thus, those definitions can be helpful tools for the development of ontologies and their maintenance [Mungall et al., 2011], for example, by evaluating the overlap and disagreement between manually asserted and automatically inferred subclass relationships [Köhler et al., 2011]. This has been implemented in a tool called *GULO* (Getting an Understanding of LOgical definitions), which checks if, e.g., the link between *Hypoglycemia* and *Decreased aldohexose concentration (blood)* would exist in both ontologies, the manually curated HPO, and the automatically reasoned ontology, which is based on the logical definitions.

4.2 Inference of a Cross Species Phenotype Ontology

4.2.1 Model organism data

The use of non-human models for the understanding of human disease has proved to be one of the most powerful approaches to understanding the pathobiology of human disease [Rosenthal and Brown, 2007, Lieschke and Currie, 2007, Schofield et al., 2010]. They are an important source of data on normal and patho-biological phenomena, because many experiments cannot be performed in humans for various reasons. Next generation sequencing has enabled the generation of the genome sequences of the most important model organisms (and their strains), such as mouse, zebrafish, and fruitfly. The amount of phenotype information on model organism is now increasing, with the literature currently being the most prominent source of phenotype annotations. For example, curators from the Mouse Genome Database annotate mouse genes with terms from the MPO based on evidence in published articles. Additionally, major international efforts were put forward recently to systematically analyse the effect of genomic variation on the model organism's phenotype. The International Knockout Mouse Consortium (IKMC) has been put forward to knockout every gene in the mouse genome and, together with, the pan-genomic phenotyping efforts of the International Mouse Phenotyping Consortium (IMPC), this allows the mapping of phenotypes together with direct validation of candidate genes through the knockout phenotype [Schofield et al., 2012]. Similar approaches are taken in other model organisms, such as zebrafish (Danio rerio) [Bradford et al., 2011].

A typical approach to predict the function of a human gene or its involvement in a pathological process using model organism data is to transfer the knowledge from model organism genes to its ortholog counterpart in human. The orthology relationships between model organism genes and human genes have extensively been studied and although this is still an active area of research it is assumed here that orthology relations to human genes from the considered model organisms are given [Fang et al., 2010]. Also, it is assumed that orthologous genes are involved in orthologous pathways and that this scaffold can then be used to transfer information on a particular gene in a model organism to its ortholog human counterpart [Schofield and Hancock, 2012]. The increasing influence of systematic phenotype-guided studies (IMPC, ZFIN) has led to a growing interest in phenotypic data, but has also revealed several challenges regarding the determination of a particular phenotype as well as the computational representation of phenotypic data. Besides the need of standardised phenotyping procedures, it is crucial to develop appropriate ontologies to describe phenotypes, so that phenotypic descriptions can be related to each other in a systematic and coherent way, even between species [Rosenthal and Brown, 2007].

A major problem is the lack of common semantics across databases, because previously mostly free-text descriptions or different vocabularies were used. This impedes sophisticated interoperable datasets, which consequently implies that the orthology scaffold can't be used in its full potential, since phenotype data is not easily transferrable from model organisms to human. As can be seen in Figure 4.2, there is a huge amount of data in model organisms that scientists wish to harvest, and the discrepancy between small number of human genes with phenotype information and the number of systematically phenotyped model organism genes is expected to increase even more in the near future.

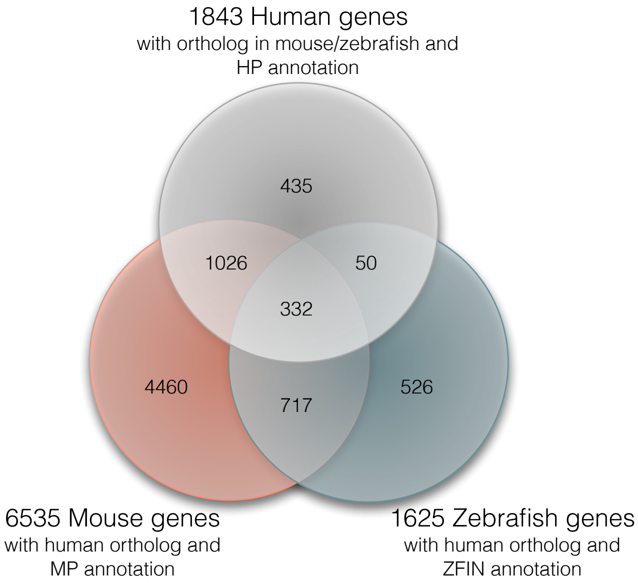


Figure 4.2: A Venn diagram illustrating the amount of human, mouse, and zebrafish orthologous genes with phenotypic information. One can see that there are 5,703 (4,460 + 717 + 526) genes with phenotype data in mouse or zebrafish for which the human ortholog gene has no phenotype data.

Cross-species ontology-based approaches offer a promising new methodology to reliably detect phenotypic similarities between human disease manifestations and model organism phenotypes [Washington et al., 2009, Hoehndorf et al., 2011b, Chen et al., 2012]. They can pave the way to gain clinically relevant insights from the over 5,000 genes for which, currently, only mouse and zebrafish phenotypic information is available (see Figure 4.2).

4.2.2 *Uberpheno* construction

In this thesis, the above mentioned EQ-approach is used to generate a single cross-species phenotype ontology (*Uberpheno*) for human, mouse, and zebrafish phenotypes. When defining phenotypes using the EQ-model, the affected entity can either be a biological function or process from GO, or an anatomical entity. Some of the ontologies used to create the definitions are largely species independent (GO, ChEBI). However, anatomical entities are mostly defined by referring anatomy ontologies that are specific for one species. In order to enable reasoning across these vertebrate anatomies, the metazoan, species independent Uberon ontology is used in constructing anatomically-based cross-products [Mungall et al., 2012, Schofield and Hancock, 2012]. In order to construct *Uberpheno*, an equivalence axiom was generated between every class in Uberon that contains a cross-reference to a species anatomy ontology class. Note that very general terms from Uberon such as `tissue` were excluded. These terms were identified by their membership to the subset `upper_level` in Uberon.

Logical definitions have been developed for GO [Mungall et al., 2011], MPO [Mungall et al., 2010], and HPO [Gkoutos et al., 2009]. Almost all logical definitions refer to classes from other ontologies. A set of logical definitions is again an ontology itself. These bridging ontologies (also called cross-product files) are available on the main OBO Foundry website[2], as well as from the individual repositories for each of the projects.

An example for a logical definition is presented in the previous section. A total of 4,874 such EQ definitions were created for HPO terms and used

[2] `http://www.obofoundry.org/index.cgi?show=mappings`

Ontology	File	Size
HPO logical definitions	hp-equivalence-axioms.obo	4,874
MPO logical definitions	mp-equivalence-axioms.obo	6,679
GO logical definitions using Uberon	biological_process_xp_uber_anatomy.obo	1,484
Behavior xp	behavior_xp.obo	110

Table 4.2: HPO and MPO files downloaded from http://code.google.com/p/phenotype-ontologies. Behaviour files downloaded from http://code.google.com/p/behavior-ontology. GO-xp file downloaded from http://obofoundry.org/cgi-bin/detail.cgi?id=biological_process_xp_uber_anatomy.

for this project. A summary of the logical definitions used can be found in Table 4.2. The phenotype ontology logical definitions provide axioms that connect phenotype classes to multiple classes in most of the ontologies listed in Table 4.3.

The HPO and MPO logical definitions were augmented with pairwise equivalence axioms generated by lexical matching. These mappings are represented in a file mp_hp-align-equiv.owl (see the project code archive[3] or the phenotype ontologies archive on Google code[4]). A total of 1,064 such lexically derived equivalence axioms were derived in this way and used to supplement the semantic analysis.

One of the files (see Table 4.2) defines GO process terms by the anatomy term to which the process is related. For example

```
id: GO:0048069 ! eye pigmentation
intersection_of: GO:0043473 ! pigmentation
intersection_of: occurs_in UBERON:0000970 ! eye
```

In order to use these definitions, the different relationships used therein, such as occurs_in, have to be made interpretable for the reasoner. For this project, an additional ontology called extra_equiv.owl was created in which these relationships are made a subPropertyOf of inheres_in.

[3] https://compbio.charite.de/svn/hpo/trunk/misc/uberpheno

[4] http://code.google.com/p/phenotype-ontologies

For zebrafish, no pre-composed ontology of phenotypic abnormalities exists (e.g. there exists no phenotype term such as *decreased width of dorsal aorta*). Instead, the ZFIN project makes use of so-called post-composed annotations, using a combination of classes in the EQ model. Thus, a translation table was implemented, as described in the publication by Mungall et al. [2010], to generate the ontology `zp.owl`. For every modified gene, a set of post-composed phenotype annotations is stored in a file (ZFIN phenotype annotation file). For every unique annotation for zebrafish genes, a class in the ZP identifier space is created. Again, the aforementioned translation to OWL is applied. For example, a zebrafish gene annotation with

```
Entity=ZFA:0000014 (dorsal aorta),
Quality=PATO:0000599 (decreased width) and
Qualifier=PATO:0000460 (abnormal)
```

generates an OWL class:

```
Class: ZP_0013789
Annotations: label "abnormally decreased width dorsal aorta"
EquivalentClassOf:
 has_part some:
   PATO_0000599 and
   inheres_in some ZFA_0000014 and
   qualifier some PATO_0000460
```

Beside generating the ZP-ontology, the annotation relation between the zebrafish genes and ZP-term is written to a file called `zp.annot`. These annotations are used later (see Section 4.3.1).

The ontologies used to construct *Uberpheno* are summarised in Table 4.3. The ontologies that are contained in the OBO Foundry[5] were downloaded on February 15, 2012.

At first, a single, merged OWL ontology is created from all the ontologies and bridging axioms. The ELK reasoner [Kazakov et al., 2011] was used to calculate subclass and equivalence relationships between classes.

[5] http://www.obofoundry.org/

Ontology	Description	File
IMR	Molecule role (INOH Protein-/family name ontology)	MoleculeRoleOntology.obo
MA	Mouse adult gross anatomy	adult_mouse_anatomy.obo
BFO	Basic Formal Ontology	bfo-1.1.owl
CL	Cell Ontology	cell.obo
ChEBI	Chemical Entities of Biological Interest	chebi.obo
GO	Gene Ontology	gene_ontology.1_2.obo
HPO	Human Phenotype Ontology	human-phenotype-ontology.obo
FMA	Foundational Model of Anatomy (adult human)	fma2_obo.obo
MPO	Mammalian Phenotype Ontology	mammalian_phenotype.obo
NBO	Neuro Behavior Ontology	behavior.obo
MPATH	Mouse Pathology	mouse_pathology.obo
PR	Protein Ontology	pro.obo
PATO	Phenotypic Qualities	quality.obo
BSPO	Spatial Ontology	spatial.obo
UBERON	Multi-species anatomy	uberon.obo
ZFA	Zebrafish anatomy and development	zebrafish_anatomy.obo

Table 4.3: Ontologies used to generate the *Uberpheno* ontology. The GO files were downloaded from `http://www.geneontology.org`. The files related to the NBO were downloaded from `http://code.google.com/p/behavior-ontology`.

These steps are implemented within the GULO framework [Köhler et al., 2011]. After the ontology has been reasoned, the Ontologizer API [Bauer et al., 2008] was used to merge all clusters of equivalent classes together into a single class. The HPO identifier is taken as the primary identifier if present.

An excerpt of the *Uberpheno* ontology is shown in Figure 4.3. There one can see how the phenotypic descriptions from different ontologies are combined and automatically organised into a single hierarchy. For instance, the fact that the mouse term *ventricular hypoplasia* is inferred to be a subclass of the human term *Hypoplastic heart* can be used to transfer the information that the mouse gene *Wasf2* is known to cause *ventricular hypoplasia*. In this case, it is possible combine this knowledge with the annotation of the human gene *TBX5* to the HPO term *Hypoplastic heart*.

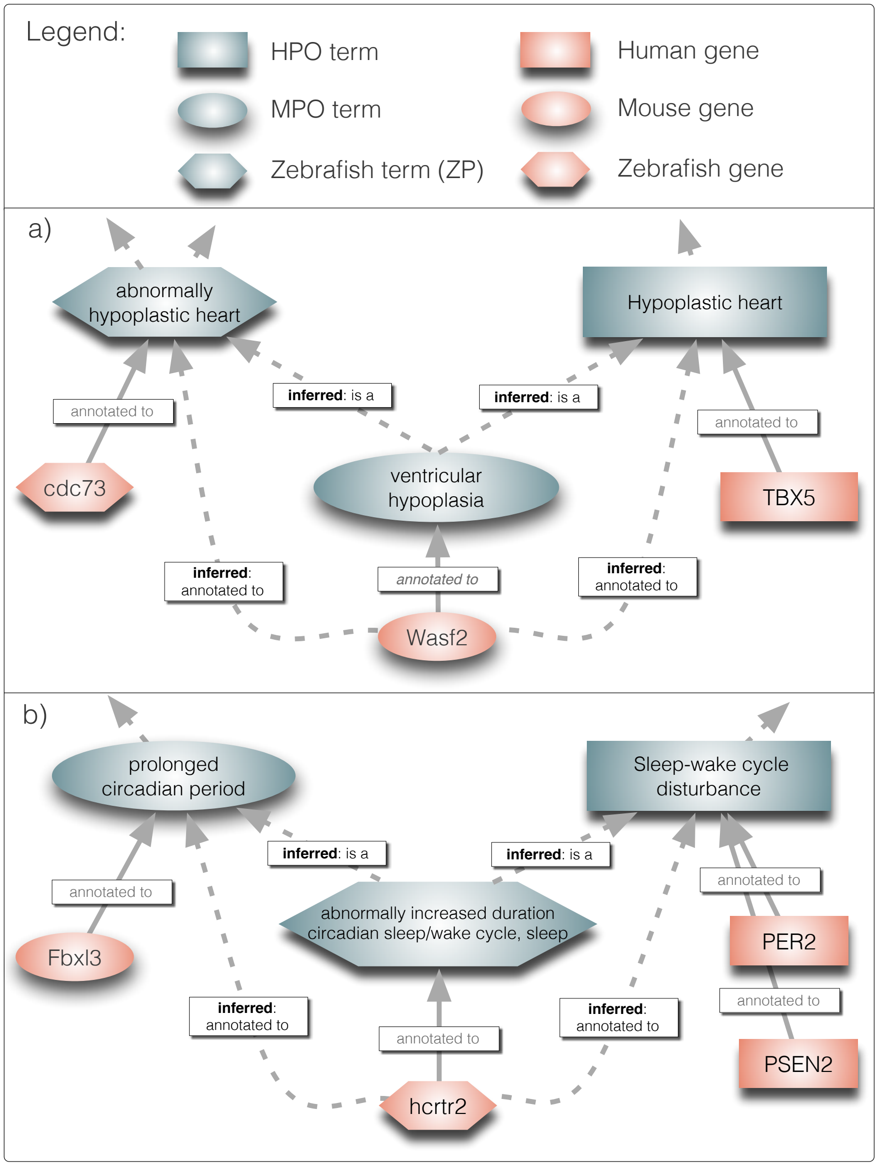


Figure 4.3: Excerpt of the *Uberpheno* ontology to illustrate how information on phenotypic abnormalities in different organisms can be combined. It also illustrates how the annotations of genes can be transferred across different species by means of orthology relationships of genes.

4.3 Cross Species Analysis of CNV Phenotypes

In section 1.1.2, copy number variations (CNVs) are introduced as an important source of human DNA polymorphism. Sequencing of individuals often reveals hundreds of CNVs, but it is often difficult to interpret their phenotypic consequences. Reliable interpretation of CNV data is often difficult and requires expertise.

In assessing a patient with a range of phenotypes and an identified CNV, it is essential to ascertain whether the CNV is causative for the disease or merely incidental. If the CNV is, in fact, the cause of the disease, it is then important to know which of the genes located within the CNV are associated with which of the phenotypic features.

Because phenotypes for mutations in single genes are often not available from human studies, the approach presented here mobilises model organism mutant phenotype data using the *Uberpheno* ontology described in the previous section. This approach makes it possible to connect human disease symptoms and observations made in model organisms. The whole approach is based upon the established premise that pathogenetic mechanisms are evolutionarily conserved [Schofield et al., 2011]. The project presented here investigates the relationship between human phenotypes associated with recurrent CNV disorders, and phenotypic abnormalities associated with human and model organism single-gene diseases whose (orthologous) genes are located within the CNVs. An overview of the complete approach presented here is given in Figure 4.4.

4.3.1 Data for CNV Analysis

In order to perform the analysis, three major data sets have to be compiled. First, the orthology relationships between model organism genes and human genes have to be set up. Secondly, the annotation of the genes with phenotype terms are required. Finally, the CNV disorders have to be annotated by the phenotypic abnormalities associated with the disorder and the set of genes affected within the aberration has to be defined.

Orthology Data

For mouse genes, the file HMD_HumanPhenotype.rpt from the MGI website was used to define the orthology relations to human genes [Blake et al., 2011]. MGI provides a curated set of mammalian orthologs which is constructed through an iterative process using both computational and manual approaches[6].

In order to assign zebrafish genes to their human ortholog, ortho.txt was used to determine those relations. The orthology relationships are manually defined by ZFIN biocurators. This file is available from the ZFIN website[7].

Annotation Data

A phenotypic annotation is a statement that a given disease is characterised by a phenotypic feature. The HPO has been used to annotate 5,035 diseases listed in OMIM. The HPO annotation file was downloaded from http://www.human-phenotype-ontology.org. Because the HPO annotates hereditary syndromes rather than genes directly, phenotype annotations are transferred to the genes that, if mutated, are known to cause the disease. For this purpose, the genemap file from OMIM was used, which associates human genes with OMIM diseases [Amberger et al., 2009]. Positional information for the human genes was obtained from *NCBI Entrez Gene*.

Similarly as for HPO, the MPO has been used to annotate genetically modified mice at the Mouse Genome Informatics (MGI) Project. The file MGI_PhenoGenoMP.rpt[8] was used to obtain mouse gene phenotype annotations [Smith et al., 2005].

As described above, EQ annotations have been used to describe the phenotypes of genetically modified zebrafish at ZFIN. Phenotypic annotations were downloaded from ZFIN[9]. In this project, the file zp.annot

[6] http://www.informatics.jax.org/orthology.shtml

[7] http://zfin.org/data_transfer/Downloads/ortho.txt

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

[9] http://zfin.org/zf_info/downloads.html

(generated by the program described in Section 4.2) was used to obtain zebrafish phenotype annotations [Sprague et al., 2008].

Note that annotations of model organism genes were included only if they applied to a gene with an identified human ortholog. The amount of genes in the three organisms, and the fractions of those that have phenotype annotations, is shown in Figure 4.2. In summary, there are 6,535 mouse genes with phenotype information for which a human ortholog could be identified using information from the MGI sequence group, and there were 1,653 zebrafish genes. Additionally, there were 1,843 human genes with phenotype information on monogenic diseases in the HPO. In all, there were 7,546 human genes with either phenotypic information in human or a phenotypic annotation associated with the ortholog gene in one of the model organisms.

CNV Data

Finally, phenotype annotations for the 27 recurrent CNV diseases analysed in this work were created by manual curation by domain experts using Phenote[10]. The incorporated CNV syndromes are listed in Table 4.5. The annotation lists can be downloaded from the SourceForge project website[11].

To define which genes are affected by the CNV, a conservative approach was chosen, which means that all genes in a maximal critical region as stated by DECIPHER [Firth et al., 2009] were included. For some diseases, a gene that was not included by DECIPHER was added by the curator to the list for the corresponding CNV because of evidence from recent publications stating involvement of the gene. The complete gene lists for the intervals of all 27 CNV disorders are available online[12].

[10] http://phenote.org

[11] http://obo.svn.sourceforge.net/svnroot/obo/phenotype-commons/annotations/OMIM/by-disease/annotated

[12] http://compbio.charite.de/svn/hpo/trunk/misc/deciphergenes

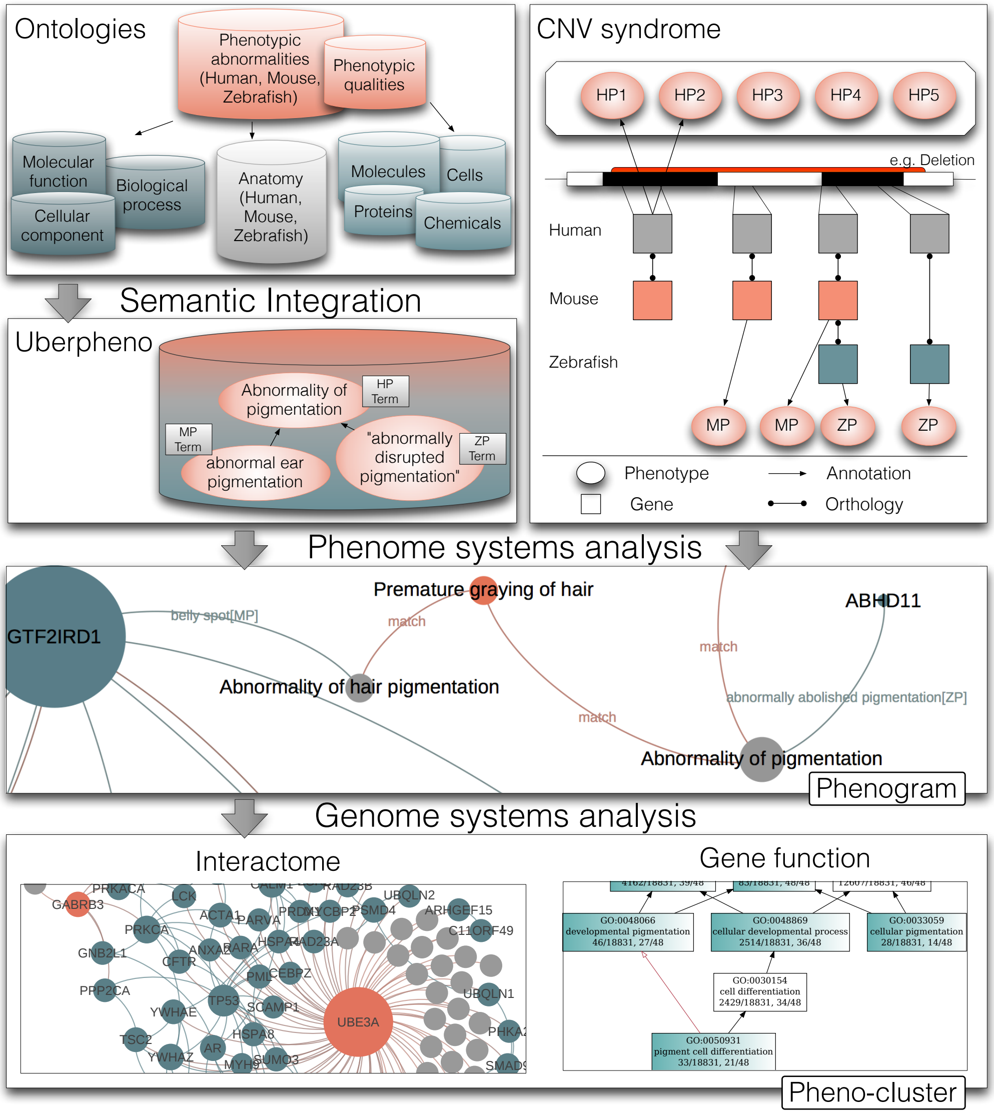

Ontologies
Phenotypic abnormalities (Human, Mouse, Zebrafish)
Phenotypic qualities
Molecular function
Biological process
Cellular component
Anatomy (Human, Mouse, Zebrafish)
Molecules
Cells
Proteins
Chemicals
CNV syndrome
HP1
HP2
HP3
HP4
HP5
e.g. Deletion
Human
Mouse
Zebrafish
MP
MP
ZP
ZP
Phenotype
Gene
Annotation
Orthology
Semantic Integration
Uberpheno
Abnormality of pigmentation
HP Term
MP Term
abnormal ear pigmentation
"abnormally disrupted pigmentation"
ZP Term
Phenome systems analysis
GTF2IRD1
belly spot[MP]
Premature graying of hair
match
Abnormality of hair pigmentation
ABHD11
match
abnormally abolished pigmentation[ZP]
match
Abnormality of pigmentation
Phenogram
Genome systems analysis
Interactome
UBE3A
GABRB3
PRKACA
PRKCA
GNB2L1
PRP2CA
TP53
UBQLN1
UBQLN2
ARHGEF15
C11ORF49
Gene function
GO:0048066 developmental pigmentation 46/18831, 27/48
GO:0048869 cellular developmental process 2514/18831, 36/48
GO:0033059 cellular pigmentation 28/18831, 14/48
GO:0030154 cell differentiation 2429/18831, 34/48
GO:0050931 pigment cell differentiation 33/18831, 21/48
Pheno-cluster

Figure 4.4: Overview of the CNV analysis performed in this chapter. Different ontologies and logical definitions are used to generate the *Uberpheno* cross-species ontology. Then the genes affected by the CNV are retrieved together with the phenotypic abnormalities related to the CNV disease. The phenotypes of the genes are collected from single-gene disorders in humans as well as the phenotypes related to the orthologous genes in model organisms. This data is then analysed in a phenome systems manner and visualised in so called *phenograms*. The *phenograms* are used to generate hypotheses, such as the concept of phenotypic multiplicities, in which multiple genes are said to influence one of the multiple phenotypic features. Thus, a so called genome systems analysis is performed to try to confirm the hypotheses generated before.

4.3.2 Phenograms for CNV Interpretation

For the analysis of copy number variant diseases (CNVs) by means of model organism phenotype data, so called *phenograms* are created. These phenograms link the genes of a CNV with one or more of the various phenotypic features seen in patients with this disease. The goal is to use this network of links in order to predict which gene or combination of genes may be responsible for which part of the phenotypic spectrum of a CNV disease. The symbols and their definition used throughout this part of this thesis are given in Table 4.4.

The analysis of a CNV disorder starts with the set of genes $\mathcal{G}_{\text{CNV}}$ that are located within the corresponding genomic region. For each of the genes $g \in \mathcal{G}_{\text{CNV}}$, there is a set $\mathcal{T}_g$ of associated phenotype terms from human single-gene disorders and of available mouse and zebrafish models. Phenotype annotations for humans, mouse, and zebrafish are mapped to the corresponding terms in *Uberpheno*. Similarly, $\mathcal{T}_{\text{CNV}}$ represents the set of phenotypes associated with the CNV. The manual phenotype curation of CNV disorders also included the frequency information with which a phenotypic features is seen in patients with a given CNV. For the analysis presented here, phenotypic features that are only rarely associated with the CNV (i.e., less than 15 % of affected persons show the feature) were removed before further analysis.

An important concept for the phenogram construction is the information content (IC) of terms in the *Uberpheno* ontology. The information content (IC) of a term t is defined as the negative logarithm of the frequency of annotations to the term [Resnik, 1995] (see Section 1.3.3). Here, the IC of a term in *Uberpheno* is calculated based on the number of genes annotated to the term in humans, mice, and zebrafish. As introduced in Section 1.3.3, the function $anc(t)$ returns the set of ancestral terms for a given term. Additionally, $anc_s(\mathcal{T})$ returns the ancestors of a *set of terms*, i.e.

$$anc_s(\mathcal{T}) = \bigcup_{t \in \mathcal{T}} anc(t) \; . \tag{4.1}$$

The set of common ancestors of an *Uberpheno* term t_g associated with gene

Symbol	*Definition*
g	A gene.
$\mathcal{G}_{\text{CNV}}$	The set of genes located within the CNV.
$\mathcal{T}_g$	The set of associated phenotype terms associated with gene g.
t	A term of *Uberpheno*.
$IC(t)$	Information content of term t.
$anc(t)$	A function that returns the set of ancestral terms for a given term t.
$anc_s(\mathcal{T})$	A function that returns the set of ancestral terms for a given set of terms $\mathcal{T}$.
$\text{CA}(t_g, \mathcal{T}_{\text{CNV}})$	The set of common ancestors of an *Uberpheno* term t_g associated with gene g and the set of *Uberpheno* terms associated with the CNV.
$t_{max}(t_g, \mathcal{T}_{\text{CNV}})$	The term with the highest information content from the set $\text{CA}(t_g, \mathcal{T}_{\text{CNV}})$.
$\mathcal{D}$	The disorder/CNV phenotypes that are explained ((in)directly connected to a gene) in the phenogram.
$\mathcal{U}$	The *Uberpheno* common ancestors shown in phenograms (gray).
$\mathcal{E}$	The edges that connect elements from $\mathcal{G}, \mathcal{U}$, and $\mathcal{D}$.
$\mathcal{G}_p = \{g_1, \ldots, g_n\}$	A *pheno-cluster*; a set of n genes associated with one particular phenotypic feature.

Table 4.4: A summary of the symbols and notations used throughout the chapter 4. Short definitions are given as well.

g and the set of *Uberpheno* terms associated with the CNV is defined as

$$\mathrm{CA}(t_g, \mathcal{T}_{\mathrm{CNV}}) = \{t | t \in anc_s(\mathcal{T}_{\mathrm{CNV}}) \cap anc(t_g)\} \ . \tag{4.2}$$

The term from the set $\mathrm{CA}(t_g, \mathcal{T}_{\mathrm{CNV}})$ with the highest information content is returned by the function $t_{max}(t_g, \mathcal{T}_{\mathrm{CNV}})$. In cases where multiple terms have the same maximal IC, one of them is randomly chosen.

Using these concepts introduced above, a phenogram is defined as a structure $(\mathcal{G},\mathcal{U},\mathcal{D},\mathcal{E},\ell)$, where $\mathcal{G}$ refers to the genes that are annotated to phenotypic features (in one of the organisms). Each of these features has a common ancestor in *Uberpheno* with a CNV feature d. Note that this common ancestor may be the root of the ontology for terms with no similarity. These common ancestors must have an IC above ℓ in order to be a member of the set $\mathcal{U}$. The IC threshold ℓ is introduced to exclude associations to relatively non-specific phenotypic features. Here, ℓ is set to a value of 2.5. The CNV features d that lead to the common ancestor are defined to be members of the set $\mathcal{D}$. The set $\mathcal{E}$ contains directed edges that connect genes with phenotypes of $\mathcal{U}$ and edges linking phenotypes of $\mathcal{U}$ to phenotypes of $\mathcal{D}$. In summary, $\mathcal{G}$ consists of all genes for which a single-gene phenotype t_g shares a common ancestor ($\mathcal{U}$) with a phenotype of the CNV (set $\mathcal{D}$), whereby the common ancestor has an information content above the threshold ℓ, i.e., $IC(t_{max}(t_g, \mathcal{T}_{\mathrm{CNV}})) \geq \ell$.

These phenograms can be nicely visualised using the Gephi-toolkit library [Bastian et al., 2009], whereby the genes are colored blue, the explained CNV phenotypes ($\mathcal{D}$) are colored red, and the above threshold common ancestors ($\mathcal{U}$) are shown in gray. The genes are connected to one or more phenotype terms t_{max}. For visualisation purposes, the thickness of edges between genes and phenotypes reflect the number of phenotype annotations (t_{max}) that support this link. Also, these edges are labeled with the amount of supporting phenotype annotations in HP (HS for Homo Sapiens), MP (MM for Mus Musculus), or ZP (DR for Danio Rerio). Note that every path that leads from a gene to a CNV phenotype represents a possible explanation, i.e. it can be hypothesised that the gene plays a role or is causative in the development of this particular phenotype of the CNV.

A representative phenogram is shown for Williams-Beuren syndrome

Syndrome	MIM ID	Genes	Phenogram matches
1q21.1 susceptibility locus (TAR)	274000	19 (3)	ITGA10, TXNIP, HFE2
1p36 microdeletion syndrome	607872	70 (12)	SKI, GABRD, HES5, DVl1, GNB1, PEX10, TP73, VWA1, NOC2L, AGRN, PRDM16, PRKCZ
3q29 microduplication syndrome	611936	22 (2)	DLG1, NCBP2
9q subtelomeric deletion syndrome	610253	8 (2)	EHMT1, CACNA1B
15q24 microdeletion syndrome	613406	36 (8)	CYP11A1, CSPG4, PTPN9, CSK, STRA6, CAX5A, MPI, NEIL1
15q26 overgrowth syndrome	D:81	29 (4)	IGF1R, ALDH1A3, PCSK6, CHSY1
17q21.3 microdeletion syndrome	610443	6 (2)	MAPT, CRHR1
Angelman syndrome	105830	50 (7)	UBE3A, OCA2, GABRB3, SNRPN, NDN, NIPA1, GABRA5
Cri du Chat syndrome	123450	42 (11)	SLC9A3, SLC6A3, NKD2, TERT, IRX1, CCT5, SDHA, SLC6A19, MTRR, SLC12A7, NDUFS6
Familial Adenomatous Polyposis	175100	3 (1)	APC
Leri-Weill dyschondrostosis	127300	1 (1)	SHOX
Miller-Dieker syndrome	247200	37 (14)	ABR, CRK, DPH1, HIC1, MNT, PAFAH1B1, PITPNA, YWHAE
NF1-microdeletion syndrome	613675	13 (2)	NF1, ATAD5
Pelizaeus-Merzbacher disease	312080	9 (2)	PLP1, GLRA4
Phelan-Mcdermid syndrome	606232	4 (3)	SHANK3, ARSA, MAPKBIP2
Potocki-Lupski syndrome	610883	47 (10)	RAI1, ULK2, TOM1L2, MYO15A, ALDH3A2, ATPAF2, PEMT, SREBF1, MAPK7, EPN2
Potocki-Shaffer syndrome	601224	15 (4)	ALX4, EXT2, CD82, SLC35C1
Prader-Willi syndrome	176270	50 (8)	GABRB3, HERC2, NDN, NIPA1, OCA2, UBE3A, GABRA5, SNRPN
RCAD (renal cysts and diabetes)	137920	11 (3)	HNF1B, ACACA, LHX1
Rubinstein-Taybi syndrome	180849	1 (1)	CREBBP
Smith-Magenis syndrome	182290	47 (13)	RAI1, SREBF1 MYO15A, LLGL1, PEMT, ALDH3A2, ATPAF2, TNFRDF13B, FLCN, MAPK7, TOM1L2, ULK2, B9D1
Sotos syndrome	117550	39 (6)	NSD1, SLC34A1, SNCB, PROP1, B4GALT7, CPLX2
Split hand/foot malformation 1	183600	6 (2)	DLX5, DLX6
WAGR 11p13 deletion syndrome	194072	5 (2)	PAX6, WT1
Williams-Beuren syndrome	194050	34 (11)	ELN, BAZ1B, LIMK1, GTF2IRD1, STX1A, NCF1, ABHD11, FZD9, CLIP2, MLXIPL, FKBP6
Wolf-Hirschhorn syndrome	194190	36 (7)	WHSC1, FGFRL1, FGFR3, IDUA, CTBP1, TACC3, PDE6B
Xq28 (MECP2) duplication	D:45	23 (6)	MECP2, BGN, L1CAM, ABCD1, AVPR2, SLC6A8

Table 4.5: Phenogram results for the 27 CNV diseases summarising candidate genes, as well as references to known genotype-phenotype associations from the literature. The column MIM ID shows the ID for the Online Mendelian Inheritance in Man database [Amberger et al., 2009] where available. Otherwise, the DECIPHER ID [Firth et al., 2009] is shown as “D:xx”. The “genes” column shows the number of genes located within the CNV, and the number of genes for which a phenogram match was obtained is shown in parentheses. “phenogram matches” shows genes identified by our method as candidates for individual phenotypic features of the CNV disorders. Previously known associations are indicated by literature citations.

in Figure 4.5. Williams-Beuren is a multisystem disorder caused by a microdeletion of 34 genes. Patients with this disorder have numerous clinical features affecting the cardiovascular, endocrine, gastrointestinal, musculoskeletal, and neurological system. They also show developmental, dentitional, ophthalmologic, behavior, and skin phenotypes [Pober, 2010]. For this CNV, 39 associations between 11 candidate genes and 26 phenotypic abnormalities were found. Following that, manual curation of literature showed that 16 of those have previously been reported by Pober [2010], and that 23 associations represent novel gene to phenotype links. For Williams syndrome, the analysis yields that many phenotypic features are predicted to be associated with more than one possible candidate gene. This can be seen, for example, in the lower part of the phenogram where CNV phenotype *Cutis laxa* is associated with both genes *ELN* and *GTF2IRD1*.

There exist other examples, such as the phenogram of Pelizaeus-Merzbacher disease shown in Figure 4.6. In contrast to the phenogram shown for Williams-Beuren Syndrome, 10 of the phenotypic features could be assigned to a single gene (*PLP1* [Boespflug-Tanguy et al., 1994]).

In total, this study identified 802 candidate genes for individual phenotypic features of the 27 investigated recurrent CNV disorders. From these associations, manual curation (Sandra Doelken, Barbara Ruef) suggests that 346 of these represent novel associations that have previously not been reported in the literature. It is furthermore interesting to investigate which organism presented the source of the explanations for the phenotypic features. The Venn diagram in Figure 4.7 shows that 431 of the 802 predictions were made only on the basis of model organism data (54 %), with the mouse being the most influential source of phenotype information.

In order to asses the statistical significance of the results of the analysis, a simple null hypothesis is introduced, which allows one to calculate P-values for the phenograms. This is achieved by generating a randomised CNV model. For this purpose, the set of CNVs and their phenotypic abnormalities is kept fixed while a randomly selected interval of genes ($\mathcal{G}_r$) is used to replace the original set of genes $\mathcal{G}$. Note that $\mathcal{G}_r$ and $\mathcal{G}$ are the

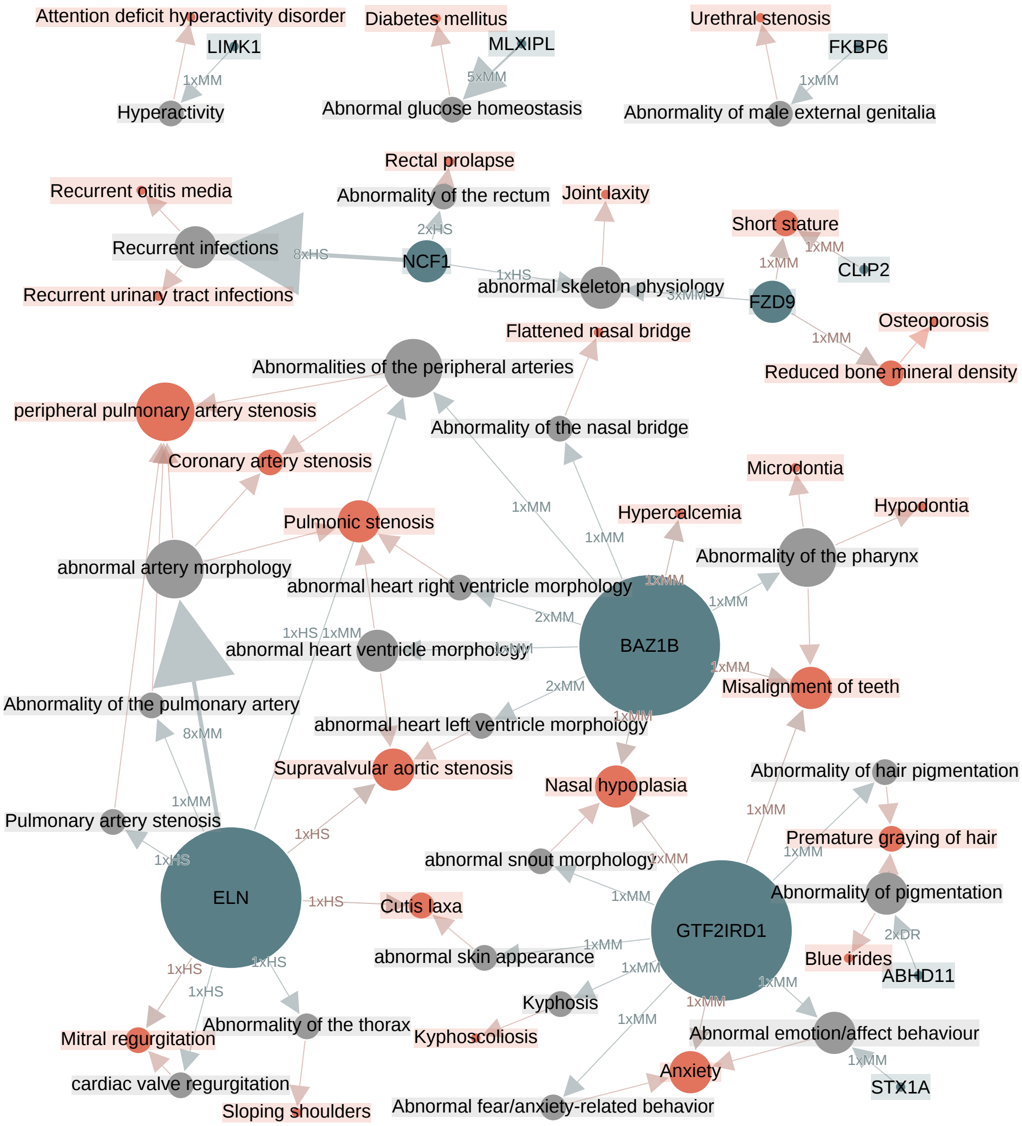

Attention deficit hyperactivity disorder
LIMK1
1xMM
Hyperactivity
Diabetes mellitus
MLXIPL
5xMM
Abnormal glucose homeostasis
Urethral stenosis
FKBP6
1xMM
Abnormality of male external genitalia
Rectal prolapse
Abnormality of the rectum
Joint laxity
Recurrent otitis media
Recurrent infections
8xHS
2xHS
NCF1
1xHS
abnormal skeleton physiology
3xMM
Recurrent urinary tract infections
Short stature
1xMM
1xMM
CLIP2
FZD9
Osteoporosis
1xMM
Reduced bone mineral density
Flattened nasal bridge
Abnormalities of the peripheral arteries
peripheral pulmonary artery stenosis
Abnormality of the nasal bridge
Coronary artery stenosis
Microdontia
1xMM
Hypodontia
1xMM
Pulmonic stenosis
Hypercalcemia
abnormal artery morphology
Abnormality of the pharynx
abnormal heart right ventricle morphology
1xMM
1xMM
2xMM
1xHS
1xMM
abnormal heart ventricle morphology
1xMM
BAZ1B
1xMM
Misalignment of teeth
2xMM
Abnormality of the pulmonary artery
abnormal heart left ventricle morphology
1xMM
8xMM
Supravalvular aortic stenosis
Nasal hypoplasia
Abnormality of hair pigmentation
1xMM
1xMM
Pulmonary artery stenosis
1xHS
Premature graying of hair
1xMM
1xHS
abnormal snout morphology
1xMM
ELN
1xHS
Cutis laxa
1xMM
GTF2IRD1
Abnormality of pigmentation
1xMM
2xDR
abnormal skin appearance
Blue irides
1xHS
1xHS
1xMM
ABHD11
1xHS
1xMM
Kyphosis
1xMM
Abnormality of the thorax
1xMM
Mitral regurgitation
Kyphoscoliosis
Abnormal emotion/affect behaviour
1xMM
Anxiety
cardiac valve regurgitation
STX1A
Sloping shoulders
Abnormal fear/anxiety-related behavior

Figure 4.5: Phenogram for Williams-Beuren Syndrome. The blue nodes represent genes that are deleted or duplicated in the CNV. The red nodes represent phenotypic features that patients with this CNV show ($\mathcal{D}$) and that could be assigned to a gene. The gray nodes show the above threshold common ancestors ($\mathcal{U}$) in the *Uberpheno*. The size of a node correlates with the number of edges that are associated with the node. The size of the edges correlate with the amount of supportive data such as the number of mouse models.

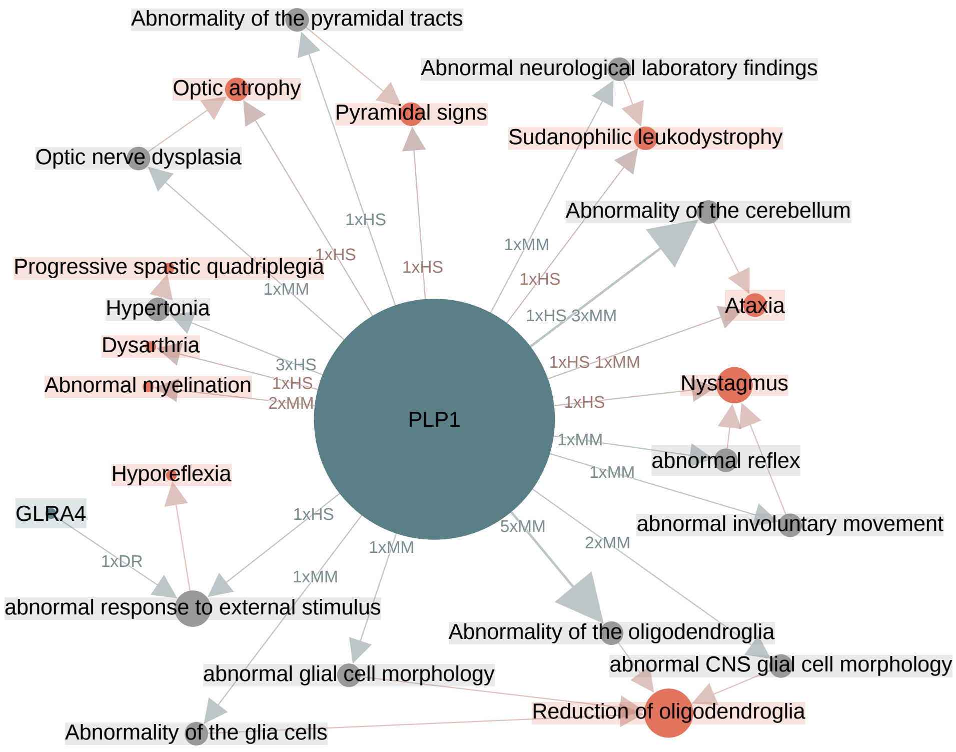

Abnormality of the pyramidal tracts
Abnormal neurological laboratory findings
Optic atrophy
Pyramidal signs
Sudanophilic leukodystrophy
Optic nerve dysplasia
Abnormality of the cerebellum
1xHS
1xMM
1xHS
Progressive spastic quadriplegia
1xHS
1xHS
1xMM
Hypertonia
Ataxia
1xHS 3xMM
Dysarthria
3xHS
1xHS 1xMM
Abnormal myelination
1xHS
Nystagmus
2xMM
1xHS
PLP1
1xMM
Hyporeflexia
abnormal reflex
1xMM
GLRA4
1xHS
abnormal involuntary movement
1xDR
5xMM
2xMM
1xMM
1xMM
abnormal response to external stimulus
Abnormality of the oligodendroglia
abnormal CNS glial cell morphology
abnormal glial cell morphology
Reduction of oligodendroglia
Abnormality of the glia cells

Figure 4.6: Phenogram for Pelizaeus-Merzbacher disease. The blue nodes represent genes that are deleted or duplicated in the CNV. The red nodes represent phenotypic features that patients with this CNV show ($\mathcal{D}$) and that could be assigned to a gene. The gray nodes show the above threshold common ancestors ($\mathcal{U}$) in the *Uberpheno*. The size of a node correlates with the number of edges that are associated with the node. The size of the edges correlate with the amount of supportive data such as the number of mouse models.

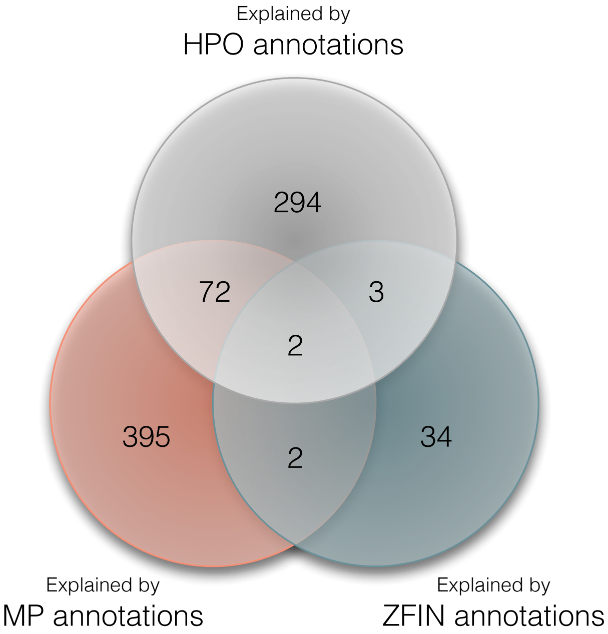


Figure 4.7: Sources of the explanations of the 802 predicted phenotypic features of the 27 CNV disorders examined in this work. 431 of the 802 predictions were made only on the basis of model organism data (54 %).

same size. From this model, two empirical P-values are calculated by performing and evaluating the random CNVs 5,000 times.

The first P-value is used to assess the overall utility of the approach. For this it is assumed that each of the phenotypic matches displayed in the phenograms represents a potential *explanation* of a phenotypic feature of the CNV. That means if a CNV-phenotype can be reached by traversing along the edges from one of the genes, it is postulated that this gene is responsible for the phenotype. Although it may be possible that individual matches may be due to chance, the total number of above-threshold matches in all CNVs is assumed to provide a useful measure of the utility of the method. The P-value is calculated by summing the size of $\mathcal{D}$ over all of the 27 CNVs. Then it was recorded how often this value was exceeded in each of the 5,000 randomisations. For the real CNVs, the set $\mathcal{D}$ contained 482 phenotypic features. This value was never reached during randomisation, which corresponds to P-value of less than 0.0002. These results are visualised in Figure 4.8 (A).

For the second P-value, it is essential to be able to quantify the phenograms for further analysis and statistical evaluation. Here, for each gene $g \in \mathcal{G}_{\text{CNV}}$, a phenomatch score S_g is defined based on the IC of the terms in $\mathcal{U}$, i.e. the matching terms with an IC above ℓ:

$$S_g(g, \mathcal{T}_{\text{CNV}}) = \sum_{\substack{t_g \in \mathcal{T}_g \\ IC(t_{max}(t_g, \mathcal{T}_{\text{CNV}})) \geq \ell}} \left[IC(t_g) \right]^k \ . \tag{4.3}$$

By choosing $k > 1$, matches based on terms with higher information content are given a higher weighting. In this analysis, k was set to 5. The full phenogram score across all genes located in the CNVs[13] is then calculated as

$$S\text{pg}(\mathcal{G}_{\text{CNV}}, \mathcal{T}_{\text{CNV}}) = \sum_{g \in \mathcal{G}_{\text{CNV}}} S_g(g, \mathcal{T}_{\text{CNV}}) \ . \tag{4.4}$$

The second P-value is calculated to investigate individual phenograms

[13] OMIM contains some entries that correspond to CNV diseases such as Rubinstein Taybi Deletion syndrome (MIM:610543). These OMIM entries are connected to the corresponding genes in `genemap` as well. To avoid bias, the HPO annotations for these syndromes were excluded from the analysis.

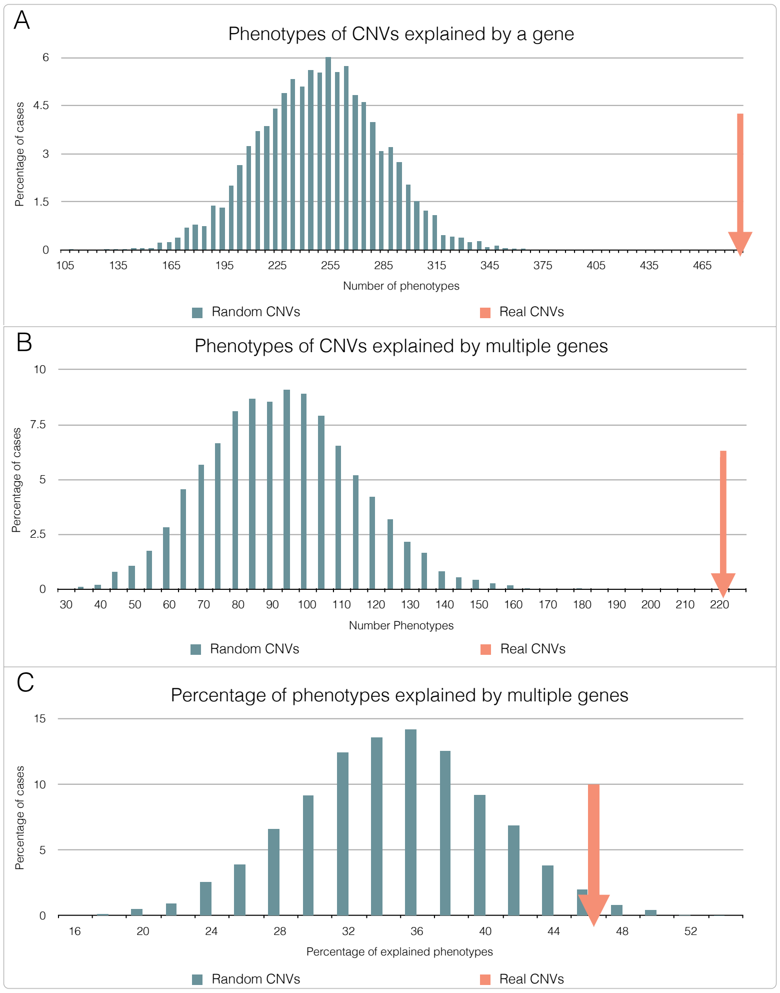


Figure 4.8: Histograms **(A)**,**(B)** and **(C)** illustrate the distribution of phenotypes, with phenogram matches for the 27 CNVs investigated in this study (red arrow) versus randomly chosen CNVs (blue bars). **(A)** Number of phenotypes explained by one gene for randomly generated versus real CNVs **(B)** Number of phenotypes explained by multiple genes (pheno-clusters) for randomly generated versus real CNVs **(C)** Percentage of phenotypes explained by multiple genes as a percentage of all matching candidate genes for randomly generated versus real CNVs. All results are statistically significant including **(C)** which supports the conclusion that pheno-clusters are not a characteristic of randomly chosen chromosomal segments **(**$P = 0.021$**)**.

of one of the 27 CNVs and their associated score (S_{pg}). In other words, a P-value for the null hypothesis that a phenogram score of S_{pg} or greater for a specific CNV has been observed by chance is calculated. This is calculated by evaluating how often $S_{\text{pg}}(\mathcal{G}_{\text{CNV}}, \mathcal{T}_{\text{CNV}})$ is equal to or exceeded by $S_{\text{pg}_r}(\mathcal{G}_{\text{CNV}}, \mathcal{T}_{\text{CNV}})$. The results of this analysis are represented in Table 4.6.

From the 27 investigated CNV syndromes, 14 have a statistically significant phenogram score (S_{pg}) with a significance threshold of 0.05. Note that a non-significant S_{pg} doesn't imply that the result isn't useful for interpretation of the genotype to phenotype relationships. This lack of significance in a statistical sense could be related to several factors. These factors include limitations of our computational approach, inadequate phenotypic annotations, or incomplete knowledge about the genes located within the CNV. It is thus investigated if specific patterns may explain the drop in the P-values.

It is easy to see that the current approach, as described above, is dependent upon the granularity of the phenotype descriptions. Unspecific, broadly used phenotype terms such as *intellectual disability* will not lead to statistical significant hits, because they are so frequently used in annotations. The provided P-values are directly dependent on the IC of the phenotypic features. For example, the term *intellectual disability* has a very low IC of 3.2 and will thus not contribute strongly to the overall S_{pg}, and randomly chosen intervals are very likely to contain genes that are annotated to those general terms. As can be seen in Figure 4.9 (A), the P-values of the phenogram scores reported in Table 4.6 correlate linearly ($R^2 = 0.28$) with the granularity of the phenotypic descriptions of the CNV disorders. The granularity is hereby measured using the average IC of the CNV phenotypes.

The same effect can be seen when correlating the P-values with the size of the CNV. The size is thereby measured as the total number of genes located within the CNV region. In Figure 4.9 (B), one can see that this correlates as well with an R^2 of 0.28.

Many recently characterised CNV disorders that have been delineated on the basis of Array-CGH screening in contrast to clinical studies have substantially less specific clinical pictures. Given those nonspecific clin-

Disease name	Annotations ($f > 15$ %)	Genes (+phen) (+match)	S_{pg}	P-value
Xq28 (MECP2) duplication	31 (14)	23 (12) (6)	1.08×10^5	< 0.0002
NF1-microdeletion syndrome	32 (27)	13 (3) (2)	3.97×10^5	< 0.0002
Leri-Weill dyschondrostosis	26 (13)	1 (1) (1)	1.56×10^5	0.0002
Familial Adenomatous Polyposis	19 (4)	3 (1) (1)	5.25×10^4	0.0002
WAGR 11p13 deletion syndrome	14 (9)	5 (2) (2)	1.10×10^5	0.0004
Pelizaeus-Merzbacher disease	26 (21)	9 (3) (2)	6.80×10^4	0.0004
Potocki-Shaffer syndrome	25 (23)	15 (9) (4)	9.83×10^4	0.0026
Split hand/foot malformation 1	11 (9)	6 (3) (2)	2.89×10^4	0.0050
Sotos syndrome	38 (21)	39 (14) (6)	8.01×10^4	0.0122
Rubinstein-Taybi syndrome	112 (73)	1 (1) (1)	2.10×10^4	0.0138
Angelman syndrome	34 (25)	50 (9) (7)	7.88×10^4	0.0184
RCAD (renal cysts and diabetes)	23 (14)	11 (4) (3)	2.41×10^4	0.0216
Williams-Beuren syndrome	92 (68)	34 (13) (11)	1.24×10^5	0.0316
Wolf-Hirschhorn syndrome	81 (64)	36 (13) (7)	1.38×10^5	0.0478
Potocki-Lupski syndrome	32 (28)	47 (22) (10)	5.06×10^4	0.0628
9q subtelomeric deletion syndrome	38 (29)	8 (2) (2)	1.33×10^4	0.0662
Phelan-Mcdermid syndrome	54 (43)	4 (4) (3)	8.19×10^3	0.0728
Prader-Willi syndrome	66 (52)	50 (9) (8)	9.39×10^4	0.0788
17q21.3 microdeletion syndrome	51 (37)	6 (2) (2)	6.93×10^3	0.1094
Miller-Dieker syndrome	42 (41)	37 (21) (14)	5.70×10^4	0.1192
15q26 overgrowth syndrome	37 (31)	29 (5) (4)	1.89×10^4	0.2028
1p36 microdeletion syndrome	86 (60)	70 (22) (12)	9.42×10^4	0.2762
Smith-Magenis syndrome	46 (40)	47 (22) (13)	3.41×10^4	0.2916
15q24 microdeletion syndrome	65 (56)	36 (15) (8)	3.80×10^4	0.2938
1q21.1 susceptibility locus (TAR)	44 (16)	19 (5) (3)	5.08×10^3	0.3178
Cri du Chat syndrome	68 (48)	42 (21) (11)	3.29×10^4	0.3374
3q29 microduplication syndrome	22 (14)	22 (6) (2)	1.31×10^3	0.5156

Table 4.6: Summary of the 27 CNV disorders used in the analysis. Included are the total number of phenotypic annotations as well as the number of annotations per disease above a frequency threshold of 15 % ($f > 15$ %), the number of genes per interval and the corresponding number of genes with phenotype information (+phen), the number of phenogram matches identified in our study (+match), as well as the phenogram score (S_{pg}) and the empirical P-values of S_{pg} for 5,000 randomisations.

ical phenotypes and high phenotypic variability, the diagnostic process may be complicated. This may represent an explanation as to why diseases associated with micro-duplications of 3q29 [Ballif et al., 2008] and micro-deletions of 15q24 [Andrieux et al., 2009] do not obtain statistical significant scores compared to more distinct CNV disorders. Of course, the target here was to identify statistically significant phenotypic matches. But again it should be noted that non-significant results for individual CNVs do not imply futility of the results. When transferrering the presented methods to a clinical decision support system, this would probably still be designed to present users the best match or matches for both specific and less specific phenotypic abnormalities.

4.3.3 Phenotypic Multiplicities of CNV Genes

In the phenogram analysis described before, groups of genes $\mathcal{G}_p$ located in the same CNV were found to be associated with the same phenotypic abnormality. Note that the size of $\mathcal{G}_p$ ($|\mathcal{G}_p| = n$) must be at least 2. In terms of the visualisation (e.g. Figure 4.5), this means that one of the phenotypes in red can be reached by traversing along the edges emanating from two or more different genes. An example for this in the Williams-Beuren syndrome are the two genes *NCF1* and *FZD9* which are predicted to be involved in the *Joint laxity* phenotype. The genes of these groups were shown to be associated with a similar phenotypic abnormality in isolation. A group of genes associated with one phenotypic features are termed a *pheno-cluster*.

These physical clusters of genes associated with particular shared phenotypes in the genome might be causative for a larger subset of the phenotypes observed in a CNV. Even genes that do not show dosage effects in isolation may contribute to a particular phenotypic abnormality if one or more pathway members are simultaneously affected. A well-known example has been described for the SHFM1 locus. There, the genes *DLX5* and *DLX6* are known to cause split-hand/split-foot malformation (SHFM). It has been shown that mouse models possess the SHFM-phenotype only if both genes are knocked out [Merlo et al., 2002, Robledo et al., 2002]. Thus,

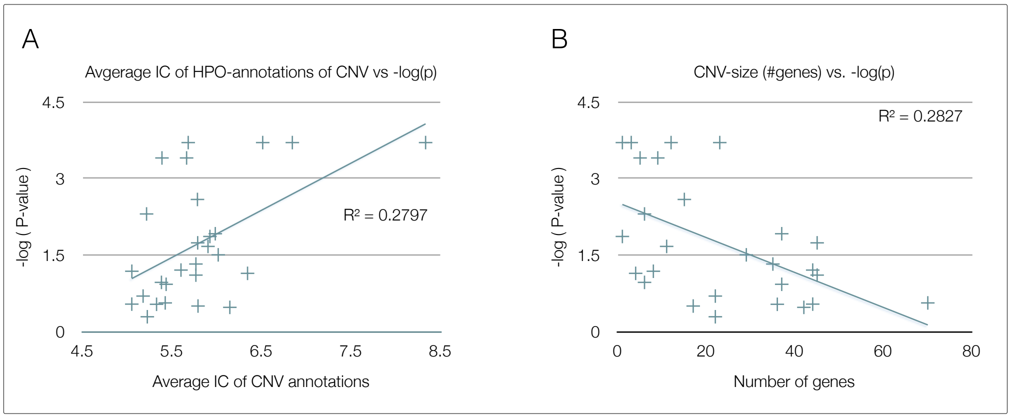


Figure 4.9: The statistical significance of the phenogram score S_{pg} is correlated with several measures. Here, the correlation for two examples is shown. Note that the negative logarithm of the P-value is taken, meaning that high values on the y-axis represent higher significance.

In **A**, it can be seen that the P-values correlate with an R^2 of 0.28 with the specificity of the phenotype information of the CNV disorders. The specificity is measured by the average information content (IC) of the CNV phenotypes. Thus, unspecific phenotypic annotations, i.e. terms with low IC, lead to less significant phenogram scores.

In **B**, the correlation to the number of genes that are affected by the CNV is shown. The correlation coefficient (R^2) amounts again to 0.28. Thus, a higher number of genes in a CNV interval is associated with a lesser degree of statistical significance of the phenogram.

interaction or vicinity of different genes affected by a CNV may be an important determinant of clinical severity. It was thus investigated whether such phenotypic summation effects due to *pheno-clusters* are encountered in the analysis more often than would be expected by chance.

The total number of *pheno-clusters* found in our analysis was more than twice that of the randomised data, which is visualised in the histogram of Figure 4.8 (B). This corresponds to P-value of less than 0.0002. One could argue that this may be, in part, a consequence of the lower overall number of genes and phenotypes identifed in the randomised data. It was thus additionally examined if the percentage of phenotypes in the randomised data explained by multiple genes is also higher than expected by chance. Even here, the percentage of pheno-clusters was significantly greater for the 27 analysed CNVs than expected according to the randomised model (P = 0.02; Figure 4.8 (C)).

In all, *pheno-clusters* were predicted for 220 phenotypes, corresponding to 135 gene clusters (in some cases, the same genes were associated with more than one phenotype). It is known that the chromosomal location of genes can be related to their function.

It was therefore investigated, if the hypothesis is true that these *pheno-clusters* of genes $\mathcal{G}_p = \{g_1, \ldots, g_n\}$ are not only related to the same phenotypic feature but also share similarity based on other biological measures. Here, the similarity based on Gene Ontology (GO) annotations of the genes in $\mathcal{G}_p$ is calculated. Furthermore, was is examined if the genes are located in close vicinity to one another within protein-protein interaction networks (PPIN) using the random-walk-based analysis presented in Chapter 2.

The homogeneity of $\mathcal{G}_p$ based on GO is computed using the average pairwise similarity for all unique pairs of genes in $\mathcal{G}_p$:

$$\underset{\mathrm{GO}}{\mathrm{HOM}}(\mathcal{G}_p) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} sim_{\mathrm{GO}}(g_i, g_j) \ . \tag{4.5}$$

For a pair of genes, the symmetric semantic similarity $sim_{\mathrm{GO}}(g_i, g_j)$ as in Equation 1.5 is calculated. The annotations of genes with GO-terms were taken from the NCBI data repository[14] without filtering by evidence code. To determine a P-value for a given homogeneity score $\mathrm{HOM}_{\mathrm{GO}}(\mathcal{G}_p)$, an empirical score distribution is computed by randomly generating 10,000 random gene groups $\mathcal{G}_r$ with the same number of genes. For these random groups $\mathrm{HOM}_{\mathrm{GO}}(\mathcal{G}_r)$ is computed, and from the distribution of these scores the P-value can be estimated. For this, the fraction of cases in which $\mathrm{HOM}_{\mathrm{GO}}(\mathcal{G}_r) \geq \mathrm{HOM}_{\mathrm{GO}}(\mathcal{G}_p)$ is computed.

In order to test the hypothesis that genes in the same pheno-cluster also tend to cluster in the human protein interaction network, a network ($\mathcal{N}$) containing 11,302 nodes was analysed. The nodes correspond to human genes coding for proteins with known interactions. The network was taken from the NCBI data repository[15] and six genes were excluded because they had more than 250 interactions[16].

The network similarity is calculated as described in Chapter 2. The random-walk matrix $\mathbf{R}$ is calculated as described in Section 2.3. Recall from this section that every entry $\mathbf{R}_{i,j}$ represents the probability of a random walker starting at node i and being at node j after an infinite number of steps. For a group of genes $\mathcal{G}_p$, the average global network proximity $\mathrm{GNP}(\mathcal{G}_p)$ is computed by:

$$\mathrm{GNP}(\mathcal{G}_p) = \frac{1}{n} \sum_{g_i \in \mathcal{G}_p} \vec{p}^{\,i}_{\infty}[g_i] \,, \tag{4.6}$$

whereby $\vec{p}^{\,i}_{\infty}$ is calculated as $\mathbf{R} \times \vec{p}^{\,i}_{0}$ (see Equation 2.13). For every gene g_i in the *pheno-cluster*, a different start vector $\vec{p}^{\,i}_{0}$ is defined. To determine this vector, the start probability of a network node k is defined as:

$$\vec{p}^{\,i}_{0}[g_k] = \begin{cases} \frac{1}{n-1} & \text{if } g_k \in \{\mathcal{G}_p \setminus g_i\} \\ 0 & \text{otherwise} \end{cases} \,. \tag{4.7}$$

[14] `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz`

[15] `ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz`

[16] This was mainly done for visualisation purposes. These hubs (nodes with a huge number of interactions) tended to occur in almost every plot (e.g. *UBC* or *CREBBP*). The results were hardly affected (only some P-values were increased, i.e. the signal was weaker).

pheno-cluster ($\mathcal{G}_p$)	Phenotype (CNV)	P-value for GNP($\mathcal{G}_p$)	P-value for HOM$_{\text{GO}}$($\mathcal{G}_p$)
SRR, CRK, PAFAH1B1, YWHAE, MYO1C	*Lissencephaly* (Miller-Dieker Syndrome)	*0.0056*	*0.0009*
TP73, PEX10	*Dilatation of lateral cerebral ventricles* (1p36 Deletion Syndrome)	*0.028*	0.19
NDN, OCA2, GABRA5, GABRB3, SNRPN	*Hyperactivity* (Angelman Syndrome)	0.17	*0.0008*

Table 4.7: Examples for the secondary analysis of *pheno-clusters*.

Put simply, when analysing a particular g_i, the random walker starts with equal probability from all nodes in G_p except from g_i. Then the random walk distance from all the start nodes to g_i is computed and the average over all $g_i \in G_p$ is taken as GNP.

Similar to the GO analysis, a P-value for a given score GNP$(\mathcal{G}_p)$ is determined by setting up the empirical score distribution. This is done by randomly generating 10,000 random gene groups $\mathcal{G}_r$ of the same size and computing GNP$(\mathcal{G}_r)$. Afterwards, the P-value is estimated as the fraction of cases in which GNP$(\mathcal{G}_r) \geq$ GNP$(\mathcal{G}_p)$.

For evaluation, a significance threshold of 0.05 was applied. When analysing the functional similarity of genes within each of the 136 *pheno-clusters* based on Gene Ontology criteria, it was shown that 48 of them demonstrated a statistically significant intracluster similarity. The random-walk analysis showed that in 24 of the *pheno-clusters* the genes are in closer proximity in the protein interactome than expected by chance. Note that 19 phenoclusters could not be analysed using the random-walk approach, because too many genes lacked interaction data, i.e. $|\mathcal{G}_p| = n < 2$. In sum, 43 % of the analysable *pheno-clusters* were shown to have a significant intracluster similarity in one of the two methods. Examples for *pheno-clusters* and the P-values are given in Table 4.7.

To further investigate the utility of the approaches presented in this thesis, one example of an identified *pheno-cluster* is depicted. During the analysis, a set $\mathcal{G}_p$ containing the five genes *SRR*, *CRK*, *PAFAH1B1*, *YWHAE*, *MYO1C* was found to be involved in the *Lissencephaly* phenotype in Miller-

pheno-cluster gene	Example Mouse Phenotype annotation
SRR	*decreased susceptibility to neuronal excitotoxicity* (MP:0008236)
CRK	*abnormal cerebral cortex morphology* (MP:0000788)
PAFAH1B1	*abnormal neuronal migration* (MP:0006009)
YWHAE	*abnormal hippocampus morphology* (MP:0000807)
MYO1C	*abnormal vestibular hair cell physiology* (MP:0004438)

Table 4.8: Examples for the mouse phenotype annotations that led to these genes being predicted as a *pheno-cluster* for *Lissencephaly*. Note that vestibular hair cells are sensory neurons.

Dieker Syndrome. *Lissencephaly* (smooth brain) is a phenotype where the patient's brain shows a lowered number of neurons. This can be seen as a dramatic decrease in the number of gyri in the cortex [Hatten, 1999]. Examples for the underlying phenotype annotations of these genes in mouse are given in Table 4.8. The five genes of this $\mathcal{G}_p$ have a significant GO homogeneity score ($\mathrm{HOM}_{\mathrm{GO}}(\mathcal{G}_p)$) and are additionally found to be in significantly close proximity in the protein interactome (see Table 4.7). It was then investigated if the members of the PPI-subnetwork connecting these genes are as well related to biological phenomenon of *Lissencephaly* so that this may be used as a tool to find novel hypothesis regarding gene-to-phenotype connections. For instance, genes in the network-vicinity of the *pheno-clusters* could be members of the same pathways or represent factors that influence related pathways by crosstalk with other cellular processes.

For this analysis, the *pheno-cluster* subnetwork of $\mathcal{N}$ is defined by retrieving all neighbors (in $\mathcal{N}$) of the members of $\mathcal{G}_p$ and all edges between them. Note that the genes of $\mathcal{G}_p$ are part of the subnetwork as well. The resulting network for the *Lissencephaly* genes is shown in Figure 4.10. There, it can be seen that genes already known to cause *Lissencephaly* are present, i.e. *TUBA1A* and *PAFAH1B1*, together with two other genes coding for subunits of Platelet-activating factor (PAF). Note that *TUBA1A* and the subunits of PAF were not part of the initial *pheno-cluster*.

It was now investigated whether this subnetwork, i.e. the surrounding nodes of $\mathcal{G}_p$, represent an interesting biological signal. The following analysis was performed with the colored (blue or red) nodes of Figure 4.10, i.e.

the ones with more than one edge in the subnetwork.

At first, these 204 nodes were imported into the *Ontologizer* [Bauer et al., 2008] for investigation of enriched GO terms. Using the method *Topology-Elim* the top hit from the "biological process" sub-ontology is *nerve growth factor receptor signaling pathway* (GO:0048011), with *P*-value of $1.24e^{-27}$ after Bonferroni correction. 28 genes in the analysis were annotated to this GO term. This reflects well the known biological connection to disturbed neuronal migration, which is assumed to be a cause of *Lissencephaly* [Dobyns et al., 1996]. The enrichment of genes annotated to this pathway adds further value to the results of the analysis based on model organism phenotypes and interaction networks.

A second analysis investigated the tissue expression patterns of the genes in the subnetwork, because genes that are not expressed in the cells or organs related to *Lissencephaly* would argue against the hypothesis that the subnetwork is meaningful for this phenotype. Using Biomart[17] [Haider et al., 2009], the attributes *HGNC symbol* and *GNF/Atlas organism part* were extracted for the complete set of human genes. The downloaded file maps 14,818 genes to one or multiple tissues in which the genes are expressed. The GNF Gene Expression Atlas uses a vocabulary of 60 organ descriptions, which were manually mapped to FMA terms. From the subnetwork, a total of 180 nodes were assigned to at least one *GNF/Atlas organism part*. For each of these genes, it was checked if they are expressed in an anatomical structure that is an asserted or inferred subclass of *Brain* (FMA:50801) or *Segment of brain* (FMA:55676). 114 (63 %) of the nodes are found to be expressed in one of these structures. The Fisher-exact test revealed a *P*-value of 0.008. Also, investigating the individual anatomical structures in which the genes of the subnetwork are expressed reveals a strong bias towards the brain. Again the Fisher-exact test was used, but accompanied with Bonferroni correction for multiple testing. Except for one, all of the significantly affected anatomical structures (indicated by * in Figure 4.11) are part of the brain[18].

[17] Biomart Portal, Version 0.7, accessed May 2012.

[18] *Cerebellum peduncles* are not a subclass of the used brain-terms and are a subclass (inferred) of *Cell part cluster of neuraxis* (FMA:83143). The *brain* is a `regional_part_of` the

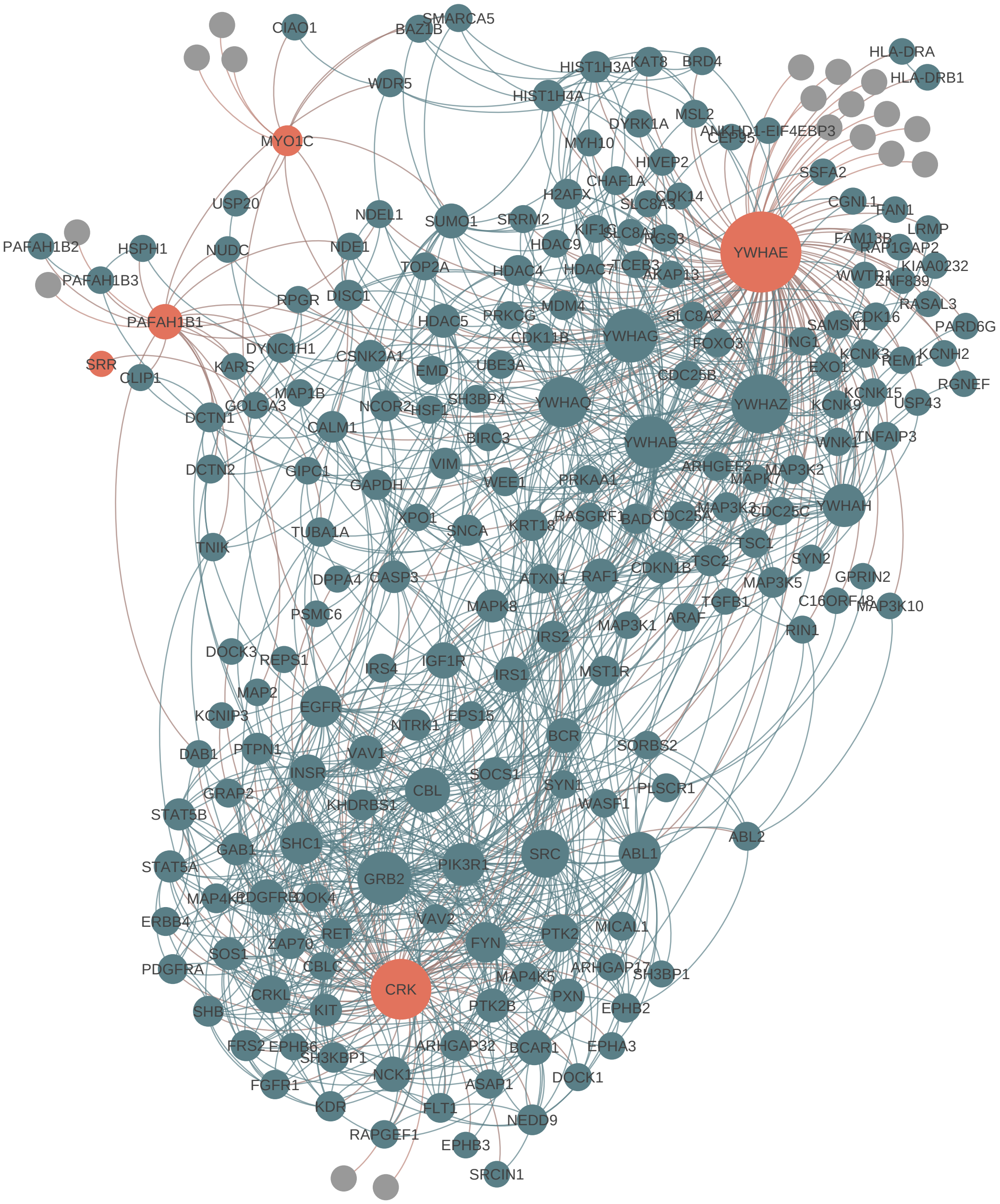
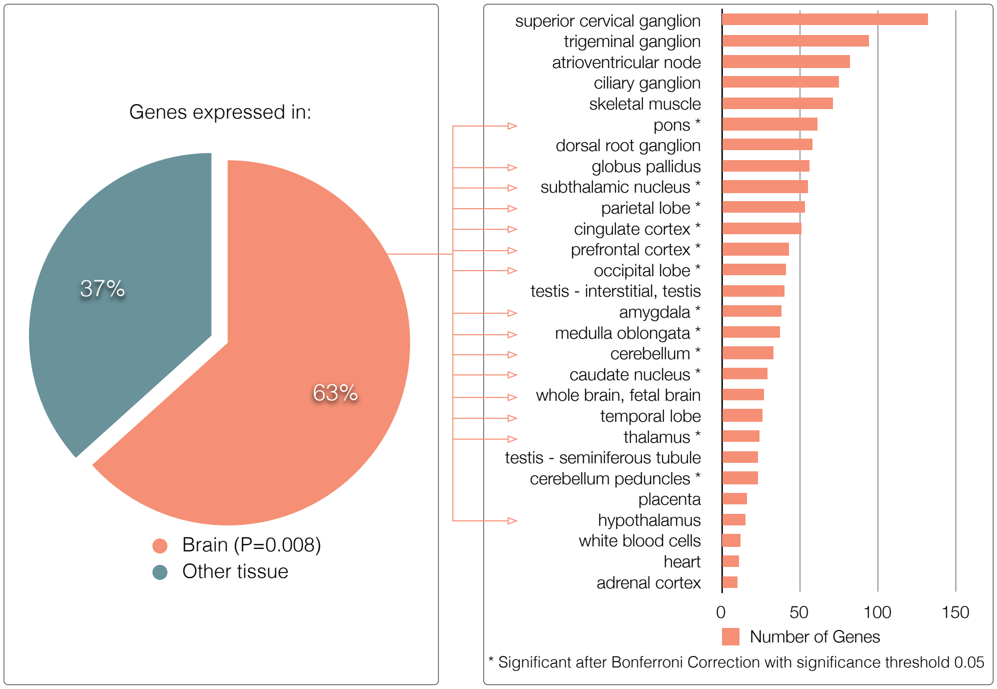


Figure 4.10: The subnetwork connecting the genes predicted to be involved in the *Lissencephaly* phenotype of Miller-Dieker Syndrome.



* Significant after Bonferroni Correction with significance threshold 0.05

Figure 4.11: Analysis of the colored genes shown in Figure 4.10, which are hypothesised to play a role in the development of *Lissencephaly*.

The combination of cross-species phenotype analysis with the analysis of high-throughput biological data, such as PPIN and gene expression data thus represents a novel approach for assessing the contribution of single genes to individual phenotypic features of a broader spectrum of phenotypes. From these relations, combinatorial effects can be discovered, in which multiple genes affect one phenotype. These multiplicities could be shown to correlate well with other measures such as GO similarity and PPIN vicinity. Especially the combination with PPINs can be a promising novel tool for the discovery of novel modules of molecular players related to specific cellular processes that are disrupted.

There exist other approaches to computational cross-species phenotype comparison. For example, PhenomicDB [Groth et al., 2007] uses only lexical matching. Also, there exists an approach that is predominantly based on lexical comparisons and aimed at mapping the UMLS Metathesaurus to the MPO and concepts of human disease [Sardana et al., 2010]. A drawback of this method is that it disregards different conceptions of a disease and a phenotype in humans and mice [Hoehndorf et al., 2011b].

The *Uberpheno* ontology presented here is different, and probably superior, to previous methods based on text mining. For instance, the *Uberpheno* method recognises that a *Nasal hypoplasia* in humans is a subclass of the mouse phenotype *abnormal snout morphology*. The major drawback of textual methods is that they do not utilise the complete semantic information contained in the relevant building block ontologies.

Similar to Hoehndorf et al. [2011b], the approach presented in this thesis uses the complete phenotypic repertoire of mouse and zebrafish, and applies automated reasoning over all of the phenotype ontologies to generate a representation that can be explored through measures of phenotypic similarity. Novel in the construction of the *Uberpheno* is that it is assisted by equivalence relations between HPO and MPO, which were identified by lexical matching. Additionally, logical definitions of GO terms are used. In these, GO cellular processes are defined by the anatomical structures that these processes affect (e.g. `inheres_in`). For example, ZFIN uses an annotation to *disruption* of the GO process *lens development in camera-*

Neuraxis (FMA:55675).

type eye (GO:0002088). Since the GO process is defined to be a developmental anomaly that inheres in the *lens* (UBERON:0000965), the reasoner infers correctly that the zebrafish phenotype is a subclass of the mouse phenotype *abnormal lens development* (MP:0005545).

The application of *Uberpheno* differs from other approaches, which were focused on the ranking of disease relevant genes. When doing this, the task is rather to completely match two phenotypic profiles consisting of two sets of phenotypic features. In this thesis, the focus is the assignment of single genes from a bigger group of genes to one particular phenotypic feature. These phenotypic features are part of a broader phenotypic spectrum that makes up a disease. This means the focus is not to find a gene that explains everything, rather the task is to explain each candidate genes individual contribution to a disease.

In summary, the herein presented methods facilitate the automated computational integration of phenotype information from the many model organism databases and several other ongoing projects aiming at genotype-phenotype analysis and model-organism research. Sophisticated knowledge representation approaches for phenotypic abnormalities enable the detection of inconsistencies and contradictions in the data. This approach enables the exploitation of the great potential of the information produced by systematic genome-wide phenotyping efforts, such as the IMPC, in order to assist gene to phenotype research.

Chapter 5

Discussion

Networks have long being assumed to be essential for understanding biological systems, and recently it has become more and more evident that human diseases result from perturbations of cellular systems such as molecular networks. In this thesis, it could be shown that network algorithms that measure not only direct and shortest-path interactions but also take the global network structure into account have a clear performance advantage in the prioritisation of candidate disease genes. These findings add weight to the assumption that phenotypically similar diseases are associated with disturbances of subnetworks within the protein interactome, and that exploration of global network structures with appropriate graph-theoretic algorithms will become an important resource for understanding the biology of disease.

The publication of the random-walk method applied to disease gene prediction has had quite a large impact[1] on the community, and also because of this it seems valuable to extend the ideas presented therein. The results have been confirmed, for example, in a publication by Navlakha and Kingsford [2010]. Li and Patra [2010] extend the random-walk ideas by including a network of disease similarities and performing the analysis on a so called heterogenous network, in which the PPIN has been augmented by a network of disease similarities calculated by text mining. Other ideas are to include multiple genome-wide data sets, such as

[1] According to Google Scholar there are 174 citations. (Accessed July 2012)

GO similarity or co-expression, into the adjacency matrix in the beginning. For these methods, it will have to be tested how much the ranking performance is influenced by each data source, and if sophisticated weighting schemes are required in order to integrate the results obtained by different data sources. This is currently work in progress. Also, the *GeneWanderer* application may profit from some extensions in the functionality of the web interface. At the moment, only Entrez Gene gene identifiers are allowed when defining the genes known to be related to the investigated phenotypes. Being allowed to use identifiers from other gene-centric databases has been requested by users as an improvement.

One drawback of the presented method is that it requires at least some knowledge of the molecular background of the disease being investigated. This a property inherent to all *guilt-by-association* approaches, because if an algorithm wants to establish guilt, there has to be something to which the unknown data can be compared against. In the setting of disease gene prioritisation a set of genes known or suspected to be part of the pathobiological process underlying the investigated phenotypes has to be set up. Very often, such a list of known genes has to be compiled manually, which is often time consuming or even impossible. In particular, there exist a huge amount of orphan diseases for which no molecular information is available. Thus, it is desirable to develop automated methods that may replace this manual procedure to assist users. The herein presented HPO may help in this task, because users should be able to define a set phenotypic features of interest. This set of HPO terms can, in turn, be used to retrieve highly similar diseases and then filter out those diseases for which the molecular cause is known. These ideas, of course, have to be evaluated in terms of usability and performance, which is again work in progress.

The HPO is a standardized, controlled vocabulary that enables phenotypic information to be described in an unambiguous fashion in medical publications and databases [Robinson and Mundlos, 2010]. As with other biomedical ontologies, the HPO specifies the meaning of terms in a vocabulary so that humans and machines can understand and process the nature of the data [Hoehndorf et al., 2011b]. The HPO is now widely adopted in the genetics community. International consortia, such ISCA, dbGAP, NCBI's Genetic Testing Registry, or DECIPHER, are now annotating all their data with terms from the HPO [Riggs et al., 2012]. Especially, given the constantly growing number of logical definitions of terms, the HPO can be a valuable tool for efficient detection of associations between genotype and phenotype.

Furthermore, it was shown that the HPO can be used in practice, by using an ontological semantic similarity analysis for clinical diagnostics. The differential diagnosis in clinical genetics can often be challenging, because of the large number of distinct syndromes and phenotypic features that need to be considered. Also, the fact that pathognomonic signs are rare, and in many cases combinations of more or less specific clinical features are needed for a diagnosis, complicates this process. For this reason, a number of commercial and freely available computational tools have been developed, including the LDDB [Fryns and de Ravel, 2002], POSSUM [Bankier and Keith, 1989], OMIM [Amberger et al., 2009], and Orphanet [Aymé, 2003]. These rely mainly on identifying lists of syndromes characterized by at least a certain number of phenotypic features entered by the user. Also, these have not provided a means of determining whether any given match is significant in a statistical sense. The method presented in this thesis procedure makes use of the semantic structure of the HPO in order to weight the importance of the query and disease terms according to their clinical specificity. Also, it was shown that the rankings of the differential diagnoses can be improved by introducing a simple statistical model, which assigns a P-values to each obtained similarity score. A freely available tool called the Phenomizer was presented. During one month in the summer 2012[2] the Phenomizer was accessed by 900 users from 38 different countries and had approximately 30 visitors per day. The tool is now being extended so that users may integrate the algorithms easily into their pipeline and retrieve Phenomizer results from a web service, where the input is part of a URL. Such a service prevents users from being forced to manually define their query using the webin-

[2] Statistics generated from 26th June to 26th July 2012 (3 years following publication) using a tool available at http://www.revolvermaps.com/?target=enlarge&i=4qfhq0flvot&color=ff0000&m=3.

terface, which is adequate for clinicians. Unfortunately, this prevents the programmatic integration of the Phenomizer in large-scale projects or external tools that wish for an automated solution. These ideas are already realised in a preliminary version developed in cooperation with Decipher. The Phenomizer was developed using the GWT-EXT library[3], which has not been further developed or extended since 2009. Thus, one should start replacing this legacy API by actively developed alternatives such as *smartgwt*[4].

Besides this, one of the most important implications will be the ability of automated, or at least computer-guided, classification of novel syndromes, based on the spectrum of phenotypic characteristics [Köhler et al., 2012]. Especially for orphan diseases the presented methods may provide a helpful tool, in particular when it is possible to identify phenotypically similar diseases, for which knowledge on pathobiology or treatment exists. Advances in high-throughput technologies in molecular genetics and the computational analysis of networks led to the concept of the diseasome [Goh et al., 2007, Barabasi, 2007], which refers to the network of complex relationships between biochemical, genetic, cellular, phenotypic, and other networks. The importance of diseasomics has become ever more obvious in the field of human genetics with the identification of biochemical or protein interaction networks whose dysfunction underlies groups of phenotypically related diseases [Brunner and van Driel, 2004, Lim et al., 2006]. Recent computational projects have shown the potential of incorporating human phenotypic data into the analysis of cellular networks [Goh et al., 2007, Gottlieb et al., 2011, Bayés et al., 2011].

However, the wide acceptance and usage of ontologies introduces novel obstacles. Also, ontologies of the biomedical domain are becoming increasingly interdependent (see Chapter 3). Thus it is advisable for the HPO developers to occupy themselves with state of the art ontological engineering strategies. This refers especially to the activities regarding the development process of the ontology and its life cycle. With growing complexity, it will be essential that tools for automatic quality control and

[3] http://code.google.com/p/gwt-ext/

[4] http://code.google.com/p/smartgwt/

release management are tightly integrated into the iterative development process in the near future. Otherwise, mistakes by individual curators may stay undetected for longer periods and propagate to other users and tools. Already being applied are ready to use systems for continuous integration such as Hudson[5] or Jenkins, which are web-based environments for running integration checks. This already does and will in the future largely improve the whole ontology and annotation process.

In recent years, systematic genome-wide phenotyping efforts such as the IMPC for mouse or the phenotyping efforts in zebrafish [Wang et al., 2007, Kettleborough et al., 2011] have been started. The data provided by those efforts are of great potential for gaining novel insights into pathobiology and for uncovering new candidates for genes involved in human disease.

As noted in a paper by Aitman et al. [2011], novel sequencing technologies promise a new golden age for human genetics. Even if it will be easier to identify new human disease genes and candidate disease genes, it will not be easier to answer the question how mutations in these genes cause disease and what biological processes are affected. In case a poorly characterized gene is linked to a human disease it does not per se make this gene better understood it just makes it 'more interesting' [Aitman et al., 2011]. This is where model organisms research can show its strength, given that reliable information extraction across species is possible. The methods presented in this thesis may represent a valuable tool in this tasks.

In general, the presented approaches may be a key ingredient of automated tools that aim towards filtering and interpreting the huge amount of variations found by comprehensive sequencing of patients. By utilising species-agnostic ontological representations in the logical definitions, phenotype-based analysis across species and domain are enabled. For example, simple queries such as the following are basically built-in features of the data provided by HPO:

> "Generate a list of all NGS-sequenced patients in my database that have a genomic variant in a gene that is related to *zinc ion*

[5] See `http://compbio.charite.de/hudson`

homeostasis (GO:0055069)".

For this, a program has to identify all the HPO-terms that are defined by referring to the given GO-term or a more specific one. In this case, it will be only *Abnormality of zinc homeostasis* (HP:0008277). Afterwards, one can simply extract all the genes related to this HPO-term and additionally make use of available model organism phenotype data. Note that this will also return genes annotated to descendants of the HPO-term, such as *Increased serum zinc* (HP:0011424). The program finally returns the patients found to have a genomic variation in one of the genes.

In this thesis, it was shown that the application of ontologies for phenotype annotation can give highly significant explanations of the complex interplay between a set of aberrated genes and specific phenotypic consequences of this aberration. Using the EQ method for the phenotypic annotations facilitates the use of a common language required for systematic comparison of phenotypes [Washington et al., 2009]. It was shown that logical definitions in combination with automated reasoning can be used to improve ontology development and facilitate the construction of consistent knowledge representation across multiple ontologies [Mungall et al., 2010, Köhler et al., 2011]. The herein presented methods make use of semantic similarities between human and model organisms. This will facilitate the computational integration of information from different phenotyping projects and enable the harvest of these rich resources in an automated way. The presented algorithms can easily be adapted to assist with interpretation and understanding of the diagnostic results from Array-CGH analyses. The ideas presented here could also be applied to the interpretation of Next-Generation-Sequencing data, thereby moving closer to the objective of a personalised genetic approach to medical care. In summary, the approach for representation of phenotype information provide a means to improve downstream analyses with the ultimate goal of uncovering molecular disruptions that cause and influence human disease.

Of course, there is space for improvement of cross-species analysis. In the near future, the Uberpheno will be extended by incorporating Least Common Subsuming (LCS) class expressions. This means that more classes are added to the Uberpheno based on the knowledge represented in the building block ontologies used by the logical definitions. For example, if a zebrafish gene is annotated to *increased concentration of copper*, the Uberpheno should then also contain the class *increased concentration of divalent metal cation*, whereby *divalent metal cation* is a superclass of *copper* in ChEBI. This should increase the sensitivity of all methods utilising the Uberpheno. A similar approach was already applied in the publication by Chen et al. [2012].

Of course, the found effect of the phenotypic multiplicities in CNVs (Chapter 3) represents an interesting biological signal, which needs to be further analysed. For example it may be tested if aberrated enhancer elements are underlying those effects. It will also be interesting to investigate the network structure of the affected subnetworks in context of the larger PPIN.

In summary, this thesis presented novel methods for making use of different networks, be it cellular PPIN or the network defined by the semantic relationships between phenotype or other biomedical concepts. These networks are used for generating possible explanations of biological phenomena such as hereditary diseases and the effect of CNVs. The combination of different network-based methods was presented in the last chapter and shown to be a valuable novel tool for unraveling the molecular basis of phenotypic abnormalities.

Bibliography

OWL tools. URL `http://code.google.com/p/owltools/`.

Protégé. URL `http://protege.stanford.edu`.

S. Aerts, D. Lambrechts, S. Maity, P. V. Loo, B. Coessens, F. D. Smet, L.-C. Tranchevent, B. D. Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537–44, 2006.

T. J. Aitman, C. Boone, G. A. Churchill, M. O. Hengartner, T. F. C. Mackay, and D. L. Stemple. The future of model organisms in human disease research. *Nat Rev Genet*, 12(8):575–82, 2011.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

C. Alfarano, C. E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo *et al.* The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, 33(Database issue):D418–24, 2005.

D. Altshuler, M. Daly, and L. Kruglyak. Guilt by association. *Nat Genet*, 26 (2):135–7, 2000.

D. Altshuler, M. J. Daly, and E. S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–8, 2008.

J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh. Mckusick's online mendelian inheritance in man (OMIM). *Nucl. Acids Res.*, 37(Database issue):D793–D796, 2009.

J. Andrieux, C. Dubourg, M. Rio, T. Attie-Bitach, E. Delaby, M. Mathieu, H. Journel, H. Copin, E. Blondeel, M. Doco-Fenzy, E. Landais, B. Delobel, S. Odent, S. Manouvrier-Hanu, and M. Holder-Espinasse. Genotype-phenotype correlation in four 15q24 deleted patients identified by array-cgh. *Am J Med Genet A*, 149A(12):2813–9, 2009.

S. Aymé. Orphanet, an information site on rare diseases. *Soins; la revue de référence infirmière*, (672):46–7, 2003.

F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *The description logic handbook: theory, implementation, and applications*. Cambridge University Press, 2003. ISBN 0-521-78176-0.

B. C. Ballif, A. Theisen, J. Coppinger, G. C. Gowans, J. H. Hersh, S. Madan-Khetarpal, K. R. Schmidt, R. Tervo, L. F. Escobar, C. A. Friedrich, M. McDonald, L. Campbell, J. E. Ming, E. H. Zackai, B. A. Bejjani *et al.* Expanding the clinical phenotype of the 3q29 microdeletion syndrome and characterization of the reciprocal microduplication. *Mol Cytogenet*, 1:8, 2008.

A. Bankier and C. G. Keith. Possum: the microcomputer laser-videodisk syndrome information system. *Ophthalmic paediatrics and genetics*, 10(1): 51–2, 1989.

A.-L. Barabasi. Network medicine–from obesity to the "diseasome". *N Engl J Med*, 357(4):404–7, 2007.

A.-L. Barabasi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12(1):56–68, 2011.

J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005.

J. B. L. Bard and S. Y. Rhee. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet*, 5(3):213–22, 2004.

M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI Conference on Weblogs and Social Media*, 2009.

S. Bauer. *Algorithms for Knowledge Integration in Biomedical Sciences*. PhD thesis, Freie Universität Berlin, 2011.

S. Bauer, S. Grossmann, M. Vingron, and P. N. Robinson. Ontologizer 2.0–a multifunctional tool for go term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650, 2008.

S. Bauer, S. Köhler, M. H. Schulz, and P. N. Robinson. Bayesian ontology querying for accurate and noise-tolerant semantic searches. *Bioinformatics (Oxford, England)*, 2012.

A. Bayés, L. N. van de Lagemaat, M. O. Collins, M. D. R. Croning, I. R. Whittle, J. S. Choudhary, and S. G. N. Grant. Characterization of the proteome, diseases and evolution of the human postsynaptic density. *Nat Neurosci*, 14(1):19–21, 2011.

G. W. Beadle and E. L. Tatum. Genetic control of biochemical reactions in neurospora. *Proc Natl Acad Sci USA*, 27(11):499–506, 1941.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

L. G. Biesecker. Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clinical genetics*, 68(4):320–6, 2005.

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 Research*, 39(Database issue):D842–8, 2011.

O. Boespflug-Tanguy, C. Mimault, J. Melki, A. Cavagna, G. Giraud, D. P. Dinh, B. Dastugue, and A. Dautigny. Genetic homogeneity of Pelizaeus-Merzbacher disease: tight linkage to the proteolipoprotein locus in 16 affected families. PMD Clinical Group. *Am J Hum Genet*, 55(3):461–7, 1994.

S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–5, 2009.

D. Botstein and N. Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genetics*, 2003.

Y. Bradford, T. Conlin, N. Dunn, D. Fashena, K. Frazer, D. G. Howe, J. Knight, P. Mani, R. Martin, S. A. T. Moxon, H. Paddock, C. Pich, S. Ramachandran, B. J. Ruef, L. Ruzicka *et al.* Zfin: enhancements and updates to the zebrafish model organism database. *Nucleic Acids Research*, 39(Database issue):D822–9, 2011.

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

T. Can, O. Çamoğlu, and A. K. Singh. Analysis of protein-protein interaction networks using random walks. In *Proceedings of the 5th international workshop on Bioinformatics*, BIOKDD '05, pages 61–68, New York, NY, USA, 2005. ACM.

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*, 33(5):858–66, 2012.

F. M. Couto, M. J. Silva, and P. M. Coutinho. Measuring semantic similarity between gene ontology terms. *Data & Knowledge Engineering*, 2007.

F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 1958.

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 (Oxford, England)*, 23(16):2198–200, 2007.

P. de Matos, R. Alcántara, A. Dekker, M. Ennis, J. Hastings, K. Haug, I. Spiteri, S. Turner, and C. Steinbeck. Chemical entities of biological interest: an update. *Nucleic Acids Research*, 38(Database issue):D249–54, 2010.

W. B. Dobyns, E. Andermann, F. Andermann, D. Czapansky-Beilman, F. Dubeau, O. Dulac, R. Guerrini, B. Hirsch, D. H. Ledbetter, N. S. Lee, J. Motte, J. M. Pinard, R. A. Radtke, M. E. Ross, D. Tampieri *et al.* X-linked malformations of neuronal migration. *Neurology*, 47(2):331–9, 1996.

G. Fang, N. Bhardwaj, R. Robilotto, and M. B. Gerstein. Getting started in gene orthology and functional analysis. *PLoS Comput Biol*, 6(3):e1000703, 2010.

J. H. Finger, C. M. Smith, T. F. Hayamizu, I. J. McCright, J. T. Eppig, J. A. Kadin, J. E. Richardson, and M. Ringwald. The mouse Gene Expression Database (GXD): 2011 update. *Nucleic Acids Research*, 39(Database issue): D835–41, 2011.

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, 2009.

L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–25, 2006.

J.-P. Fryns and T. J. L. de Ravel. London Dysmorphology Database, London Neurogenetics Database and Dysmorphology Photo Library on CD-ROM [Version 3] 2001 R. M. Winter, M. Baraitser, Oxford University Press. *Hum Genet*, 111(1):113, 2002.

T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–93, 2006.

R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Research*, 34(19):e130, 2006.

M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. What is a gene, post-encode? history and updated definition. *Genome Res*, 17(6): 669–81, 2007.

G. V. Gkoutos, E. C. J. Green, A.-M. Mallon, J. M. Hancock, and D. Davidson. Using ontologies to describe mouse phenotypes. *Genome Biol*, 6(1): R8, 2004.

G. V. Gkoutos, C. J. Mungall, S. Dolken, M. Ashburner, S. E. Lewis, J. M. Hancock, P. N. Schofield, S. Köhler, 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, 2009.

A. M. Glazier, J. H. Nadeau, and T. J. Aitman. Finding genes that underlie complex traits. *Science*, 298(5602):2345–9, 2002.

K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi. The human disease network. *Proc Natl Acad Sci USA*, 2007.

A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol*, 7:496, 2011.

J. Gross and J. Yellen. *Graph theory and its applications*. CRC Press, Inc., 2006. ISBN 1-58488-505-x.

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 Research*, 35(Database issue):D696–9, 2007.

T. Gruber. A translation approach to portable ontology specifications. *Knowledge acquisition*, 1993.

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.

S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, and A. Kasprzyk. Biomart central portal–unified access to biological data. *Nucleic Acids Research*, 37(Web Server issue):W23–7, 2009.

J. M. Hancock, A.-M. Mallon, T. Beck, G. V. Gkoutos, C. J. Mungall, and P. N. Schofield. Mouse, man, and meaning: bridging the semantics of mouse phenotype and human disease. *Mammalian Genome*, pages 1–5, 2009.

M. E. Hatten. Central nervous system neuronal migration. *Annu Rev Neurosci*, 22:511–39, 1999.

P. Hitzler, M. Krötzsch, and S. Rudolph. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2009.

R. Hoehndorf, M. Dumontier, A. Oellrich, D. Rebholz-Schuhmann, P. N. Schofield, and G. V. Gkoutos. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLoS ONE*, 6(7):e22006, 2011a.

R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos. Phenomenet: a whole-phenome approach to disease gene discovery. *Nucleic Acids Research*, 39 (18):e119, 2011b.

D. Horn, P. Krawitz, A. Mannhardt, G. C. Korenke, and P. Meinecke. Hyperphosphatasia-mental retardation syndrome due to PIGV mutations: Expanded clinical spectrum. *Am J Med Genet A*, 155(8):1917–22, 2011.

M. Horridge and S. Bechhofer. The OWL API: A Java API for working with OWL 2 ontologies. *6th OWL Experienced and Directions Workshop, Chantilly, Virginia*, 2009.

M. Horridge, N. Drummond, J. Goodwin, A. Rector, R. Stevens, and H. Wang. The manchester OWL syntax. *OWL: Experiences and Directions*, pages 10–11, 2006.

W. L. Johannsen. *Elemente der exakten Erblichkeitslehre*. G. Fischer, Jena, 1909.

Y. Kazakov, M. Krötzsch, and F. Simančík. Concurrent classification of $\mathcal{EL}$ ontologies. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist, editors, *Proceedings of the 10th International Semantic Web Conference (ISWC'11)*, volume 7032 of *LNCS*. Springer, 2011.

S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink *et al.* Intact–open source resource for molecular interaction data. *Nucleic Acids Research*, 35 (Database issue):D561–5, 2007.

R. N. W. Kettleborough, E. de Bruijn, F. van Eeden, E. Cuppen, and D. L. Stemple. High-throughput target-selected gene inactivation in zebrafish. *Methods Cell Biol*, 104:121–7, 2011.

S. Köhler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4): 949–58, 2008.

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, 2009.

S. Köhler, S. Bauer, C. J. Mungall, G. Carletti, C. L. Smith, P. Schofield, G. V. Gkoutos, and P. N. Robinson. Improving ontologies by automatic reasoning and evaluation of logical definitions. *BMC Bioinformatics*, 12 (1):418, 2011.

S. Köhler, S. C. Doelken, A. Rath, S. Aymé, and P. N. Robinson. Ontological phenotype standards for neurogenetics. *Human mutation*, 33(9):1333–1339, 2012.

P. M. Krawitz, M. R. Schweiger, C. Rödelsperger, C. Marcelis, U. Kölsch, C. Meisel, F. Stephani, T. Kinoshita, Y. Murakami, S. Bauer, M. Isau, A. Fischer, A. Dahl, M. Kerick, J. Hecht *et al.* Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet*, 42(10):827–9, 2010.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

Y. Li and J. C. Patra. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics (Oxford, England)*, 26(9):1219–24, 2010.

G. J. Lieschke and P. D. Currie. Animal models of human disease: zebrafish swim into view. *Nat Rev Genet*, 8(5):353–67, 2007.

J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J.-F. Rual, C. J. Fisk, N. Li, A. Smolyar, D. E. Hill, A.-L. Barabási, M. Vidal, and H. Y. Zoghbi. A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125(4):801–14, 2006.

F. Loebe, F. Stumpf, R. Hoehndorf, and H. Herre. Towards improving phenotype representation in OWL. *J Biomed Semantics*, 3 Suppl 2(3):S5, 2012.

Y. A. Lussier and Y. Liu. Computational approaches to phenotyping: high-throughput phenomics. *Proc Am Thorac Soc*, 4(1):18–25, 2007.

P. M. Mabee, M. Ashburner, Q. Cronk, G. V. Gkoutos, M. A. Haendel, E. Segerdell, C. J. Mungall, and M. Westerfield. Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol (Amst)*, 22 (7):345–50, 2007.

M. Mahner and M. Kary. What exactly are genomes, genotypes and phenotypes? And what about phenomes? *J Theor Biol*, 186(1):55–63, 1997.

M. Masseroli, O. Galati, M. Manzotti, K. Gibert, and F. Pinciroli. Inherited disorder phenotypes: controlled annotation and statistical analysis for knowledge mining from gene lists. *BMC Bioinformatics*, 6 Suppl 4:S18, 2005.

T. F. Meehan, A. M. Masci, A. Abdulla, L. G. Cowell, J. A. Blake, C. J. Mungall, and A. D. Diehl. Logical development of the cell ontology. *BMC Bioinformatics*, 12:6, 2011.

G. J. Mendel. Versuche über Pflanzenhybriden. *Verhandlungen des Naturforschenden Vereines in Brünn*, 4:3–47, 1866.

G. R. Merlo, L. Paleari, S. Mantero, F. Genova, A. Beverdam, G. L. Palmisano, O. Barbieri, and G. Levi. Mouse model of split hand/foot malformation type i. *Genesis*, 33(2):97–101, 2002.

B. Motik, R. Shearer, and I. Horrocks. Optimized reasoning in description logics using hypertableaux. In *Proc. of CADE-21*, volume 4603 of *LNCS (LNAI)*, pages 67–83, Heidelberg, 2007. Springer.

B. Motik, P. F. Patel-Schneider, and B. Parsia. OWL 2 Web Ontology Language: structural specification and functional-syle syntax. `http://www.w3.org/TR/owl2-syntax/`, 2008.

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, 2010.

C. J. Mungall, M. Bada, T. Z. Berardini, J. Deegan, A. Ireland, M. A. Harris, D. P. Hill, and J. Lomax. Cross-product extensions of the gene ontology. *J Biomed Inform*, 44(1):80–6, 2011.

C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13(1):R5, 2012.

D. A. Natale, C. N. Arighi, W. C. Barker, J. A. Blake, C. J. Bult, M. Caudy, H. J. Drabkin, P. D'Eustachio, A. V. Evsikov, H. Huang, J. Nchoutmboube, N. V. Roberts, B. Smith, J. Zhang, and C. H. Wu. The protein ontology: a structured representation of protein forms and complexes. *Nucleic Acids Research*, 39(Database issue):D539–45, 2011.

S. Navlakha and C. Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics (Oxford, England)*, 26 (8):1057–63, 2010.

K. P. O'Brien, M. Remm, and E. L. L. Sonnhammer. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, 33 (Database issue):D476–80, 2005.

M. C. O'Donovan, G. Kirov, and M. J. Owen. Phenotypic variations on the theme of CNVs. *Nat Genet*, 40(12):1392–3, 2008.

S. Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–3, 2000.

G. Ostlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, O. Frings, and E. L. L. Sonnhammer. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Research*, 38 (Database issue):D196–203, 2010.

M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–8, 2006.

L. Page, S. Brin, M. Rajeev, and W. Terry. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

J. Pelz, V. Arendt, and J. Kunze. Computer assisted diagnosis of malformation syndromes: an evaluation of three databases (lddb, possum, and syndroc). *Am J Med Genet*, 63(1):257–67, 1996.

E. Pennisi. Breakthrough of the year. human genetic variation. *Science*, 318 (5858):1842–3, 2007.

S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. K. B. Gandhi, K. N. Chandrika, N. Deshpande, S. Suresh, B. P. Rashmi, K. Shanker, N. Padma, V. Niranjan, H. C. Harsha *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Research*, 32(Database issue):D497–501, 2004.

C. Pesquita, D. Faria, A. O. Falcão, P. Lord, F. M. Couto, and P. E. Bourne. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 5(7), 2009.

B. R. Pober. Williams-beuren syndrome. *N Engl J Med*, 362(3):239–52, 2010.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *Arxiv preprint cmp-lg*, 1995.

E. R. Riggs, L. Jackson, D. T. Miller, and S. V. Vooren. Phenotypic information in genomic variant databases enhances clinical care and research: The international standards for cytogenomic arrays consortium experience. *Human mutation*, 33(5):787–96, 2012.

J. D. L. Rivas and C. Fontanillo. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, 2010.

P. N. Robinson. Deep phenotyping for precision medicine. *Human mutation*, 33(5):777–80, 2012.

P. N. Robinson and S. Bauer. *Introduction to Bio-Ontologies*. CRC Press Inc, 2011.

P. N. Robinson and S. Mundlos. The human phenotype ontology. *Clinical Genetics*, 77(6):525–534, 2010.

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, 2008.

R. F. Robledo, L. Rajan, X. Li, and T. Lufkin. The dlx5 and dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. *Genes Dev*, 16(9):1089–101, 2002.

C. Rödelsperger, S. Köhler, M. H. Schulz, T. Manke, S. Bauer, and P. N. Robinson. Short ultraconserved promoter regions delineate a class of preferentially expressed alternatively spliced transcripts. *Genomics*, 94 (5):308–16, 2009.

C. Rödelsperger, G. Guo, M. Kolanczyk, A. Pletschacher, S. Köhler, S. Bauer, M. H. Schulz, and P. N. Robinson. Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Research*, 39(7):2492–502, 2011.

N. Rosenthal and S. Brown. The mouse ascending: perspectives for human-disease models. *Nat Cell Biol*, 9(9):993–9, 2007.

C. Rosse and J. L. V. Mejino. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform*, 36(6):478–500, 2003.

L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database issue):D449–51, 2004.

D. Sardana, S. Vasa, N. Vepachedu, J. Chen, R. C. Gudivada, B. J. Aronow, and A. G. Jegga. Phenohm: human-mouse comparative phenome-genome server. *Nucleic Acids Research*, 38(Web Server issue):W165–74, 2010.

G. Schindelman, J. S. Fernandes, C. A. Bastiani, K. Yook, and P. W. Sternberg. Worm phenotype ontology: integrating phenotype data within and beyond the c. elegans community. *BMC Bioinformatics*, 12:32, 2011.

P. N. Schofield and J. M. Hancock. Integration of global resources for human genetic variation and disease. *Human mutation*, 33(5):813–6, 2012.

P. N. Schofield, G. V. Gkoutos, M. Gruenberger, J. P. Sundberg, and J. M. Hancock. Phenotype ontologies for mouse and man: bridging the semantic gap. *Dis Model Mech*, 3(5-6):281–9, 2010.

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*, 10(5): 258–65, 2011.

P. N. Schofield, R. Hoehndorf, and G. V. Gkoutos. Mouse genetic and phenotypic resources for human genetics. *Human mutation*, 33(5):826–36, 2012.

M. H. Schulz, S. Köhler, S. Bauer, M. Vingron, and P. N. Robinson. Exact score distribution computation for similarity searches in ontologies. *Algorithms in Bioinformatics*, pages 298–309, 2009.

M. H. Schulz, S. Köhler, S. Bauer, and P. N. Robinson. Exact score distribution computation for ontological similarity searches. *BMC Bioinformatics*, 12(1):441, 2011.

S. Schulze-Kremer. Adding semantics to genome databases: towards an ontology for molecular biology. *Proc Int Conf Intell Syst Mol Biol*, 5:272–5, 1997.

B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part i. experimental techniques and databases. *PLoS Comput Biol*, 3(3):e42, 2007.

E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics*, 5(2):51–53, 2007.

B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, O. B. I. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone *et al.* The obo foundry:

coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, 2007.

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, 2005.

T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981.

M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–2, 2011.

L. N. Soldatova and R. D. King. Are the current ontologies in biology good ontologies? *Nat Biotechnol*, 23(9):1095–8, 2005.

J. Sprague, L. Bayraktaroglu, Y. Bradford, T. Conlin, N. Dunn, D. Fashena, K. Frazer, M. Haendel, D. G. Howe, J. Knight, P. Mani, S. A. T. Moxon, C. Pich, S. Ramachandran, K. Schaper *et al.* The zebrafish information network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Research*, 36(Database issue):D768–72, 2008.

J. Spranger. Pattern recognition in bone dysplasias. *Prog Clin Biol Res*, 200: 315–42, 1985.

C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. V. Auken, X. Wang, X. Shi, T. Reguly, J. M. Rust, A. Winter, K. Dolinski, and M. Tyers. The biogrid interaction database: 2011 update. *Nucleic Acids Research*, 39(Database issue):D698–704, 2011.

U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122 (6):957–68, 2005.

M. Strickberger. *Genetics*. Macmillan, 1985. ISBN 9780024181206.

D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(Database issue):D561–8, 2011.

A. Tarski and J. Corcoran. *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Hackett Publishing Company, 1983. ISBN 9780915144761.

H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 613–622, Washington, DC, USA, 2006. IEEE Computer Society.

D. Tsarkov and I. Horrocks. FaCT++ description logic reasoner: system description. In *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer, 2006.

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, 2006.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman *et al.* The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

D. Wang, L.-E. Jao, N. Zheng, K. Dolan, J. Ivey, S. Zonies, X. Wu, K. Wu, H. Yang, Q. Meng, Z. Zhu, B. Zhang, S. Lin, and S. M. Burgess. Efficient genome-wide mutagenesis of zebrafish genes by retroviral insertions. *Proc Natl Acad Sci USA*, 104(30):12428–33, 2007.

J. Wang, X. Zhou, J. Zhu, C. Zhou, and Z. Guo. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC Bioinformatics*, 11:290, 2010.

G. Warsow, B. Greber, S. S. I. Falk, C. Harder, M. Siatkowski, S. Schordan, A. Som, N. Endlich, H. Schöler, D. Repsilber, K. Endlich, and G. Fuellen. Expressence–revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst Biol*, 4:164, 2010.

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, 2009.

J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.

J. Xu and Y. Li. Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics*, 22(22):2800–5, 2006.

Appendix A

Glossary

Asserted axiom

An axiom that was manually asserted by a curator.

Axiom

In the context of KR used to refer to statements that say what is true in the domain (see `http://www.w3.org/TR/owl2-syntax/#Axioms`).

CNV

Genomic variation in that a segment of DNA has an abnormal number of copies. The segment can be deleted, inverted (turned around), or added (duplicated) in an individuals genome.

Diagnosis

Assigning the patient's conditions to a classification, which can then give insight into possible ways of treatment or information on prognoses [Pelz et al., 1996].

Disease gene family

A concept for grouping clinically distinct diseases into small groups of diseases that share important phenotypic features. Pathogenetic similarities are assumed to underlie the phenotypic commonalities.

DNA

Encodes the heritable information.

EQ-approach

Description schema in which a phenotype can be abstracted into two parts. First, an entity that is affected (e.g. an enzyme, an anatomical structure) and the quality of that entity. In the typical setting, a phenotype is described using a class expression consisting of a PATO quality class differentiated by a bearer entity class (a term from an OBO ontology) using the `inheres_in` relation [Hancock et al., 2009].

Gene

A union of genomic sequences encoding a coherent set of potentially overlapping functional products [Gerstein et al., 2007].

Genome

The complete set of genetic material of an organism.

Genotype

The complete constitution or makeup of the genetic material belonging to a cell or an individual.

Graph

A graph $G = (V, E)$ is defined as a mathematical structure that consists of two finite sets V and E. The elements of V are referred to as vertices or nodes. The elements of E are called the edges or arcs.

Inferred axiom

An axiom that was deduced by automated reasoning (based on asserted axioms).

Interactome

The complete set of interactions and interactants of cellular networks.

Mendelian disease

Phenotypes that are mostly determined by a mutation (or mutations) in a single gene and that follow a dominant, recessive, or X-linked inheritance pattern.

Mutation

Alteration within the DNA.

Ontology

An ontology is an *explicit*, *formal* specification of a *shared conceptualization* [Gruber, 1993].

Phenotype

The manifold biological appearances, including chemical, structural and behavioral attributes, that we can observe about an organism but excludes its genetic constitution [Strickberger, 1985].

Pleiotropy

Condition in which a gene affects more than one phenotype.

Reasoning

Inference of implicitly represented knowledge from the knowledge that is explicitly contained (asserted) in the knowledge base [Baader et al., 2003].

RNA

Polymeric, single stranded nucleic acid, involved as intermediate for transmission of the DNA information to the protein level. Also involved in the control of several chemical processes.

SNP

Change in the DNA of one single base pair (nucleotide).

Appendix B

Acronyms

AUROC	Area under the ROC curve.
BMA	Best-match average (used for semantic similarity).
ChEBI	Chemical entities of biological interest ontology.
CNV	Copy number variation.
DECIPHER	Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources.
DI	Direct interaction (method for disease gene prioritisation).
DNA	Deoxyribonucleic acid.
DR	Danio rerio (Zebrafish).
EQ	Entity/Quality approach.
FMA	The Foundational Model of Anatomy ontology.
FPR	False-positive rate.

GNF	Genomics Institute of the Novartis Research Foundation.
GO	The Gene Ontology.
HPO	The Human Phenotype Ontology.
HS	Homo sapiens (Human).
IC	Information content (used for semantic similarity).
IKMC	International Knockout Mouse Consortium.
IMPC	International Mouse Phenotyping Consortium.
KB	Knowledge Base.
KR	Knowledge Representation.
LDDB	London Dysmorphology Database.
MA	Mouse adult gross anatomy ontology.
MGI	Mouse Genome Informatics.
MM	Mus musculus (House mouse).
MPO	The Mammalian Phenotype Ontology, also MP.
NCBI	National Center for Biotechnology Information.
OMIM	Online Mendelian Inheritance in Man.
OWL	Web Ontology Language.
PATO	Phenotype, Attribute and Trait Ontology (Currently Phenotypic Quality Ontology).
POSSUM	Pictures of Standard Syndromes and Undiagnosed Malformations.
PPIN	Protein-protein interaction network.
PRO	Protein Ontology.
RDF	Resource Description Framework.

RDFS	RDF Schema.
RNA	Ribonucleic acid.
ROC	Receiver-operating characteristic.
RWR	Random Walk with restart.
SNP	Single nucleotide polymorphism (pronounced as ′snip′).
SP	Shortest path (method for disease gene prioritisation).
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins.
TPR	True-positive rate.
UMLS	Unified Medical Language System.
XML	Extensible Markup Language.
ZFA	Zebrafish anatomy and development ontology.
ZFIN	Zebrafish Information Network (Now: The Zebrafish Model Organism Database).

Appendix C

Abstract

Understanding the relationships between human phenotypic abnormalities and their underlying genes is an important subject in biomedicine. Comprehensive data sets on interactions between gene products enable novel systems approaches to be applied for elucidating those associations. Recently, neighborhood approaches, analysing the local shortest-path distances between network nodes, have been applied to the problem of disease gene prediction. Here it is shown that a global network-similarity measure based on random walks, is well suited for analysing vicinity in protein-protein interaction networks, and that this boosts the performance of guilt-by-association approaches for gene-to-phenotype research.

Analysing disease information has long been hampered by the lack of standards and semantics in knowledge representation on associated phenotypic abnormalities. Often, phenotype descriptions were stored as a part of free text, making automated mining very difficult. This work presents the Human Phenotype Ontology (HPO) and its application to disease similarity calculation based on semantic similarity between phenotypic spectra. A tool is presented that uses the HPO to aid with clinical diagnostics in medical genetics and makes use of a novel statistical model assigning P-values to semantic similarity scores.

Motivated by the aim of revealing genotype-to-phenotype associations directly, high-throughput projects are now exploring complete phenomes of model organisms such as mouse. Especially the transfer of knowledge across species is important to understand the relations between genotypes and phenotypes in model organisms on the one hand, and those seen in human diseases on the other hand. Cross-species phenotype analysis is of major importance, given that there are currently more than 5,000 human genes with no phenotype information, but for which detailed phenotypes are available for their mouse and/or zebrafish orthologs.

In this work, the development of cross-species, harmonised, semantic representation of phenotype information is presented. A computational framework is developed to comprehensively harvest phenotypic information from model organisms and single-gene human hereditary disorders annotated to HPO terms. It is exemplified how this can speed the interpretation of the complex phenotypes of CNV disorders, and how this ontology-based approach is used to identify similarities between human phenotypes and the mutational phenotypes in known model organism genes.

Using this approach, phenotypic multiplicities are identified as common characteristic of CNVs, in which multiple genes are said to influence a particular phenotypic feature of a broader spectrum of phenotypes. Often, the association between the genes of these multiplicities represent novel hypotheses and are supported by correlation with Gene Ontology similarity and random walk vicinity in protein interaction networks.

Appendix D

Zusammenfassung

Das Verständnis der Beziehungen zwischen menschlichen Phänotypen und den zugrunde liegenden Genen ist ein zentrales Thema der modernen Biomedizin. Umfangreiche Datensätze über Interaktionen zwischen Gen-Produkten ermöglichen neue Ansätze zur Untersuchung dieser Beziehungen, indem die Zellen als komplexe Systeme bzw. Netzwerke betrachtet werden. Bisherige Ansätze betrachten dabei die lokale Nachbarschaft zu identifizierten Krankheitsgenen mittels Berechnung der kürzesten Pfade, um Krankheitsgene vorherzusagen. In dieser Arbeit wird hingegen gezeigt, dass Maße, die die gesamte Netzwerkstruktur einbeziehen, sehr gut für diese Problemstellung geeignet sind. Dafür wird eine Methode basierend auf dem Random-Walk vorgestellt und verglichen.

Die Analyse von Krankheiten und deren Symptomen wird seit langem durch das Fehlen von Standards behindert. Phänotypische Beschreibungen wurden bisher lediglich in Textform abgelegt, wodurch automatisierte computerbasierte Analysen behindert werden. Diese Arbeit stellt die Human Phenotype Ontology (HPO) vor und beschreibt deren Anwendung auf Ähnlichkeitsberechnungen zwischen Krankheiten auf Basis von semantischer Ähnlichkeit zwischen den phänotypischen Spektren der Krankheitsbilder. Des Weiteren wird ein Programm vorgestellt, welches mit Hilfe der HPO das Finden von klinischen Diagnosen in der Humangenetik unterstützt. Dieses wurde mit einem neuen statistischen Modell unterlegt, welches die Zuweisung von *P*-Werten für semantische Ähnlichkeiten er-

laubt.

Motiviert durch das Ziel, Genotyp-zu-Phänotyp Assoziationen direkt zu finden, sollen neue Hochdurchsatz-Projekte nach und nach alle Gene von Modellorganismen ausschalten und die phänotypischen Konsequenzen aufzeichnen. Vor allem die speziesübergreifende Übertragung von solchen phänotypischen Daten ist wichtig, um die Beziehungen zwischen den Genotypen und Phänotypen beim Menschen besser zu verstehen. Speziesübergreifende Phänotyp-Analysen sind von enormer Bedeutung, da es derzeit mehr als 5.000 menschliche Gene ohne Phänotyp-Informationen gibt, für die es allerdings detaillierte Phänotypen-Information für die orthologen Gene in Maus und/oder Zebrafisch gibt.

In dieser Arbeit wird die Entwicklung einer speziesübergreifenden, harmonisierten, semantischen Repräsentation von phänotypischen Abnormalitäten vorgestellt. Es wird beschrieben, wie man dadurch systematisch phänotypische Informationen von Modellorganismen und menschlichen monogenen Krankheiten integrieren kann. Weiterhin wird gezeigt, wie dies genutzt werden kann, um komplexe Phänotypen in Krankheiten, die durch Copy-Number-Variations (CNV) ausgelöst werden, einzelnen betroffenen Genen zuzuordnen.

Mit diesem Ansatz wurden phänotypische „Vervielfachungen“ als Charakteristikum von CNVs gefunden, bei denen mehrere Gene ein bestimmtes phänotypisches Merkmal beeinflussen. Häufig stellen diese „Vervielfachungen“ neuartige Hypothesen für die gemeinsame Funktion der beteiligten Gene dar. Diese Hypothesen werden durch Korrelation mit Gene Ontology-Ähnlichkeit unterstützt und zeigen eine statistisch signifikante Nähe im Protein-Interaktions-Netzwerk, welche durch die Random-Walk-Methode ermittelt wurde.

Appendix E

Acknowledgments

First and foremost I want to thank my supervisor Peter N Robinson for supporting me during the last 5 years, sharing several dozens of ideas and providing me with the freedom to plan and realise these projects on my own. I also would like to thank him for giving me the opportunity to visit Lisbon during a four-month period in 2010 and work in the group of Francisco M Couto.

I also want to thank my colleague Sebastian Bauer for all his guidance, ideas, and constructive criticism during my PhD. He has basically been my most important project advisor throughout long periods in the last years. My knowledge on algorithm development, optimisation, statistics, and software development grew enormously throught those years.

I also want to thank Prof. Martin Vingron for reviewing my thesis and all the constructive criticism and ideas for this thesis.

For very, very extensive proofreading I want to thank Barbara Ruef from ZFIN. I am also thankful for the positive feedback and comments by Chris J Mungall and Nicole Washington from LBNL.

I want to thank Francisco Couto for giving me the opportunity of working in his group, see a different working environment, and meet a lot of nice people at his group.

I am thankful to have been able to work with a lot of international collaborators, namely Marcel Schulz, Damian Smedley, Monte Westerfield, Paul Schofield, George Gkoutos, Michael Ashburner, Robert Hoehndorf, Suzanna E Lewis, John Hancock, Cynthia Smith, Ana Rath, and Ségolène Aymé.

There are a lot of people at the Institute of Medical Genetics in Berlin, that I want to thank as well. Especially Sandra Doelken, Marten Jäger, Claus-Eric Ott, Verena Heinrich, Christian Rödelsperger, Dominik Seelow, Peter Krawitz, Denise Horn, Christine Mundlos, and Stefan Mundlos.

I also want to thank the scientist and graphics designer Lude Franke, whose work always inspired me to put some amount of extra work into graphics. The gentle reader may notice that it was his PhD Thesis, that inspired the color choice of the figures in my thesis.

I want to thank my parents for supporting throughout the last years. I think in total you did a great job in parenting. I want to thank my grandparents, since without you I wouldn't have had the possibilities that I had now. (I will never forget you Helga and Gerhard.)

Last but not least, I want to thank my girlfriend Maria for all those years of support and encouragement during the hard times. Thanks for listening and trying to let me look at things in a different light when times were really difficult.

Appendix F

Software Availability

In chapter 2, a new method to prioritise genes that are a involved in particular human phenotyes is described. In order to make this easily accessible for the community, the *GeneWanderer* was implemented as a webtool [Köhler et al., 2008]. The homepage is located at `http://compbio.charite.de/genewanderer` and allows user-defined genomic intervals to be analysed by different methods and data sets.

A new strategy to represent abnormal human phenotypes in described in chapter 3. The *Human Phenotype Ontology* (HPO) itself, as well as tools for exploration and navigation of the ontology, can be found on the website `http://www.human-phenotype-ontology.org`. This chapter additionally describes new ways of applying the HPO in the setting of clinical genetics. The *PhenExplorer*, a tool to explore the HPO, disease genes, and diseases, is available at `http://compbio.charite.de/phenexplorer`. The software that implements semantic similarity searches in the HPO in order to support clinical diagnostics is called *The Phenomizer*. It is freely available at `http://compbio.charite.de/phenomizer`.

Semantic integration of phenotype data across domain and species is the topic of chapter 4. Therein, a novel method for consistency and completeness of inspections of ontologies is depicted. This tool is also used for constructing the *Uberpheno* ontology. The software called *GULO* is available at `https://compbio.charite.de/svn/hpo/trunk/src/tools/gulo` as an executable jar-file.

Appendix G

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, September 2012