Ontology-based Similarity Measures and their Application in Bioinformatics

DISSERTATION

ZUR ERLANGUNG DES GRADES DES
DOKTORS DER NATURWISSENSCHAFFTEN DER
NATURWISSENSCHAFTLICH-TECHNISCHEN FAKULTÄTEN DER
UNIVERSITÄT DES SAARLANDES

eingereicht von
ANDREAS SCHLICKER

Saarbruecken, Mai 2010

Tag des Kolloquiums:	02.11.2010
Dekan der Fakultät:	Prof. Dr. Holger Hermanns
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Gerhard Weikum
Gutachter:	Prof. Dr. Dr. Thomas Lengauer Dr. Mario Albrecht Dr. Marie-Dominique Devignes
Beisitzer:	Dr. Francisco S. Domingues

Abstract

Genome-wide sequencing projects of many different organisms produce large numbers of sequences that are functionally characterized using experimental and bioinformatics methods. Following the development of the first bio-ontologies, knowledge of the functions of genes and proteins is increasingly made available in a standardized format. This allows for devising approaches that directly exploit functional information using semantic and functional similarity measures. This thesis addresses different aspects of the development and application of such similarity measures.

First, we analyze semantic and functional similarity measures and apply them for investigating the functional space in different taxa. Second, a new software program and a new database are described, which overcome limitations of existing tools and simplify the utilization of similarity measures for different applications.

Third, we delineate two applications of our functional similarity measures. We utilize them for analyzing domain and protein interaction datasets and derive thresholds for grouping predicted domain interactions into low- and high-confidence subsets. We also present the new MedSim method for prioritization of candidate disease genes, which is based on the observation that genes and proteins contributing to similar diseases are functionally related. We demonstrate that the MedSim method performs at least as well as more complex state-of-the-art methods and significantly outperforms current methods that also utilize functional annotation.

Kurzfassung

Die Sequenzierung der kompletten Genome vieler verschiedener Organismen liefert eine große Anzahl an Sequenzen, die mit Hilfe experimenteller und bioinformatischer Methoden funktionell charakterisiert werden. Nach der Entwicklung der ersten Bio-Ontologien wird das Wissen über die Funktionen von Genen und Proteinen zunehmend in einem standardisierten Format zur Verfügung gestellt. Dadurch wird die Entwicklung von Verfahren ermöglicht, die funktionelle Informationen direkt mit Hilfe semantischer und funktioneller Ähnlichkeit verwenden. Diese Doktorarbeit befasst sich mit verschiedenen Aspekten der Entwicklung und Anwendung solcher Ähnlichkeitsmaße.

Zuerst analysieren wir semantische und funktionelle Ähnlichkeitsmaße und verwenden sie für eine Analyse des funktionellen Raumes verschiedener Organismengruppen. Danach beschreiben wir eine neue Software und eine neue Datenbank, die Limitationen existierender Programme überwinden und den Einsatz von Ähnlichkeitsmaßen in verschiedenen Anwendungen vereinfachen.

Drittens schildern wir zwei Anwendungen unserer funktionellen Ähnlichkeitsmaße. Wir verwenden sie zur Analyse von Domän- und Proteininteraktionsdatensätzen und leiten Grenzwerte ab, um die Domäninteraktionen in Teilmengen mit niedriger und hoher Konfidenz einzuteilen. Außerdem präsentieren wir die MedSim-Methode zur Priorisierung von potentiellen Krankheitsgenen. Sie beruht auf der Beobachtung, dass Gene und Proteine, die zu ähnlichen Krankheiten beitragen, funktionell verwandt sind. Wir zeigen, dass die MedSim-Methode mindestens so gut funktioniert wie komplexere moderne Methoden und die Leistung anderer aktueller Methoden signifikant übertrifft, die auch funktionelle Annotationen verwenden.

Acknowledgements

First, I would like to thank my supervisor Thomas Lengauer for his invaluable support and advice during my PhD studies. I am grateful that he provided me with the possibility to pursue my own ideas in such a great environment. I also want to thank Mario Albrecht for guiding me during this time, for his help and advice on all problems. I am also grateful to Marie-Dominique Devignes for her willingness to act as an additional reviewer of this thesis.

I especially want to thank my office mates Dorothea Emig and Fidel Ramírez and all other members of our team, Hagen Blankenburg, Sven-Eric Schelhorn, Gabriele Mayr, Sarah Diehl, and Christoph Welsch, for the discussions, their input, and the good time. Thanks go also to Alejandro Pironti with whom I worked on issues regarding the prioritization of candidate disease genes. I am also grateful to Franisco Domingues and Ingolf Sommer for critically reading manuscripts and providing suggestions. I would also like to thank all other people on floor five of the MPII I didn't mention yet for the joy of working with them and all the time we spent together. Thanks also to Susanne Eyrisch for her suggestions on this thesis.

I would especially like to acknowledge Achim Büch for his help with all technical problems and his support in setting up databases and web services. Special thanks go also to Ruth Schneppen-Christmann for her support with all administrative issues and for organizing numerous trips to conferences.

Finally, I would like to thank my family and especially my mother for always believing in me and encouraging me. I am grateful to Jasmin for lighting up my life. Last but not least, I would like to thank all my friends, without whom I wouldn't be here today, for knowingly and unknowingly supporting me whenever needed, and the countless hours of fun.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Until the 1990s, molecular-biological research has mostly focused on studying specific molecules at a time. Single genes and proteins were analyzed in detail to determine their sequence, structure, or functions. One central assumption for this research was the "one-gene-one-protein" hypothesis. It states that each gene is transcribed into an unique protein, which has one specific function. In recent years, this central hypothesis has been challenged more and more by the biological discoveries made using large-scale analyses.

The sequencing of whole genomes marked the beginning of high-throughput research in biology. Since the first complete genome of a bacterium was made available in 1995 (*Haemophilus influenzae*, Fleischmann *et al.* 1995), the rate with which data are produced has increased steadily. International research consortia like the 1000 Genomes Project or the Cancer Genome Project have started to produce several terabytes of new sequences per year. In addition, experimental techniques were developed to collect genome- and proteome-wide datasets that complement the genomic sequence. Microarrays and increasingly deep sequencing are utilized for monitoring the expression of all genes in a cell. The complex networks of interactions between all proteins of an organism are probed by several experimental high-throughput techniques. These are only two examples for a variety of large-scale datasets that are being generated, and that have to be integrated in modern biological and medical research.

The introduction of high-throughput methods also led to an increasingly interdisciplinary character of biomedical research. Fields that developed independently in the past are now combined to increase knowledge, for example in genetics. Genome-wide association studies have shown that inherited diseases are often influenced by alterations in a variety of different genes each of which contributes only little to the overall disease risk. This makes it necessary to combine different fields of genetics research for determining the genetic influence of complex diseases.

The transformation of biology from a science of small datasets to a science dealing

with large, complex datasets creates two major challenges. First, it is important to develop new methods for efficiently storing and processing these data. Genome-wide datasets cannot be handled manually, and computational approaches are necessary for their analysis. Second, the available datasets become increasingly complex and multidimensional. Therefore, new approaches for combining disparate information need to be devised for gaining insights into biological processes and phenomena.

Bioinformatics plays an important role in providing solutions for both questions. Research in bioinformatics aims at the development of algorithms and software programs that allow for handling genome-wide, interconnected datasets. Computational methods can test biological hypotheses by analyzing datasets, but they can also be applied to generate hypotheses that are amenable to experimental verification.

Ontologies are one particular example of a technique that has been recently introduced into biomedical research. Ontologies contain controlled vocabularies that are generally organized in a tree-like structure. This way, they also define the relationships between different terms. Ontologies facilitate representing knowledge in a specific area in a standardized, automatically accessible form. Their development and application to biological datasets requires the close collaboration of experts from different fields, and the use of ontologies opens several new possibilities. The utilization of a controlled vocabulary allows for developing computational methods that integrate and analyze disparate biological datasets. Furthermore, they have the potential to standardize the use of terminology across different research areas, which is a prerequisite for interdisciplinary research.

Today, ontologies are quickly gaining importance in computer science and biomedical research. One particular research area that requires their widespread adoption is the development of the semantic web, which aims at improving the interoperability of data repositories by adding semantic annotation. In the biomedical domain, the ontologies provided by the Gene Ontology Consortium are of major importance. They are widely used to annotate genes and gene products with functional information. The Reference Genome Annotation Project of the Gene Ontology Consortium, for instance, strives to provide functional annotation with ontology terms for the complete genomes of twelve organisms including human, mouse, yeast, and fruit fly. This functional annotation is utilized in many important applications, for example, analysis of gene expression data, prediction and validation of molecular interactions, and prioritization of disease gene candidates.

The extensive utilization of ontologies creates several tasks. First, similarity measures for comparing ontological annotation need to be devised. Second, new software programs for calculating similarity scores are required to efficiently cope with the growing amount of ontological annotations. Third, new methods need to be developed that exploit these annotations for automatically generating new testable hypotheses. In this thesis, we describe methods that deliver solutions for all these tasks.

1.2 Overview

This work focuses on the development of computational approaches for the application of ontologies in biomedical research. We devised new methods for integrating semantic similarity values for different ontologies into a single score. The developed software programs are targeted at bioinformaticians as well as biological and medical users. They were designed to be easy to use but also to be versatile and allow for their flexible application.

Utilizing these tools, we employed functional similarity measures in several new applications. First, we analyzed predicted domain-domain interactions and derived thresholds to classify them into sets of varying confidence. Second, we developed a new method for prioritizing candidate disease genes and proteins. This novel method improves on the performance of existing methods and the prediction model allows for easily interpreting the results.

This thesis is based on nine publications in reputable bioinformatics journals and conference proceedings. A list of all papers is given in the Appendix. The work was financially supported by the German Research Foundation (DFG) for the clinical research group KFO 129, the Federal Ministry of Education and Research (BMBF) for projects within the German National Genome Research Network (NGFN), and the European Commission for the BioSapiens Network of Excellence for genome annotation.

1.3 Outline

The remainder of this thesis is structured into six chapters followed by the bibliography and an appendix. Chapter 2 gives a comprehensive overview on ontologies in general and in the biomedical domain. Furthermore, we provide a review of important semantic and functional similarity methods.

In Chapter 3, we outline the extensive evaluation of our semantic and functional similarity measures and compare them to previous work. Further, a comparison highlights differences and commonalities in the molecular functions and biological processes from different taxa. Using functional similarity, we derive maps of the functional space of protein domains and yeast proteins.

Chapter 4 introduces the Functional Similarity Search Tool (FSST) and the Functional Similarity Matrix (FunSimMat). FSST is a stand-alone tool for calculating functional similarity values between annotated entities. It also allows for incorporating private ontological annotations into the comparison. FunSimMat is the first comprehensive database of precomputed semantic and functional similarity values.

Chapter 5 describes the application of functional similarity measures for assessing molecular interaction data. Using the similarity measures, different sets of predicted domain-domain interactions are classified according to their confidence. Additionally, human protein interaction datasets are analyzed using functional measures.

Chapter 6 deals with the problem of finding genes and proteins that are associated with diseases. Human inherited genetic diseases are often caused by different functionally related genes. Here, we outline the MedSim method that automatically builds functional profiles of disease phenotypes with terms from the Gene Ontology. These profiles are applied for improving the prioritization of candidate disease genes for further experimental validation. Furthermore, we show the application of the functional profiles for performing a functional comparison between different disease phenotypes.

Chapter 7 draws conclusions on the conducted research and summarizes the main achievements. Additionally, possible improvements and methodological perspectives are discussed.

In the Appendix, we provide a complete list of disease phenotypes and proteins used for validation in Chapter 6 as well as a list of publications that describe work related to this thesis.

Chapter 2

Ontologies and Similarity Computation

The field of Ontology arose historically as a branch of philosophical research and was introduced into computer science in the last 20 years as the theory of entities and their relationships in a specific domain. The comprehensive lists of all entities and their relationships pertaining to a single domain are called ontologies. Recently, such ontologies have quickly gained importance in biological and medical research as an instrument for knowledge representation and integration. They are increasingly utilized as a means of annotating database entries with additional information and of integrating different data sources.

The first section of this chapter briefly introduces the philosophical origin of ontologies and their application in computer science and biomedical research. The second section outlines methods for measuring the semantic similarity between terms in an ontology. Finally, methods are described for quantifying the functional similarity of entities that are annotated with terms from a single ontology and for integrating similarity values from different ontologies. We will restrict the description of similarity measures to approaches that were applied to annotation with Gene Ontology (GO) terms.

2.1 Ontology

2.1.1 The Study of the Nature of Existence

There are many different notions of *ontology* as a philosophical discipline. Gottfried Wilhelm Leibniz, for instance, defined ontology as the science of all things that do and do not exist, and their modes (Leibniz, 1988).

Historically, ontology evolved as the branch of philosophy that is concerned with the study of objects, their properties and relationships (Smith, 2003). Its foundations were laid by Aristotle, who referred to the field as *first philosophy*. Later, his students used the name *metaphysics*, which is still widely used as a synonym for ontology in

philosophy. In 1613, the term *ontology* was independently introduced by Rudolph Göckel and Jacob Lorhard (Corazzon, 2009). One of the main goals of the research field ontology has always been to provide a definitive and exhaustive classification of all things, which should help answer questions like: "What classes of entities are needed for a complete description of all things in the universe?" (Smith, 2003). In particular, this includes the question which things exist or are conceivable to exist, and how these objects can be grouped in a hierarchy by respecting the relationships between them.

The lists of entities and their relationships in a specific domain of interest are called ontologies. The terms taxonomy and ontology are very often used interchangeably (Pidcock, 2010). More strictly defined, a taxonomy is a controlled vocabulary of terms that are hierarchically structured. An ontology additionally defines properties of the objects and imposes constraints on how the terms in the vocabulary can be used in a meaningful way (Pidcock, 2010). From a philosophical point of view, one can distinguish between four types of concepts in an ontology: universals, particulars, continuants, and occurrents (Smith *et al.*, 2003). Particulars, on the one hand, are concrete instances of universals, for example, an individual cell as opposed to cells in general. Orthogonal to this distinction is the difference between continuants and occurrents. Occurrents progress and develop over time while continuants exist through time. For instance, a chromosome is an occurrent, and the process of its duplication is a continuant.

2.1.2 Ontologies in Computer Science and Biomedical Research

Integration of diverse information from a multitude of sources is commonly required as part of current research in bioinformatics and biomedicine. This task, however, is complicated by many factors including the size of the datasets, their heterogeneity, and the diversity of data types (Soldatova and King, 2005). One particular problem is that different research communities developed terminology in which the same words are used in different contexts implicating slightly diverging meanings. A good example from the biomedical domain is the word *pseudogene* (Goble and Stevens, 2008). Depending on the context or the database, a pseudogene is defined as a gene-like structure containing in-frame stop codons, a transposable cassette that is rearranged, or a full reading frame that is not transcribed. Ontologies may be applied for solving such problems by providing a conceptualization of terms and their relationships.

The main focus of ontological research in computer and information sciences differs from the primary focus in philosophy. While the later discipline aims at a logically rigorous formalization, computer science concentrates on reasoning efficiency (Smith *et al.*, 2003). In addition to representing a classification of objects in a specific domain, an ontology makes the attempt to describe these objects with their constituting properties and relationships to other objects. In analogy to the object-oriented programming paradigm, a concept is also called class, and objects in the real world are denoted as instances. A more precise characterization of the classes is achieved by including a textual definition

and attributes, which are intended to unify the usage of the class and allow for expressing its features. Class definitions may contain explicit statements about the domain of interest, called propositions. The definition of the term "ATP binding" from the molecular function ontology of GO, for instance, contains the proposition "Interacting selectively and non-covalently with ATP". Ontologies can also contain axioms that are unproven but accepted assumptions about the domain and serve as starting points for building the ontology (Kemerling, 2010; Schulze-Kremer, 2002). The structure, the axioms, and the propositions of an ontology provide the possibility for performing automatic consistency checks, reasoning, and inferences. Several languages have been developed for encoding ontologies. The Web Ontology Language (OWL), an official W3C recommendation, is commonly used in semantic web applications, while the Open Biomedical Ontologies (OBO) format developed by the OBO Foundry (Smith *et al.*, 2007) is widespread in the biomedical domain.

Ontologies are usually represented as undirected or directed graphs that may contain cycles (Figure 2.1). The concepts of an ontology are represented by nodes, which are connected by edges representing their relationships. Each concept can have more than one parent and more than one child node. Several types of relationships are used to link concepts to their parent and child terms in an ontology. The most common type in taxonomies are subsumption relationships that include specializations ("is a") as well as partitive links ("part of" or "has component") (Stevens *et al.*, 2000). In special cases, the graph structure simplifies to a tree. The taxonomy of species, one of the oldest taxonomies in the life sciences, is an example containing only subsumption relationships organized in a tree structure (Sayers *et al.*, 2009).

Classifications of Ontologies

Based on their properties, ontologies can be discriminated in several ways. A coarse classification distinguishes formal and non-formal ontologies (Antezana *et al.*, 2009). While logical frameworks are applied for building formal ontologies, natural language is commonly utilized for specifying non-formal ontologies. Therefore, formal ontologies are better suited for computational purposes. A second discrimination is between top-level, task-oriented, application, and domain ontologies (Gómez-Pérez *et al.*, 2004; Antezana *et al.*, 2009).

A top-level ontology represents the most generic type and models universal concepts that are not specific to a single field, for instance, *entity* and *object*. Such ontologies are mainly applied for integrating different domain-specific ontologies. The Basic Formal Ontology (BFO, Grenon *et al.*, 2004) and the General Formal Ontology (GFO, Heller and Herre, 2004) are examples of top-level ontologies.

Task-oriented ontologies define concepts for generalizing tasks. The BioMOBY project (Wilkinson and Links, 2002) developed a task-oriented ontology consisting of bioinformatics analysis types that is used for describing services referenced by MOBY Central.

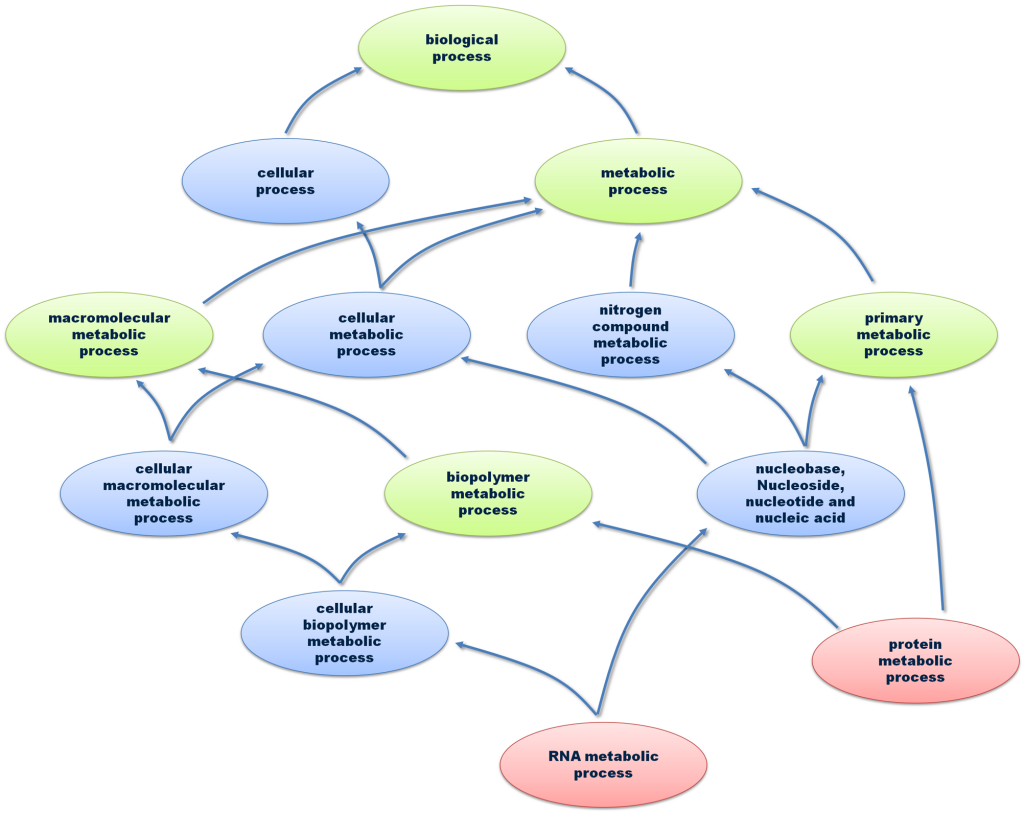

Figure 2.1: Part of the GO BP ontology. The leaf nodes are colored red, their common ancestors are colored green, and all other nodes are colored blue. The leaf nodes represent the most specific terms and the root node the most generic term. The edges represent "is a" relationships between the terms. For example, the relationship "cellular metabolic process is a metabolic process" indicates that "cellular metabolic process" is a subclass of "metabolic process". Therefore, all edges are directed upwards from the more specific node to the more generic node.

As another example, the Multiple Alignment Ontology (MAP, Thompson *et al.*, 2005) defines standards for retrieval and exchange of sequence and structure alignment data. The third type, application ontologies, focuses on concepts that are relevant for a particular application. The Cell Cycle Ontology (CCO, Antezana *et al.*, 2006) contains cell cycle related terms and aims at facilitating systems biology approaches to cell cycle research. Last, domain-oriented ontologies map the knowledge of one specific research domain. The ontologies for describing different aspects of gene product functions developed by the Gene Ontology consortium are an example for this category (GO, Ashburner *et al.*, 2000).

It is not possible to uniquely group all ontologies into one of these four categories. WordNet (Al-Halimi *et al.*, 1998) is a dictionary of the English language consisting of so called synsets, groups of synonymous words, along with a definition. Specialization

links, "is a" relationships, are used to represent the hierarchy of synsets. On the one hand, WordNet can be considered a domain-ontology of the English language. However, since it also contains ubiquitous concepts like *entity* and *object*, WordNet may also be classified as top-level ontology.

2.1.3 Biomedical Ontologies

One of the first large classifications in biology was the taxonomy of species developed by Linnaeus in the 18th century (McCray, 2006). At the end of the 1990s, ontologies were introduced to bioinformatics and later also to molecular biology (Schulze-Kremer, 2002; Soldatova and King, 2005). Since then, well over 100 ontologies have been developed in the biomedical domain, covering different categories such as anatomy, biological function, and phenotype. Today, the National Center for Biomedical Ontology (NCBO, www.bioontology.org) and the Open Biomedical Ontologies Foundry (OBO, www.obofoundry.org) are the two major initiatives supporting biomedical research through the development of ontologies and related tools.

National Center for Biomedical Ontology

The NCBO is an international research consortium aiming at fostering the application of ontologies in the biomedical domain for improving human health. To this end, the consortium provides tools for supporting the development and use of ontologies, which also entails the recommendation of formats and methods for working with ontologies (Musen *et al.*, 2006). One of these tools is BioPortal, a repository for accessing biomedical ontologies through web browsers and web services (Noy *et al.*, 2009). As of March 2010, BioPortal lists more than 190 ontologies along with additional metadata, for instance, a description, user comments, and information about the release. Besides the ontologies themselves, BioPortal offers cross-references of the contained ontologies to external data repositories, for example the Gene Expression Omnibus (GEO, Barrett *et al.*, 2007). The available mappings between different ontologies are especially important for integrating and working with several ontologies.

Open Biomedical Ontologies Foundry

The OBO Foundry is the second main ontology development effort in the biomedical domain (Smith *et al.*, 2007). This collaboration of developers from science-based domains aims at establishing a set of orthogonal reference ontologies. In order to achieve this goal, a set of design principles and best practices have been defined that are mandatory if an ontology is to be included into the foundry. One of the basic assumptions underlying these principles is that it is always possible to improve any given ontology, for instance, by including additional concepts, relationships, or definitions. The OBO foundry specifies standards concerning all levels of the process of ontology building (Bodenreider and

Stevens, 2006). On the design level, the requirements include that the ontology has to be expressed using either the OBO or the OWL syntax and the identifier space must be unique within the foundry. An additional prerequisite is that an ontology is developed in collaboration with other foundry members, and given appropriate acknowledgment, its use is free for everybody. These requirements ensure that all ontologies within the OBO foundry are interoperable and can be combined, for instance, through inter-ontology relationships. As of March 2010, the OBO Foundry contains 57 candidate ontologies, which can be grouped according to their main topic from genotype to phenotype, and 36 other ontologies and terminologies of interest. While genotype-related topics are more universal and are covered only by one ontology, several species-specific ontologies exist for anatomy, development, and phenotype (Bodenreider and Stevens, 2006).

Gene Ontology

In 1998, the Gene Ontology (GO) consortium started with the aim of producing structured and well-defined vocabularies for describing gene products in eukaryotes with respect to their molecular functions and their occurrence in biological processes and cellular components (Ashburner *et al.*, 2000; Hill *et al.*, 2009). Specifically, these descriptions should be easy to transfer between sequences with high similarity in different organisms. To this end, the GO developers mainly focused on the production of a controlled vocabulary instead of implementing or logically designing a theoretically centered ontology (The Gene Ontology Consortium, 2001). Initially, the GO ontologies were designed to annotate gene products from a generic eukaryotic cell, but this scope has been extended to all taxonomic lineages after many new members joined the consortium.

GO architecture GO consists of three orthogonal ontologies: molecular function (MF), biological process (BP), and cellular component (CC). Figure 2.1 depicts a small example from the BP ontology. Terms in the MF ontology describe the biochemical activity of a gene product. "Transporter activity" and "receptor binding" are examples for such functions. A biological process is an ordered assembly of several molecular functions that are accomplished by more than one gene product or a complex of gene products, for instance, "translation" and "cAMP biosynthesis". It is important to note that a process is generally not equal to a metabolic pathway. The "pantothenate and coenzymeA biosynthesis II" pathway (PWY-4221) from the MetaCyc (Caspi *et al.*, 2008) database maps to two different processes, "coenzyme A biosynthetic process" (GO:0015937) and "pantothenate biosynthetic process" (GO:0015940). The CC ontology describes the cellular substructure and complexes found inside or outside a cell, e.g. "ribosome" and "nucleus".

Each ontology is organized as separate directed acyclic graph (DAG), in which nodes represent the terms and the edges their relationships. The root node represents the most general term in an ontology graph, and the leaf nodes are the most specific ones. Edges are directed upwards from the more specific term to the more general term. Some applications introduce an artificial root node for combining the three ontologies into one single graph.

Every node may have several different parent nodes and more than one child node. Terms are defined by a set of mandatory attributes. These include the unique identifier, the term name, the ontology to which the term belongs (namespace), and its relationships to other terms. Since a definition of terms was initially not mandatory, there are still many terms that have none. In addition to the compulsory attributes, a number of optional ones can be attached to a term, including secondary identifiers, synonyms, database cross references, and comments.

Currently, there are three different types of relationships: "is a", "part of", and "regulates". The "regulates" relationship is further divided into two sub-relationships, "positively regulates" and "negatively regulates". The subsumption relationship "A is a B" indicates that *A* is a subclass of *B*. A "part of" edge, "C part of D", implies that *C* is always a part of *D* whenever *C* is present. However, if *D* exists, *C* does not necessarily exist. The regulation relationships express a regulatory interaction between one biological process and a second biological process or biological quality, such as cell size. The most frequent relationship is "is a" (almost 84 % in December 2009), and the highest fraction of "part of" relationships can be found in the cellular component ontology. Recently, a new relationship, "has part", was introduced. It is the logical complement of the "part of" relationship; "D has part C" expresses that whenever *D* exists, it necessarily has *C* as part. Relationships of this type have not yet been included in the general release.

From the definition of the relationships and their properties, logical rules can be derived that allow for automatic reasoning and for inferring new relationships between terms that are not directly connected by an edge. One important property of the "is a", "part of", and "has part" relationships is that all are transitive. A path like "A part of B part of C" implies that "A part of C", for example. Additionally, the general precedence of the relationships can be deduced; "regulates" has higher precedence than "part of", which in turn has higher precedence than "is a". Consequently, from the relationship "A regulates B is a C", it can be concluded that *A* regulates *B* and *C*.

GO Annotations The GO ontologies are widely used for annotating functional information to a variety of entities, for instance, genes, proteins, and protein families. Each annotation of a gene product with a GO term is supplemented with one of 17 evidence codes (EC) that indicates how the association was derived. The ECs can be divided into three classes: experimental, computational, and statement. An experimental EC implies that the annotation is based on the results of a biological experiment while a computational EC is used if the annotation is based on *in silico* analysis. If an annotation is due to a statement by an author or curator, one of the statement ECs is utilized. Common to all ECs is that they indicate that the annotation was reviewed by a curator. The only exception from this rule is the "Inferred from Electronic Annotation" (IEA) code, which is used for computational annotations that have not been reviewed. This is by far the most commonly used EC with more than 98 % (December 2009). A complete list of evidence codes can be found in Table 2.1. Of particular importance for all annotations with terms

Table 2.1: Complete list of GO evidence codes (EC). Each annotation of an entity with a GO term is attributed with one of these evidence code.

Class	EC
experimental	Inferred from Experiment (EXP)
	Inferred from Direct Assay (IDA)
	Inferred from Physical Interaction (IPI)
	Inferred from Mutant Phenotype (IMP)
	Inferred from Genetic Interaction (IGI)
	Inferred from Expression Pattern (IEP)
computational	Inferred from Sequence or Structural Similarity (ISS)
	Inferred from Sequence Orthology (ISO)
	Inferred from Sequence Alignment (ISA)
	Inferred from Sequence Model (ISM)
	Inferred from Genetic Context (IGC)
	Inferred from Reviewed Computational Analysis (RCA)
	Inferred from Electronic Annotation (IEA)
statement	Traceable Author Statement (TAS)
	Non-traceable Author Statement (NAS)
	Inferred from Curator (IC)
	No biological Data available (ND)

from GO is the *true path rule*. This rule states that if an entity is annotated with a GO term, all annotations with the ancestors of this term must also be valid.

Problems with the use of GO Ontologies in general and GO in particular are often misinterpreted and misused (Schulze-Kremer, 2002; Rhee *et al.*, 2008). Often, an ontology is interpreted as a collection of facts pertaining to a specific situation. However, the ontology provides and describes the classes that are necessary for describing the situation. Furthermore, an ontology is neither a database schema nor a knowledge base gathering information about objects, but it may be used to derive such a schema.

Focusing on the ontologies provided by the GO consortium, several shortcomings affect their construction and application. The main relationships, "is a" and "part of", are not used consistently throughout the ontologies leading to potential misinterpretations, for instance, "part of" relationships can have the meaning "causes" or "subprocess of" (Schulze-Kremer, 2002). By attaching a definition to each term, the consortium sought to standardize its usage in different communities. In the beginning, however, this was not mandatory, and many terms still are not defined properly. Moreover, the ontology lacks clear principles for including new terms and integrity constraints for checking the correct-

ness of the ontologies.

Annotation of entities with GO terms is performed by a large number of curators and with a variety of automated methods. While curators make annotations based on published information, their decisions are affected by a number of subjective influences such as their personal knowledge of the ontologies and of the biological detail (Camon *et al.*, 2005; Alterovitz *et al.*, 2010). Automatic methods often exploit mappings to other databases like cross-references to protein family resources or derive annotations by evaluating defined rules (Camon *et al.*, 2003). In both cases, the resulting annotations have varying confidence and level of detail. Moreover, the evidence code "ND" (Table 2.1) can be used to distinguish unannotated entities from entities for which no data are available. However, this is not utilized consistently among databases (Rhee *et al.*, 2008). In order to emphasize that an entity is lacking a specific function, the annotation can be modified using the "NOT" qualifier. Although this is essential for the correct interpretation of the annotation, this qualifier is ignored by several tools (Rhee *et al.*, 2008).

2.2 Semantic Similarity

As outlined previously, the utilization of ontologies for representing biomedical knowledge holds great promise for many applications, such as cross-database searches, automatic inference, and hypothesis generation. A necessary prerequisite for taking full advantage of ontological annotation, however, is the ability to quantify the semantic similarity between terms in an ontology. To this end, several methods have been proposed that assess the commonalities and differences of terms in an ontology (Pesquita *et al.*, 2009).

There are two elementary approaches for measuring the semantic similarity between two ontology terms, edge-based and node-based. Edge-based methods rely on the relationships in the hierarchy while node-based approaches utilize information on the terms themselves and their properties. In the following description of the similarity measures, several definitions are used:

An ontology consists of a set of terms and a set of relationships, which are represented by nodes and edges, respectively, in the ontology graph. In general, nodes can have several parents and children; nodes without parents are called *root* nodes, and nodes without children are called *leaf* nodes. The length of the path between two nodes t_1 and t_2 is defined as the number of edges on this path. The depth of a node t is given by the length of the shortest path from the root node to t. The depth of an edge e connecting nodes t_1 and t_2 is equal to the minimum of the depths of t_1 and t_2. The descendants of a node t are all nodes that lie on a path from t to a leaf node. The set of common ancestors (CA) of two nodes t_1 and t_2 consists of all nodes that lie on both a path from t_1 to the root node and a path from t_2 to the root node (Figure 2.1). The set of disjoint common ancestors (DCA) of two nodes contains all common ancestors that are not ancestors of any other common ancestors (Figure 2.2, Pesquita *et al.*, 2009). The lowest common ancestor (LCA) of t_1 and t_2 is the common ancestor with the largest depth. If two terms have several parents with

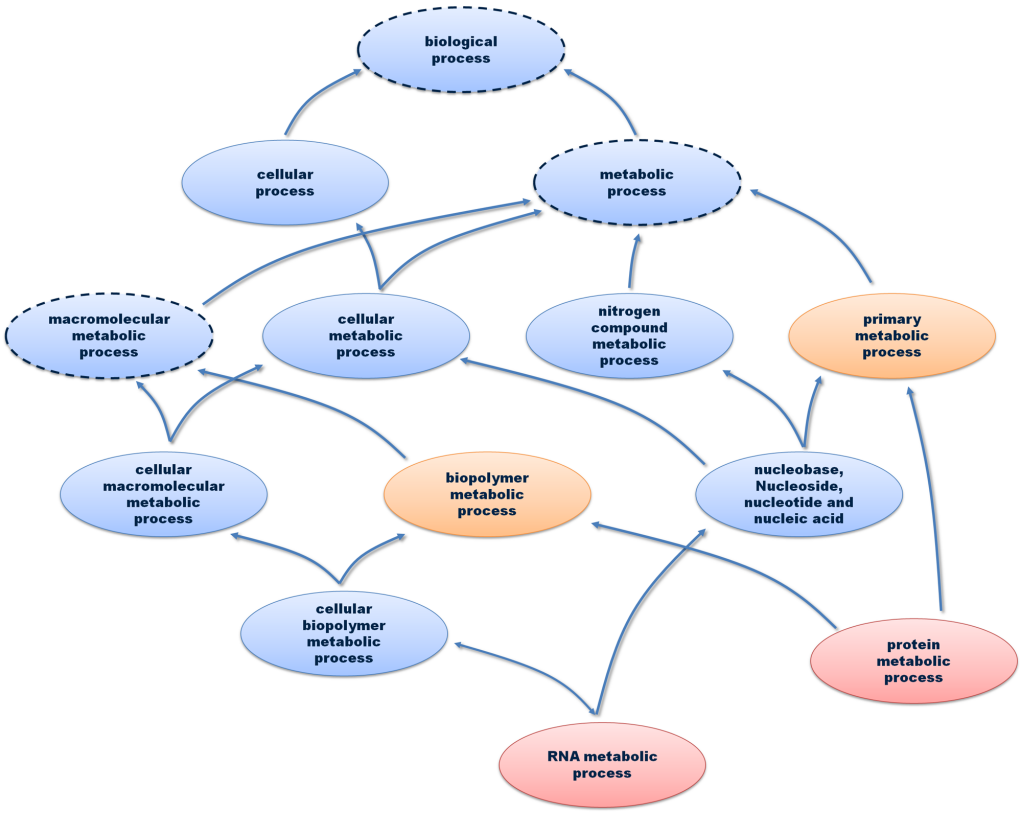

Figure 2.2: Part of the GO BP ontology. In contrast to Figure 2.1, the set of disjoint common ancestors (*DCA*) of the red leaf nodes are colored orange. Common ancestors that do not belong to the set of *DCAs* are drawn with a dashed outline.

the same maximal depth, one is randomly chosen or distances to all *LCAs* are averaged. Since there is a one-to-one mapping of ontology terms and nodes in the corresponding graph, the two words are commonly used interchangeably.

2.2.1 Edge-Based Measures

Edge-based measures rely on the hierarchical structure of the ontology for quantifying the semantic similarity between two terms t_1 and t_2. The most common approaches are counting the number of edges on the shortest path between t_1 and t_2, or taking the average length of all paths between the two nodes (Rada *et al.*, 1989). In both cases, the longer the path from one term to the other, the lower their similarity. A related approach is to define the semantic similarity of two terms as the depth of their *LCA* (Wu and Palmer, 1994; Wu *et al.*, 2005). In this case, a longer path signifies a higher similarity. Several variants of this approach have been proposed. Chiang *et al.*, and Pekar and Staab both introduced similarity measures that take into account the depth of the *LCA* of the two

terms and the length of the paths from the two terms to their *LCA* (Pekar and Staab, 2002; Chiang *et al.*, 2008). Wu *et al*. combined the depth of the *LCA*, the length of the paths from the two terms to their *LCA*, and the distance of the two terms to any leaf nodes into one measure (Wu *et al.*, 2006b). Yu *et al*. defined semantic similarity of two terms t_1 and t_2 as the ratio of their depth in the graph (Yu *et al.*, 2005).

These edge-counting approaches make several assumption that are not always true in biomedical ontologies (Jiang and Conrath, 1997; Lord *et al.*, 2003). First, they assume that nodes and edges are uniformly distributed over all parts of the ontology, and that nodes with the same depth have the same level of detail. However, both of these assumptions are commonly violated as the population of different regions of the ontology with terms represents the current state of biological knowledge (Alterovitz *et al.*, 2010). Thus, the number of nodes and edges is highly variable in different parts of the GO ontologies. The BP sub-ontology rooted at "reproduction", for example, contains about 800 nodes and 1,500 edges translating into approximately two edges per node whereas the "metabolic process" sub-ontology contains approximately three edges per node (about 6,000 nodes and 16,000 edges). Second, many edge-counting approaches take only "is a" edges into account although other relationship types may represent a substantial fraction of the total number of edges. Third, not all links between a parent node and its child nodes represent an equal semantic difference because the children may vary in the level of detail.

Different approaches have been introduced for accounting for these effects. Sussna, and Richardson and Smeaton introduced edge weights that were computed from network density, edge type, and edge strength (Sussna, 1993; Richardson and Smeaton, 1995). The sum of these edge weights on the shortest path between the two terms is then defined as semantic similarity. Cheng *et al*. introduced an edge weighting factor that is proportional to the depth of the edge and set the similarity to the sum of the weights on the shortest path from the root to the *LCA* (Cheng *et al.*, 2004). Li *et al*. defined the semantic similarity between two terms as a function of the length of the shortest path between them, the depth of their *LCA*, and the maximum information content of any of their common ancestors (Li *et al.*, 2003). Recently, Pozo *et al*. devised a very different step-wise approach (Pozo *et al.*, 2008). First, they created a co-occurrence vector for each MF term, which counted how many times two terms were both annotated to the same InterPro (Hunter *et al.*, 2009) entry. Second, the similarity between any two MF terms was computed as the cosine between the two corresponding co-occurrence vectors. Third, the matrix with all pairwise similarity values from step 2 was used as input for a spectral clustering algorithm, which projected the terms to a lower dimensional space. In the fourth step, Pozo *et al*. utilized an hierarchical clustering method to obtain a functional tree and defined the semantic similarity between two terms as the depth of their *LCA* in this functional tree.

2.2.2 Node-Based Measures

In contrast to using the hierarchy, node-based methods exploit properties of the terms in the ontology. In this respect, the most commonly used term property is the information content (*IC*), which is based on the consideration that a term conveys only little information if it is annotated to many entities. The *IC* of an ontology term t is defined as the negative logarithm of the term's probability (Resnik, 1995):

$$IC(t) = -\log p(t). \tag{2.1}$$

In order to calculate the *IC*, two empirically derived probability measures for ontologies have been introduced. First, Resnik defined the probability of a term as its relative frequency of occurrence in a large annotation database (Resnik, 1995). If all children are more specific than their parents, the total number of occurrences of any given term is the sum of its occurrences plus the number of occurrences of its children. The probability measure is defined accordingly:

$$p_{anno}(t) = \frac{occur(t)}{occur(root)}, \tag{2.2}$$

where $occur(t)$ is the number of occurrences of term t, and $occur(root)$ is the number of occurrences of the root term. The resulting probability increases monotonically from the leafs to the root, which has probability $p_{anno}(root) = 1$ if it is unique.

Zhang *et al*. proposed a second probability measure that is based on the hierarchy (Zhang *et al.*, 2006). Each leaf is assigned a distribution value (D) of 1 and the D values of the other terms are calculated as the sum of the D values of their children. Then, the probability of term t is given by dividing its D value by the D value of the root:

$$p_{graph}(t) = \frac{D(t)}{D(root)}. \tag{2.3}$$

Essentially, the distribution value of a term t measures the number of terms in the subontology rooted at t. The resulting probability measure $p_{graph}(t)$ has the same properties as $p_{anno}(t)$.

Resnik introduced the first semantic similarity measure based on *IC* (Resnik, 1995). The underlying intuition is that two terms are more similar if they share more information. Since this shared information is represented by the common ancestors of the two terms, Resnik defined the semantic similarity of two terms as follows:

$$sim_{Resnik}(t_1, t_2) = \max_{c \in CA}(-logp(c)) = IC(MICA), \tag{2.4}$$

where the most informative common ancestor (*MICA*) of terms t_1 and t_2 is the term with the highest information content in CA. Consequently, sim_{Resnik} has a minimum of 0 but

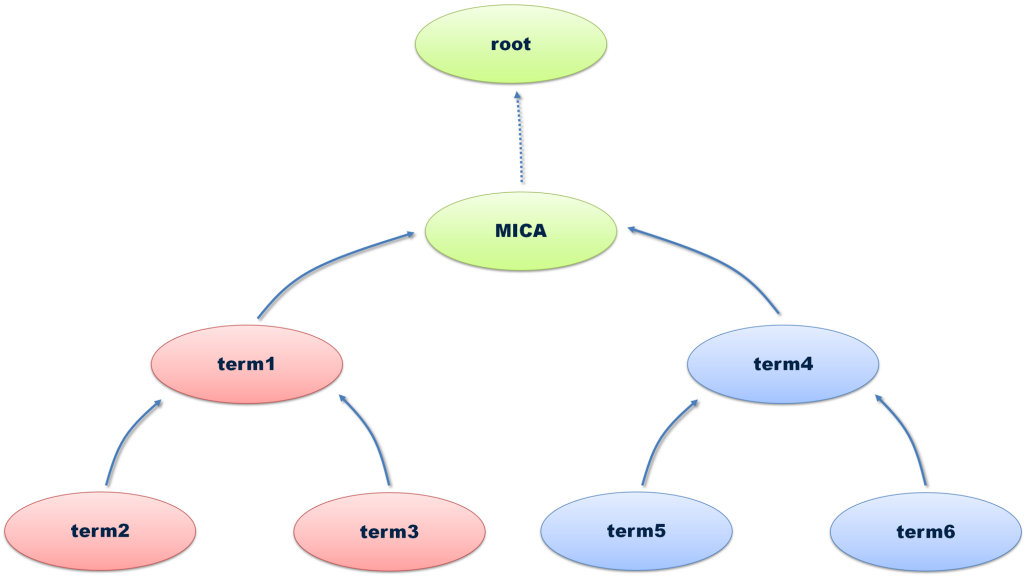

Figure 2.3: Example of an ontology illustrating the problem with sim_{Resnik}. The dotted edge represents a path of arbitrary length. In the shown case, any term pair with $t_1 \in \{term1, term2, term3\}$ and $t_2 \in \{term4, term5, term6\}$ has $sim_{Resnik} = IC(MICA)$. Therefore, these term pairs cannot be distinguished using this measure.

has no maximum. From Figure 2.3, the following issue of this approach becomes immediately apparent. Any pair of terms t_1 and t_2 with $t_1 \in \{term1, term2, term3\}$ and $t_2 \in \{term4, term5, term6\}$ are assigned the same sim_{Resnik}. Therefore, these term pairs are indistinguishable from each other when this measure is used for ranking term pairs.

Resnik's measure utilizes only the information two terms have in common. Intuitively, however, the similarity should also be inversely related to their differences, and the maximum similarity should be assigned if a term is compared to itself. This was taken into account by Jiang and Conrath (1997) and Lin (1998). Jiang and Conrath defined a distance measure between two terms as follows:

$$dist_{JC}(t_1, t_2) = IC(t_1) + IC(t_2) - 2 \cdot IC(MICA). \tag{2.5}$$

Later, this was transformed into a similarity measure (Couto *et al.*, 2007):

$$sim_{JC}(t_1, t_2) = \frac{1}{IC(t_1) + IC(t_2) - 2 \cdot IC(MICA) + 1}. \tag{2.6}$$

Lin defined the similarity of two terms as the ratio of their commonalities and the information needed to fully describe the two concepts, i.e., the sum of the information contents of the two terms (Lin, 1998):

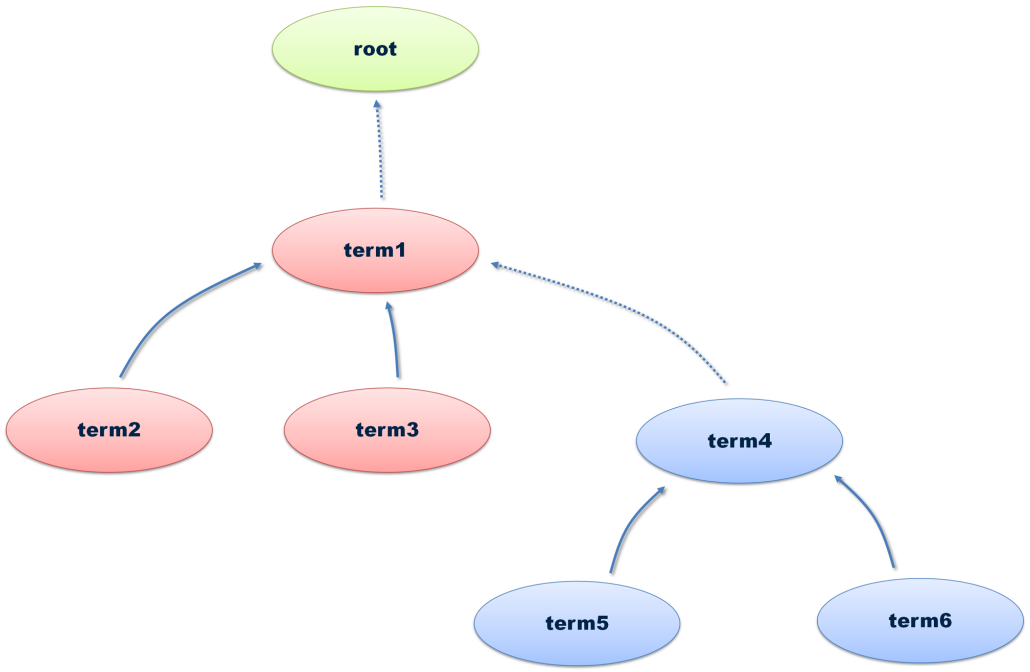

Figure 2.4: Example of an ontology illustrating the problem with sim_{Lin}. The dotted edges represent paths of arbitrary length. If $p(term1)$, $p(term2)$, and $p(term3)$ are multiples of a factor x with $p(term4)$, $p(term5)$, and $p(term6)$, respectively, it follows that $sim_{Lin}(term2, term3) = sim_{Lin}(term5, term6)$ although the first two terms are more general, and thus should receive a lower similarity value.

$$sim_{Lin}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1) + IC(t_2)}. \tag{2.7}$$

The values of both sim_{JC} and sim_{Lin} range from 0 to 1. The similarity defined by Lin quantifies the information of the *MICA* relative to the information of the two terms, but it does not account for the location of the terms in the graph. Therefore, it may assign high similarity values to very generic terms (Figure 2.4).

In order to overcome the limitations of Resnik's and Lin's measures, we introduced the relevance similarity (Schlicker, 2005). The relevance similarity weights sim_{Lin} with the probability of the *MICA*, thus incorporating information on how detailed the *MICA* is. The relevance similarity is defined as follows:

$$sim_{Rel}(t_1,t_2) = \frac{2 \cdot IC(MICA)}{IC(t_1) + IC(t_2)} \cdot (1 - p(MICA)). \tag{2.8}$$

In the measures described above, $IC(MICA)$ is utilized for assessing the maximum amount of commonality between the two terms. However, given that terms can have

more than one common ancestor, the information shared by two terms can be higher than $IC(MICA)$. Therefore, Couto *et al*. developed the GraSM approach, in which the $IC(MICA)$ is replaced with the average *IC* of all disjoint common ancestors (Figure 2.2). This approach can be applied to all node-based semantic similarity measures (Couto *et al.*, 2007).

All node-based semantic similarity measures described so far utilize the occurrence probability of an ontology term for determining the amount of information it carries. A different approach was taken by Bodenreider *et al*. who used an annotation database for deriving for each GO term a co-annotation vector of gene products that are annotated with this term (Bodenreider *et al.*, 2005). The semantic similarity between two terms is then computed as cosine between the two corresponding co-annotation vectors:

$$sim_{Bodenreider} = \frac{\mathbf{t}_1 \cdot \mathbf{t}_2}{|\mathbf{t}_1| \cdot |\mathbf{t}_2|}, \tag{2.9}$$

where $\mathbf{t}_1 \cdot \mathbf{t}_2$ represents the dot product of the two co-annotation vectors, and $|\mathbf{t}_1|$ and $|\mathbf{t}_2|$ their norm. Later, Sanfilippo *et al*. mapped terms from different ontologies with the help of co-annotation vectors and introduced a weighting factor for other node-based similarity measures (Sanfilippo *et al.*, 2007). Chagoyen and colleagues also measured semantic similarity as cosine between vectors, but they created term vectors from titles and abstracts of scientific publications that were associated with BP terms annotated to *Saccharomyces cerevisiae* proteins (Chagoyen *et al.*, 2006). Marthur and Dinakarpandian introduced a modification of the Jaccard distance as semantic similarity measure (Mathur and Dinakarpandian, 2007). Their measure is defined as follows:

$$sim_{MD}(t_1, t_2) = \frac{\frac{n(t_1 \cap t_2)}{n(t_1 \cup t_2)}}{\frac{n(t_1)}{N} \cdot \frac{n(t_2)}{N}}, \tag{2.10}$$

where $n(t_1 \cap t_2)$ is the number of gene products annotated with both terms, $n(t_1 \cup t_2)$ is the number of gene products annotated with any of the two terms, $n(t_1)$ and $n(t_2)$ the number of gene products annotated with t_1 and t_2, respectively, and N is the total number of gene products.

2.2.3 Hybrid Approaches

Several semantic similarity measures have been published that integrate edge-based and node-based approaches. Wang *et al*. proposed the first hybrid approach that is based on the semantic contribution of an ancestor a to a term t (Wang *et al.*, 2007). Each edge is assigned a weight that is determined by the edge type, and the weights of the edges on a path from t to a are multiplied. The maximum of these path scores is then defined as semantic contribution of a to t. Subsequently, the semantic similarity between two terms is calculated as sum of the semantic contributions of all common ancestors divided by

the sum of the semantic contributions of the ancestors of each term. Recently, Othman *et al*. introduced a distance measure that is based on edge weights (Othman *et al.*, 2008). They calculated the edge weights from the rank of the source node of the edge, the number of children this source node has, and the *IC* difference of the nodes that are connected by this edge. The semantic similarity between terms t_1 and t_2 is then calculated as sum of the edge weights on the shortest paths between t_1 and the *LCA* as well as t_2 and the *LCA*.

2.3 Functional Similarity

Since entities are usually annotated with multiple terms from the same ontology, it is necessary to develop measures that compare sets of ontology terms. These measures are primarily applied in the context of GO annotation and are often called functional similarity measures. In the following, they are divided into two groups: pairwise and groupwise measures (Pesquita *et al.*, 2009).

2.3.1 Pairwise Measures

For computing the functional similarity between two sets of terms, pairwise functional similarity methods integrate the semantic similarity between all possible term pairs. Given two entities A and B that are annotated with the sets GO^A and GO^B of ontology terms with sizes N and M, respectively, the similarity matrix S containing all pairwise similarity values is calculated as follows:

$$s_{ij} = sim(GO_i^A, GO_j^B), \forall i \in 1,...,N, \forall j \in 1,...,M. \tag{2.11}$$

This matrix S can be calculated using any semantic similarity measure. Three basic approaches are applied for computing the functional similarity from the matrix S: maximum, average, and best-match average (BMA). Lord *et al*. were the first to apply semantic similarity measures in the context of GO and devised the average approach for quantifying functional similarity (Lord *et al.*, 2003):

$$GOscore_{avg}(A,B) = \frac{1}{N*M} \sum_{i=1}^{N} \sum_{j=1}^{M} s_{ij}. \tag{2.12}$$

The maximum method assigns the highest pairwise semantic similarity value as overall functional similarity (Speer *et al.*, 2004; Wu *et al.*, 2005; Riensche *et al.*, 2007):

$$GOscore_{max}(A,B) = max\ s_{ij}, \forall i \in 1,...,N, j \in 1,...,M. \tag{2.13}$$

The best-match average (BMA) method is computed in several steps (Schlicker, 2005). First, the maximum values in the rows and the columns of matrix S are averaged giving the *rowScore* and *columnScore*:

$$rowScore(A,B) = \frac{1}{N}\sum_{i=1}^{N} \max_{1\leq j\leq M} s_{ij}, \tag{2.14}$$

$$columnScore(A,B) = \frac{1}{M}\sum_{j=1}^{M} \max_{1\leq i\leq N} s_{ij}. \tag{2.15}$$

The $rowScore(A,B)$ can be interpreted as comparing entity A to entity B, and the $columnScore(A,B)$ as a comparison of B to A. The final functional similarity is calculated as the average of these two scores (Schlicker, 2005; Azuaje *et al.*, 2005; Couto *et al.*, 2007; Wang *et al.*, 2007; Mathur and Dinakarpandian, 2007):

$$GOscore^{BMA}_{avg}(A,B) = \frac{1}{2}\cdot(rowScore(A,B) + columnScore(A,B)), \tag{2.16}$$

or as their maximum (Schlicker, 2005; Wu *et al.*, 2005; Pozo *et al.*, 2008):

$$GOscore^{BMA}_{max}(A,B) = max(rowScore(A,B), columnScore(A,B)). \tag{2.17}$$

In order to assign a high functional similarity, the $GOscore^{BMA}_{avg}$ requires that, for each function annotated to one entity, a similar function is also annotated to the other entity. In contrast, the $GOscore^{BMA}_{max}$ allows one entity to be annotated with additional terms. Tao *et al*. suggested a variation of the BMA approach that requires two terms to be reciprocal best matches, and that their semantic similarity exceeds a threshold for including it in the average calculation (Tao *et al.*, 2007).

The various pairwise approaches can be used to address different problems. The maximum approach gives high scores if two entities share one similar term. Thus, it allows for finding proteins that have one function in common. However, it is not suited for assessing the overall functional similarity because it disregards most annotations, leading to an overestimation of the true similarity. The average approach assesses whether two entities are annotated only with similar terms. In general, it will underestimate the true similarity since it requires each entity to be annotated only with similar terms to assign a high value. The BMA method is best suited for assessing overall functional similarity because it takes into account all annotations, but it uses only the highest similarities.

2.3.2 Groupwise Measures

Groupwise approaches do not rely on a semantic similarity measure for term pairs. Instead, most of these methods use the set of directly annotated terms augmented with the term's ancestors from the ontology graph. The simplest approaches perform an exact matching of terms. Lee *et al*. proposed the term overlap (*TO*), which counts the number of terms contained in both augmented sets (Lee *et al.*, 2004):

$$TO(A,B) = \mathbf{GO^A} \cap \mathbf{GO^B}, \tag{2.18}$$

where $\mathbf{GO^A}$ and $\mathbf{GO^B}$ contain the terms annotated to A and B, respectively, and their ancestors. Recently, Mistry and Pavlidis developed a variant of the *TO* measure that is normalized by the size of the smaller set, called the normalized term overlap (*NTO*) (Mistry and Pavlidis, 2008). Martin *et al*. implemented the Czekanowski-Dice distance and Jaccard similarity in their GOToolBox (Martin *et al.*, 2004). The Czekanowski-Dice distance is defined as follows:

$$dist_{CD}(A,B) = \frac{|GO^A \Delta GO^B|}{|GO^A \cap GO^B| + |GO^A \cup GO^B|}, \tag{2.19}$$

where $|GO^A \Delta GO^B|$ is the number of terms in the symmetrical difference of the two sets of annotated terms, and $|GO^A \cap GO^B|$ and $|GO^A \cup GO^B|$ are the number of terms in the union and intersection, respectively, of the two sets. The Jaccard similarity is defined as the ratio of the number of terms in the intersection and the number of terms in the union of the two sets:

$$sim_{Jaccard}(A,B) = \frac{|GO^A \cap GO^B|}{|GO^A \cup GO^B|}. \tag{2.20}$$

Gentleman implemented the *simUI* and the *simLP* measures in a software package (Gentleman, 2007). The *simUI* is equal to the Jaccard distance, and the *simLP* score is defined as the length of the longest path that is shared by the two sets. Ye *et al*. proposed a variant of the *simLP* measure that is normalized by the minimum and maximum depth in the GO hierarchy (Ye *et al.*, 2005). Pesquita *et al*. proposed a further variant of the *simUI* measure (Pesquita *et al.*, 2008). They defined the *simGIC* score that weights each term with its *IC* before calculating the Jaccard distance. Cho *et al*. simply used the *IC* of the most informative term that is contained in both sets of annotated terms as functional similarity (Cho *et al.*, 2007).

A different group of methods calculates the probability that an entity is annotated with a specific set of terms. Yu *et al*. introduced the total ancestry similarity measure that is based on the set of disjoint common ancestors (Yu *et al.*, 2007). The similarity between two entities is defined as the relative frequency of entities that share the exact same *DCA* set. Sheehan *et al*. utilized the probability of a gene product to be annotated with the nearest common annotation of two entities for deriving the *IC*, and then applied Resnik's (Equation 2.4) or Lin's (Equation 2.7) similarity for calculating the functional similarity (Sheehan *et al.*, 2008).

A third class of groupwise methods describe entities as vectors of terms. In general, the term vector contains a component for every term in the ontology. In the binary case, the component of a term is set to one if the term is annotated to the entity and to zero otherwise. It is also possible to assign a weight to each term in the vector. Chabalier *et*

al. used the inverse document frequency (*idf*) for weighting annotated terms (Chabalier *et al.*, 2005). The *idf* of term t is defined as the logarithm of the number of total entities divided by the number of entities annotated with the term t. The functional similarity between two gene products is then defined as the cosine between their respective term vectors. In addition to GO terms, Huang *et al*. included other annotation sources in the binary term vector and calculate the functional similarity using kappa-statistics (Huang *et al.*, 2007).

A unique approach to functional similarity was taken by Popescu *et al*. who defined a fuzzy similarity measure based on the information content (Popescu *et al.*, 2006). First, they augmented the term sets, GO^A and GO^B, with the lowest common ancestors of any $t_1 \in GO^A$ and any $t_2 \in GO^B$. Then, they calculated the fuzzy similarity of the terms shared between the augmented sets GO^A and GO^B.

2.3.3 Measures Combining Different Ontologies

Semantic similarity measures allow for a comparison of terms within the same ontology. However, gene products are generally annotated with terms from the BP, CC, and MF ontologies. Combining the annotations with terms from distinct ontologies can either be achieved by combining scores computed for different ontologies or by directly quantifying the semantic similarity between terms from different ontologies. We developed the *funSim* score, which combines the similarity according to BP and MF annotation (Schlicker, 2005). It is defined as follows:

$$funsim(A,B) = \frac{1}{2} \cdot \left[\left(\frac{BPscore}{\max(BPscore)} \right)^2 + \left(\frac{MFscore}{\max(MFscore)} \right)^2 \right], \tag{2.21}$$

where the *BPscore* and the *MFscore* are the functional similarity scores based on BP and MF annotations, respectively, and $\max(BPscore)$ and $\max(MFscore)$ are the maximal *BPscore* and *MFscore*, respectively. The normalization ensures that the *funSim* score is always between 0 and 1.

The XOA methodology introduced by Riensche and colleagues combines a semantic similarity measure for terms from one ontology with a co-annotation approach for comparing terms from different ontologies (Riensche *et al.*, 2007). This allows for computing inter-ontology semantic similarities. The overall functional similarity of two entities is then computed using the maximum of all pairwise semantic similarity values.

2.4 Summary

The use of ontologies in the biomedical domain is still in its early stages. Nevertheless, they have already proven to be valuable tools for biomedical research. Especially the GO

ontologies are utilized in a number of different approaches, for example, for finding functionally similar proteins (Lord *et al.*, 2003; Cao *et al.*, 2004), analysis of gene expression data (Speer *et al.*, 2004; Lee and Lee, 2005; Alexa *et al.*, 2006; Cho *et al.*, 2009), analysis of protein interactions (Ramírez *et al.*, 2007; Lu *et al.*, 2005), and disease gene prioritization (Yilmaz *et al.*, 2009). A more detailed reviewed of such methods is provided in the following chapters.

Major initiatives like the NCBO (Section 2.1.3) and the OBO Foundry (Section 2.1.3) have been initiated to unify and drive forward the field of bio-ontologies. To this end, they define standards for the development and application of bio-ontologies. A multitude of different ontologies is already available that aim at the description of biological phenomena ranging from molecular features to the development of organisms and complex phenotypes. Examples are the vocabularies developed by the Gene Ontology Consortium (Section 2.1.3), the Zebrafish Anatomical Ontology (Sprague *et al.*, 2008), the Human Phenotype Ontology (Robinson *et al.*, 2008), and the Mammalian Phenotype Ontology (Smith *et al.*, 2005). Other ontologies have been developed to facilitate the development of bioinformatics methods including the Systems Biology Ontology (Novère, 2006).

Currently, one of the main problems of many bio-ontologies is that they are controlled vocabularies rather than rigorously defined ontologies (Bodenreider and Stevens, 2006; Alterovitz *et al.*, 2010). However, in order to take full advantage of automatic consistency checking and reasoning, it will be important to formally define ontology terms as well as their relationships. The semantic and functional similarity measures described in this chapter are an example for automatic reasoning methods that were made possible through the adoption of ontologies. However, their widespread application depends on several points. Foremost, it is important to make biomedical knowledge available in form of ontological annotation. Second, it is to be expected that different semantic and functional similarity measures are suited differently for various concrete applications. In addition to BP annotation, at least annotation with MF terms has to be taken into account in order to find functionally similar proteins in different organisms, for instance. To find the best method for a specific problem, extensive benchmarks need to be carried out. In the next chapter, we provide a detailed analysis of our previously developed semantic and functional similarity measures. We also assess their performance in several medically important application scenarios.

Chapter 3

Analysis of Semantic and Functional Similarity

In this chapter, we describe the comprehensive analysis of our previously developed approaches for measuring the semantic similarity between GO terms (sim_{Rel}) and the functional similarity (*funSim*) of annotated proteins or protein families (Section 2.1). The experiments described in this chapter show that our new functional similarity measure is robust with respect to incomplete functional annotation. Using four sets of protein pairs with varying levels of sequence similarity, we analyzed the relationship between sequence similarity and our *funSim* measure of functional similarity.

We provide examples for utilizing semantic and functional similarity measures in various medically relevant applications. First, a comparison highlights differences and commonalities in the functions and processes in different taxa. Second, taking advantage of available MF annotation, we derive two-dimensional maps of the functional space of yeast proteins and Pfam protein families. This work was published in the journal BMC Bioinformatics (Schlicker *et al.*, 2006a).

3.1 Introduction

Today, genome annotation relies heavily on bioinformatics methods. The identification of homologous relationships is a powerful and frequently used approach for protein-level annotation (Stein, 2001), where query protein sequences are compared to sequences of characterized proteins in order to find homologies. Based on this comparison, proteins of unknown function are assigned to characterized protein families, generating testable hypotheses of their molecular function. However, this established annotation approach has several limitations and there is no obvious, simple relationship between sequence similarity and function (Devos and Valencia, 2000, 2001). More direct approaches for the functional characterization of gene products have been proposed. In particular, genomic context methods predict which gene products are involved in common biological pro-

cesses (Gabaldon and Huynen, 2004; von Mering *et al.*, 2005). Complementary methods use different protein features or structural information to predict the function of a gene product (Jensen *et al.*, 2003; Watson *et al.*, 2005; Domingues and Lengauer, 2007).

Several methods for computing functional similarity have been developed to take advantage of the increasing availability of annotation with GO terms. A detailed description of these methods is given in Chapter 2. Some issues have to be considered when devising a robust method based on GO annotation. First, a number of problems arise from the fact that GO is an ongoing project, and new terms are continuously added to the ontologies (Alterovitz *et al.*, 2010). Since not all parts of the ontologies are equally well developed, the depth of a term in the GO graph is not fully representative of the specificity of the underlying concept. Distinct terms with the same rank usually are not equally specific, and functional terms may still be missing. Second, apart from the ontologies themselves, the annotation of proteins and protein families with these terms is neither complete nor free of errors. The manual annotation with GO terms is based on knowledge found in the scientific literature or public databases. Nevertheless, it relies on human decisions, and therefore, it is considerably subjective (Wu *et al.*, 2006a). Moreover, the annotation of proteins and protein families is far from complete and many entities are completely lacking annotation.

Functional similarity based on GO annotation has been used in different applications. Lord *et al*. developed the first such approach (Lord *et al.*, 2003). They implemented GOGraph, a tool for calculating the functional similarity of protein pairs. Cao *et al*. integrated a semantic similarity search into the Bio-Data Warehouse, which uses Resnik's measure (Equation 2.4) for quantifying the similarity between two single GO terms (Cao *et al.*, 2004). Speer *et al*. employed a distance measure based on Lin's similarity (Equation 2.7) for clustering genes on a microarray according to their function (Speer *et al.*, 2004). Friedberg and Godzik used the MF annotation of protein structures in the Protein Data Bank (Berman *et al.*, 2000) for comparing protein folds on the functional level (Friedberg and Godzik, 2005). Lee and Lee applied Resnik's semantic similarity measure to functional annotations in order to infer modularized gene networks (Lee and Lee, 2005). Shalgi *et al*. utilized Lord's definition for a subcellular clustering score based on the cellular component ontology (Shalgi *et al.*, 2005). Björklund *et al*. developed a domain distance score for assessing the similarity of two domain architectures (Bjorklund *et al.*, 2005). They showed that this domain distance correlates well with Lord's approach to semantic similarity of proteins. Sevilla *et al*. analyzed the correlation of gene expression with Resnik's and Lin's measures of semantic similarity (Sevilla *et al.*, 2005) and concluded that Resnik's measure correlates well with gene expression.

Gene products are functionally similar if they have comparable molecular functions and are involved in similar biological processes. Such gene products did not necessarily evolve from a common ancestor and do not necessarily show sequence similarity. GO annotations capture the functional information that is available for a gene product and can be used as basis for defining a measure of functional similarity between them. Here, we provide a detailed analysis of the semantic similarity measure sim_{Rel} and the functional

similarity score *funSim* in different medically relevant applications. Processes and functions that are unique to pathogens and absent in the hosts present potential targets for new drugs. Therefore, we use sim_{Rel} to identify all processes and functions from two different, medically important groups of organisms. First, we determine BPs from fungi that do not appear in mammals, and second, we find MFs from *Mycobacteria* that do not occur in mammals. Furthermore, a detailed comparison of the *funSim* score with sequence similarity is provided. Then, this score is applied to find human proteins that are functionally related to yeast proteins. Maps of the functional space of yeast proteins and Pfam protein families are presented that were obtained using multidimensional scaling based on the comparison of MF annotations.

3.2 Materials and Methods

3.2.1 Database

The presented analyses were performed using the GOTax platform (Schlicker, 2005). This is a comparative genomics platform consisting of an integrated database, GOTaxDB, and a query tool, GOTaxExplorer. The used version of GOTaxDB contained the NCBI Taxonomy (Wheeler *et al.*, 2000, downloaded on 22 August 2005), Pfam release 18.0 (Finn *et al.*, 2006), SMART domains (Letunic *et al.*, 2006) from the InterPro release 11.0 (Mulder *et al.*, 2005), GO (Ashburner *et al.*, 2000) term definitions from the monthly release from August 2005, and protein information and annotations imported from UniProtKB release 5.8 (Wu *et al.*, 2006a). GOTaxExplorer provides a simple query language for accessing GOTaxDB. Importantly, it allows for performing semantic and functional similarity searches. The GOTax platform is freely accessible over the Internet at http://gotax.bioinf.mpi-inf.mpg.de.

3.2.2 Datasets with Protein Pairs

In order to be able to compare our functional similarity measures with sequence similarity (Section 3.4), we derived four sets of protein pairs, each pair consisting of one protein from the yeast *Saccharomyces cerevisiae* and one human protein. The different sets represent varying levels of evolutionary relationship: no sequence similarity (NSS), low sequence similarity (LSS), high sequence similarity (HSS), and orthology according to Inparanoid (IO, Remm *et al.*, 2001). Sequences of yeast and human proteins were taken from Inparanoid version 4.0.

Definition of the set IO A set with orthologous proteins (IO) from yeast and human was extracted from Inparanoid version 4.0 (Remm *et al.*, 2001). Inparanoid contains clusters of orthologous and paralogous proteins from two species. Each cluster is seeded

with one protein from the two species, which are reciprocally best matches. These two proteins are called main orthologs and receive an inparalog score of 1.0. Proteins from both species are added to the cluster if their sequence similarity to the main ortholog of the same species is higher than the sequence similarity between the two main orthologs. These added proteins are called in-paralogs and have an inparalog score that is smaller than 1.0. We extracted the pair of main orthologs from each Inparanoid cluster. Only pairs such that both proteins were annotated with BP and MF terms were considered, resulting in the final set consisting of 682 protein pairs.

Definition of the sets LSS and HSS For the two sets of protein pairs with low sequence similarity (LSS) and high sequence similarity (HSS), a BLAST (McGinnis and Madden, 2004; Tatusova and Madden, 1999) comparison of all yeast proteins against all human proteins from Inparanoid was performed. Human sequences without BP or MF annotation were excluded from this comparison. The Ensembl (Hubbard *et al.*, 2005) BioMart tool was applied for mapping the Ensembl accession numbers in Inparanoid to UniProtKB accessions (26 October 2005). We mapped the SGD accession numbers of the yeast protein sequences to UniProtKB accession numbers with the files from UniProtKB release 5.8. The sequence comparison was carried out with version 2.2.12 of the blastp program using default parameters and an E-value threshold of 0.001. For each yeast protein, the LSS data set contains the human protein with the highest E-value that is not the ortholog from the IO set. The human protein with the lowest E-value that is not the ortholog from the IO set was paired with each yeast protein in the HSS dataset. Both sets contain 989 protein pairs.

Definition of the set NSS In order to compile a set of protein pairs with no sequence similarity (NSS), we selected all human proteins with BP and MF annotation that are not contained in the IO set. Each yeast protein with BP and MF annotation in the IO set was randomly assigned to one of these human proteins. Using BLAST, we verified that none of the resulting protein pairs had significant sequence similarity. The NSS set contains 1356 protein pairs.

3.2.3 Comparison with Lord *et al.*

For this analysis, the previously defined datasets IO, HSS, LSS, and NSS were used. The semantic similarity between single GO terms was calculated using the sim_{Rel} measure (Equation 2.8). According to the original method by Lord *et al.* (2003), the comparison of proteins was performed applying the average approach (Equation 2.12), and $BPscore_{avg}$ and $MFscore_{avg}$ correspond to functional similarity according to the BP and MF annotations, respectively.

3.2.4 Functional Comparisons

The sim_{Rel} measure (Equation 2.8) was used for two comparisons. First, we carried out a comparison of biological processes annotated to fungal and mammalian proteins. Second, molecular functions of proteins from *Mycobacteria* were compared to functions of proteins from mammals. In order to investigate the functional similarity of protein pairs with varying sequence similarity, we calculated the $BPscore^{BMA}_{max}$ (Equation 2.17), the $MFscore^{BMA}_{max}$ (Equation 2.17), and the *funSim* score (Equation 2.21) for all pairs in the sets IO, HSS, LSS, and NSS. Additionally, the *funSim* score was applied for identifying functionally similar proteins in yeast and human. The 7,356 yeast proteins from UniProtKB release 5.8 were compared to the 70,447 proteins from human from the same release. From the yeast proteins, 3,000 could not be analyzed because they had no GO annotation, and another 1,300 proteins had either no BP or no MF terms assigned, resulting in an incomplete score. The data files for the comparison of biological processes from fungi and mammals ("bp_fungi_mammals.txt"), the comparison of molecular functions from *Mycobacteria* and mammals ("mf_myco_mammals.txt"), and the *funSim* comparison of yeast with human ("sc_hs.txt") are available for download at http://gotax.bioinf.mpi-inf.mpg.de/raw_data/.

3.2.5 Multidimensional Scaling

The statistical software environment R (http://www.r-project.org) was utilized for performing metric multidimensional scaling (MDS). The pairwise comparison of all yeast proteins with MF annotation yielded a symmetric similarity matrix. For performing the MDS, the functional similarity of two proteins was transformed into a distance by computing $d_{MF} = 1 - MFscore$. The same procedure was applied for computing the functional distance matrix of all Pfam families based on their MF annotation. The symmetric d_{MF} matrix was used as input for the *cmdscale* method in R to perform a metric MDS. Normalized stress (NS) was calculated as follows:

$$NS = \frac{\sum_{ij}(d'_{ij} - d_{ij})^2}{\sum_{ij} d^2_{ij}}, \tag{3.1}$$

where d'_{ij} and d_{ij} are the distances of proteins i and j in the low-dimensional space and in the original space, respectively. The change rate of normalized stress (CR) was calculated as follows:

$$CR_k = \frac{(NS_k - NS_{k-1})}{(NS_{k+1} - NS_k)}, \tag{3.2}$$

with k being the number of dimensions. Densities were estimated with a two-dimensional Gaussian kernel estimation by the *kde2d* function from the R software.

3.2.6 Hierarchical Clustering

A hierarchical clustering was computed with Pycluster version 1.29 and Python 2.4.2 (http://www.python.org) using a maximum linkage clustering algorithm. The distance matrix was the same as utilized for the MDS.

3.3 Comparing Biological Processes and Molecular Functions

The similarities and differences of the molecular biology between different taxonomic groups were investigated with the help of the sim_{Rel} measure. The sim_{Rel} score ranges from 0 to 1; pairs of GO terms with a score above 0.9 correspond to highly similar functions, between 0.5 and 0.7, the two GO terms may be considered functionally related, and below 0.3, they have no functional similarity. The following examples illustrate the relationship between the sim_{Rel} score and functional similarity. Comparing the GO term "biotin biosynthesis" (GO:0009102) with itself results in a sim_{Rel} score of almost 1. It is not exactly 1 because the definition of the sim_{Rel} score contains the factor $1 - p(MICA)$, where $p(MICA)$ is the probability to occur of the most informative common ancestor (Equation 2.8). If a term is compared to itself, it is its own *MICA*. The probability of the term is greater than 0 if it occurs at least once in the annotation database used to compute the probability, which leads to a sim_{Rel} score that is slightly smaller than 1. The terms "ATP-dependent chromatin remodeling" (GO:0043044) and "chromatin silencing at telomere" (GO:0006348) have a similarity score of 0.75. Both terms are descendants of "chromatin remodeling" (GO:0006338) and represent related processes. The processes "aromatic amino acid transport" (GO:0015801) and "L-glutamate transport" (GO:0015813) have a score of 0.56. The lowest common ancestor of these two terms, "amino acid transport" (GO:0006865), is rather generic, which results in a low sim_{Rel} score. The processes "chitin localization" (GO:0006033) and "ATP synthesis coupled proton transport" (GO:0015986) are completely unrelated, which is reflected by their low similarity score (0.30).

3.3.1 Comparison of Processes from Fungi and Mammals

Proteins participating in BPs that are unique to pathogens and absent in the hosts are potential targets for developing new drugs. We utilized the sim_{Rel} measure for identifying processes from fungi (NCBI Taxonomy id: 4751) that are not present in mammals (NCBI Taxonomy id: 40674). Table 3.1 contains the forty most dissimilar BPs annotated to fungal proteins when compared to BPs annotated to mammalian proteins. The two BPs with the lowest sim_{Rel} scores are "plasmid partitioning" (GO:0030541) and "chitin localization" (GO:0006033) with sim_{Rel} scores of about 0.16 and 0.30, respectively. They

Table 3.1: The 40 BPs from fungi with lowest sim_{Rel} values compared to mammalian BPs. The first and the second columns contain the accessions and the names of the BP terms. The third column contains the maximum sim_{Rel} value of this term compared to any mammalian BP.

GO accession	GO name	sim_{Rel}
GO:0030541	plasmid partitioning	0.15808
GO:0006033	chitin localization	0.30027
GO:0046713	boron transport	0.31932
GO:0009302	snoRNA transcription	0.38639
GO:0006089	lactate metabolism	0.38782
GO:0019630	quinate metabolism	0.39903
GO:0019541	propionate metabolism	0.42775
GO:0042128	nitrate assimilation	0.45020
GO:0009305	protein amino acid biotinylation	0.45870
GO:0016926	protein desumoylation	0.47222
GO:0031144	proteasome localization	0.49160
GO:0045116	protein neddylation	0.49535
GO:0000338	protein deneddylation	0.51801
GO:0006279	premeiotic DNA synthesis	0.52348
GO:0006522	alanine metabolism	0.53138
GO:0019985	bypass DNA synthesis	0.53997
GO:0048309	endoplasmic reticulum inheritance	0.54037
GO:0015847	putrescine transport	0.55201
GO:0031291	Ran protein signal transduction	0.55494
GO:0015801	aromatic amino acid transport	0.55565
GO:0042762	regulation of sulfur metabolism	0.55802
GO:0045458	recombination within rDNA repeats	0.55879
GO:0018298	protein-chromophore linkage	0.56308
GO:0000256	allantoin catabolism	0.56748
GO:0040031	snRNA modification	0.58001
GO:0042545	cell wall modification	0.58265
GO:0000373	Group II intron splicing	0.58438
GO:0000358	formation of catalytic U2-type spliceosome for second transesterification step	0.58438
GO:0000396	U2-type spliceosome conformational change to release U4 and U1	0.58438
GO:0046459	short-chain fatty acid metabolism	0.58763

are both unique to fungi, and in particular, "chitin localization" is a promising candidate for finding new drug targets (Ruiz-Herrera and San-Blas, 2003). The next step for identifying potential drug targets would be to assess the essentiality of the individual proteins associated with the selected processes for the survival of the organism.

The low scores of the processes "Boron transport" (GO:0046713) and "snoRNA transcription" (GO:0009302) show that the results of the comparison depend on the quality and the availability of the functional annotations. The human protein with the UniProtKB accession Q8NBS3 is actually involved in "boron transport" (Park *et al.*, 2004), but it was not annotated with GO terms in the used release of UniProtKB. The process "snoRNA transcription" is annotated to the yeast protein with the UniProtKB accession P53538 (Ganem *et al.*, 2003). Ensembl contains a predicted human ortholog (ENSG00000160075) that belongs to the same InterPro family (Mulder *et al.*, 2005) (IPR006811) as the yeast protein, but the human gene product was not annotated with GO terms.

3.3.2 Comparison of Functions from Mycobacteria and Mammals

Targeting proteins with molecular functions that are unique to pathogens holds the promise to be able to develop drugs with few side effects. Therefore, we applied the sim_{Rel} score for identifying molecular functions from the genus *Mycobacterium* (NCBI Taxonomy id: 1763) that cannot be found in mammals (NCBI Taxonomy id: 40674). Our database contains annotations for proteins of several *Mycobacterium* pathogens. *M. avium paratuberculosis* (NCBI Taxonomy id: 1770) is the causative agent for Johne's disease in ruminants and is possibly linked to Crohn's disease in humans. *M. bovis* (NCBI Taxonomy id: 1765) causes tuberculosis in most animals and particularly in cattle. *M. tuberculosis* (NCBI Taxonomy id: 1773) and *M. leprae* (NCBI Taxonomy id: 1769) are human pathogens causing tuberculosis and leprosy, respectively.

A list of the 30 most dissimilar functions according to sim_{Rel} is given in Table 3.2. The MF with the lowest sim_{Rel} score (0.05) is "3,4-dihydroxy-2-butanone-4-phos-phate synthase activity" (GO:0008686), indicating a function in *Mycobacteria* that is absent in mammals. In fact, this catalytic activity corresponds to one of the first steps in riboflavin biosynthesis. Riboflavin is the precursor of flavocoenzymes, which are essential for the catalysis of a variety of redox-reactions. Riboflavin is produced in microorganisms, fungi, and plants, but it is an essential nutrient for animals. The riboflavin biosynthetic pathway has recently been considered as potential drug target for anti-infectives against pathogenic fungi, bacteria, and mycobacteria in particular (Fischer and Bacher, 2005; Morgunova *et al.*, 2005). There has also been some specific interest in developing inhibitors of the 3,4-dihydroxy-2-butanone-4-phosphate synthase from different fungi (Echt *et al.*, 2004; Liao *et al.*, 2000). So far, however, it has not been studied for targeting mycobacteria. Another MF of potential interest that is not found in mammals is "UDP-N-acetylmuramate dehydrogenase activity" (GO:0008762). It has a sim_{Rel} score of 0.60 to the most similar function in mammals and represents a step in the synthesis of bacterial peptidoglycan.

Table 3.2: The 30 MFs from *Mycobacterium* with lowest sim_{Rel} values compared to mammalian MFs. The first and the second columns contain the accessions and the names of the MF terms. The third column contains the maximum sim_{Rel} value of this term compared to any mammalian MF.

GO accession	GO name	sim_{Rel}
GO:0008686	3&4-dihydroxy-2-butanone-4-phosphate synthase activity	0.05293
GO:0018786	haloalkane dehalogenase activity	0.13931
GO:0004125	L-seryl-tRNASec selenium transferase activity	0.18767
GO:0043365	[formate-C-acetyltransferase]-activating enzyme	0.30076
GO:0008862	formate acetyltransferase activating enzyme activity	0.33830
GO:0016216	isopenicillin-N synthase activity	0.35215
GO:0004475	mannose-1-phosphate guanylyltransferase activity	0.37570
GO:0008773	[protein-PII] uridylyltransferase activity	0.39112
GO:0003919	FMN adenylyltransferase activity	0.39503
GO:0050348	trehalose O-mycolyltransferase activity	0.40932
GO:0004654	polyribonucleotide nucleotidyltransferase activity	0.41180
GO:0047330	polyphosphate-glucose phosphotransferase activity	0.41820
GO:0016210	naringenin-chalcone synthase activity	0.42160
GO:0030401	transcription antiterminator activity	0.42698
GO:0008910	kanamycin kinase activity	0.42998
GO:0008928	mannose-1-phosphate guanylyltransferase (GDP) activity	0.43487
GO:0008879	glucose-1-phosphate thymidylyltransferase activity	0.43877
GO:0008710	8-amino-7-oxononanoate synthase activity	0.46909
GO:0016852	sirohydrochlorin cobaltochelatase activity	0.48446
GO:0008968	phosphoheptose isomerase activity	0.51468
GO:0008887	glycerate kinase activity	0.52137
GO:0016851	magnesium chelatase activity	0.52842
GO:0004063	aryldialkylphosphatase activity	0.52996
GO:0000036	acyl carrier activity	0.53992
GO:0046025	precorrin-6Y C5&15-methyltransferase (decarboxylating) activity	0.55599
GO:0046026	precorrin-4 C11-methyltransferase activity	0.56021
GO:0008832	dGTPase activity	0.57983
GO:0008691	3-hydroxybutyryl-CoA dehydrogenase activity	0.58093
GO:0045156	electron transporter & transferring electrons within the cyclic electron transport pathway of photosynthesis activity	0.58192
GO:0008949	oxalyl-CoA decarboxylase activity	0.59180

A further example is "adenosylmethionine-8-amino-7-oxononanoate transaminase activity" (GO:0004015), which forms part of the biotin synthesis, and has a maximum sim_{Rel} score of 0.65.

3.4 Comparison of the *funSim* Score with Sequence Similarity

The *funSim* score ranges from 0 to 1, which translates into an increasing degree of functional similarity. A *funSim* value close to 1 indicates high functional similarity whereas a score close to 0 indicates low similarity. This is comparable to the sim_{Rel} score since the *funSim* score is a combination of sim_{Rel} scores. We analyzed the distribution of the *funSim* score and its two components, the $MFscore_{max}^{BMA}$ (for MF annotation) and the $BPscore_{max}^{BMA}$ (for BP annotation), with four different sets of protein pairs representing varying levels of evolutionary relationship: no sequence similarity (NSS), low sequence similarity (LSS), high sequence similarity (HSS), and orthology according to Inparanoid (IO) (Remm *et al.*, 2001). Sequence similarity is often applied to automatically annotate proteins with GO terms. In order to exclude a potential circular argument when comparing functional and sequence similarity, we performed two comparisons. For the first analysis, we disregarded all GO annotations with the evidence codes IEA (inferred from electronic annotation) and ISS (inferred from sequence or structural similarity). The second comparison includes all available GO annotations.

Figures 3.1A and B show the distributions of the $BPscore_{max}^{BMA}$ and the $MFscore_{max}^{BMA}$ for the four datasets after removing GO annotations with IEA and ISS evidence codes. Almost 60 % of the protein pairs in the IO dataset have an $MFscore_{max}^{BMA}$ above 0.8 and 45 % have a $BPscore_{max}^{BMA}$ in the same range. This indicates that orthologous proteins from Inparanoid tend to have similar functions, and to a smaller extent, are also involved in similar BPs. Some protein pairs in the IO set have scores below 0.2, indicating no semantic similarity of the available annotation. It can be seen in all four datasets that there are more protein pairs with a $BPscore_{max}^{BMA}$ between 0.2 and 0.8 than with an $MFscore_{max}^{BMA}$ in the same range. This is caused by the lower density of the MF ontology. High-level terms in the latter ontology are less connected by edges between each other than high-level terms in the BP ontology, which results in lower scores for MF. The percentage of proteins with high functional similarity (S0.8) is highest for the IO category, and decreases for HSS and LSS, to almost no protein pairs in the NSS set. The reverse is observed for the proteins without functional similarity (S0.0) where the highest percentage is observed for NSS and then in decreasing order LSS, HSS, and IO. This effect is more pronounced for the $MFscore_{max}^{BMA}$ than for the $BPscore_{max}^{BMA}$.

Figure 3.1C shows the distribution of the *funSim* score for the different datasets. Since the *funSim* score combines the other two scores, it exhibits an intermediate distribution. About half of the orthologous protein pairs have a score above 0.6 indicating some func-

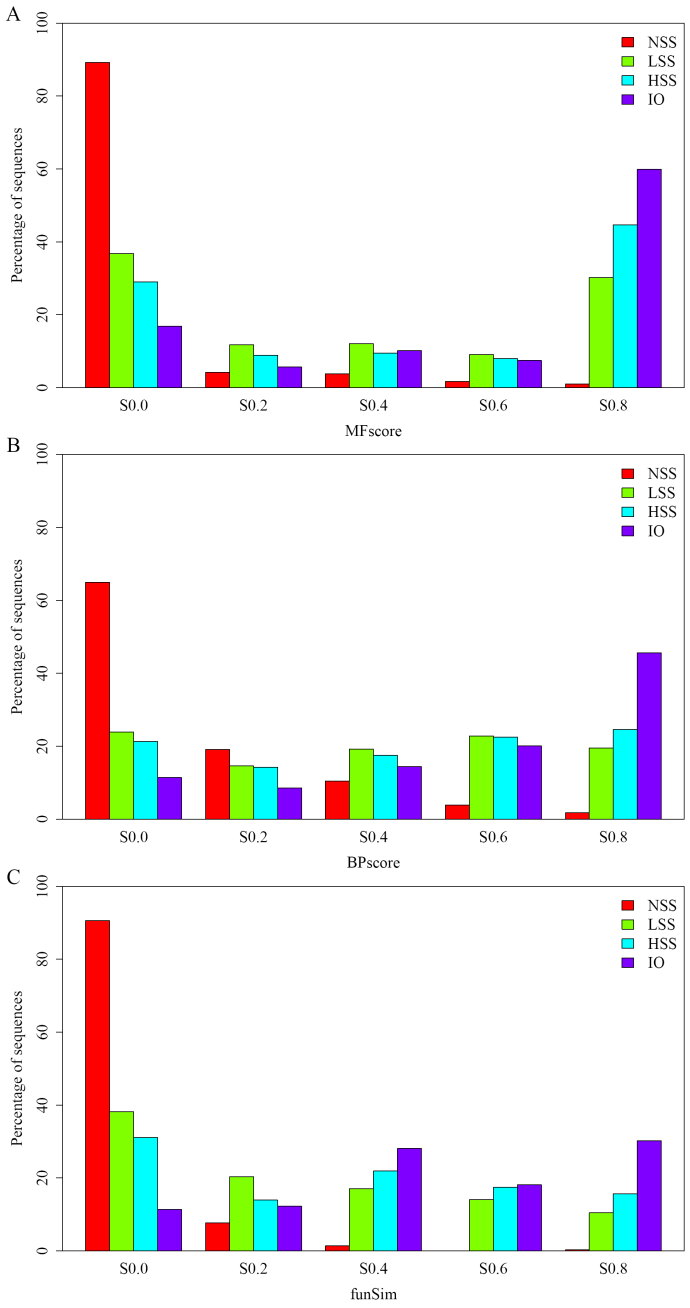

Figure 3.1: Distribution of (A) $MFscore^{BMA}_{max}$, (B) $BPscore^{BMA}_{max}$, and (C) *funSim* score for different sets of protein pairs. The bins correspond to the following intervals of functional similarity values: S0.0: $[0.0, 0.2[$; S0.2: $[0.2, 0.4[$; S0.4: $[0.4, 0.6[$; S0.6: $[0.6, 0.8[$; S0.8: $[0.8, 1.0]$. GO annotation using the evidence codes IEA (inferred from electronic annotation) and ISS (inferred from sequence or structural similarity) was excluded from the analysis, and thus the sets contain the following numbers of protein pairs: NSS 288, LSS 364, HSS 338, and IO 563. Percentages were calculated according to the size of the different sets.

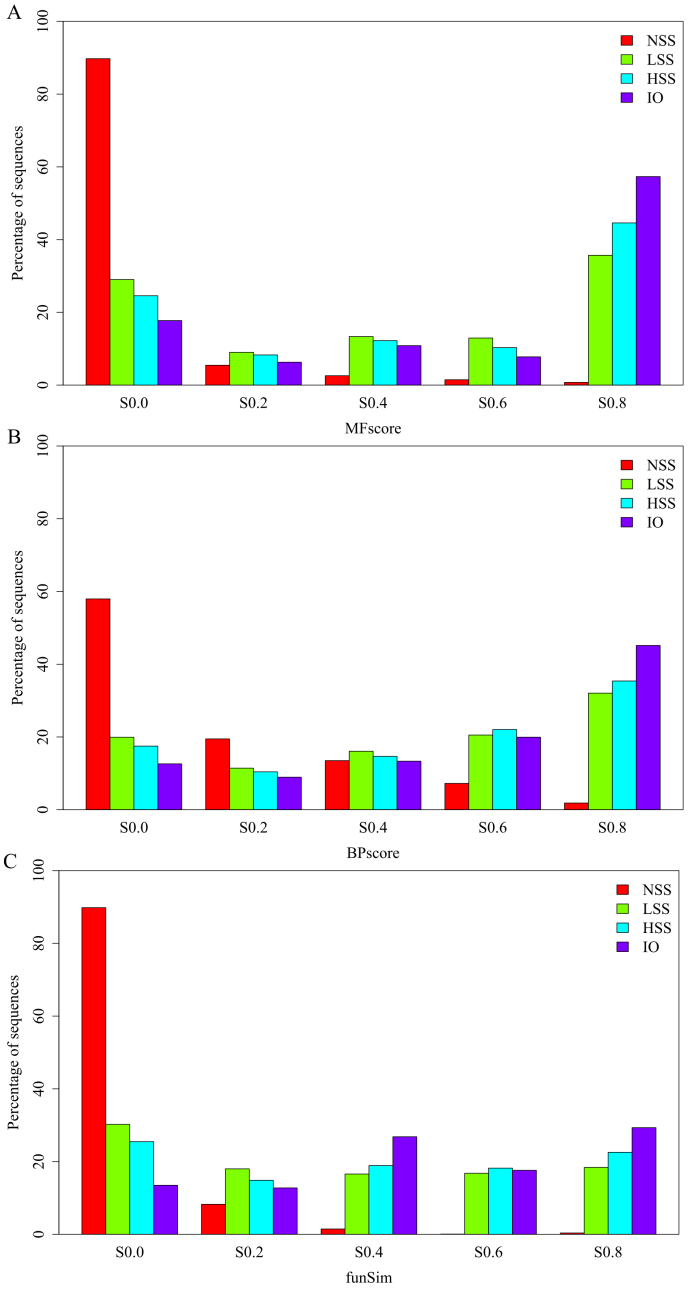

Figure 3.2: Distribution of (A) $MFscore_{max}^{BMA}$, (B) $BPscore_{max}^{BMA}$, and (C) *funSim* score for different sets of protein pairs. The bins correspond to the same intervals as in Figure 3.1. Percentages were calculated according to the total number of protein pairs in the different sets. The different sets contain the following numbers of protein pairs: NSS 1356, LSS 989, HSS 989, and IO 682.

tional relationship between the proteins. In particular, the highest peak is at S0.8, which suggests high functional relatedness of the proteins. This was to be expected given the high sequence similarity between orthologous proteins. Nevertheless, 25 % of the orthologous protein pairs have a *funSim* value below 0.4. This indicates that the corresponding protein pairs are annotated with BP and MF terms that have a very low semantic similarity to each other. One possible explanation is that the paired proteins are functionally different although they have a high sequence similarity, or the available GO annotation is incomplete. The IO distribution shows a local peak at S0.4, which is the result of combining the $MFscore_{max}^{BMA}$ and the $BPscore_{max}^{BMA}$. A considerable number of protein pairs have a high $MFscore_{max}^{BMA}$ and a low $BPscore_{max}^{BMA}$ or vice versa, resulting in *funSim* scores in the range between 0.4 and 0.6. With few exceptions, the protein pairs in the set NSS have very low scores. This indicates that there is almost no functional relationship between random protein pairs without sequence similarity. The distributions for the LSS and the HSS sets show considerable similarity. However, there is a shift in the LSS distribution towards lower scores if compared to the HSS distribution. This was to be expected since protein pairs in the HSS set have a higher sequence similarity than pairs in the LSS set.

Figure 3.2 depicts the distributions of the $BPscore_{max}^{BMA}$, the $MFscore_{max}^{BMA}$, and the *funSim* score for the four datasets including all available annotation. The comparison to Figure 3.1 shows that there is no large difference between the distributions in Figures 3.1 and 3.2. The $MFscore_{max}^{BMA}$ distributions of the LSS and HSS datasets have a lower percentage of very low scores (S0.0) but a higher number of middle MF scores (S0.4 and S0.6). These two datasets also have a higher percentage of protein pairs with high BP scores (S0.8). The same trend is also observable for the *funSim* score distributions, although to a lower extent. In general, excluding the electronic annotations does not have a noticeable effect on the distribution of the similarity scores. This shows that sequence similarity-based methods for annotating GO terms do no lead to biased annotations.

Figure 3.3 shows the relationship between $BPscore_{max}^{BMA}$ and $MFscore_{max}^{BMA}$ for protein pairs in the IO dataset. The bars are colored according to the *funSim* score of the protein pairs. The highest peak is observed at M0.9 and B0.9 indicating that many Inparanoid orthologous pairs perform the same function and are involved in the same processes. A considerable number of protein pairs have a high score (higher than 0.8) in one of the ontologies and a low score (lower than 0.2) in the other ontology. This corresponds to the upper left and the lower right corners of the plot. These proteins have either similar functions but take part in different biological processes, or they perform different molecular functions in similar biological processes. The *funSim* score of these protein pairs lies between 0.4 and 0.6, which explains the local peak for orthologous proteins at S0.4 in Figure 3.1C.

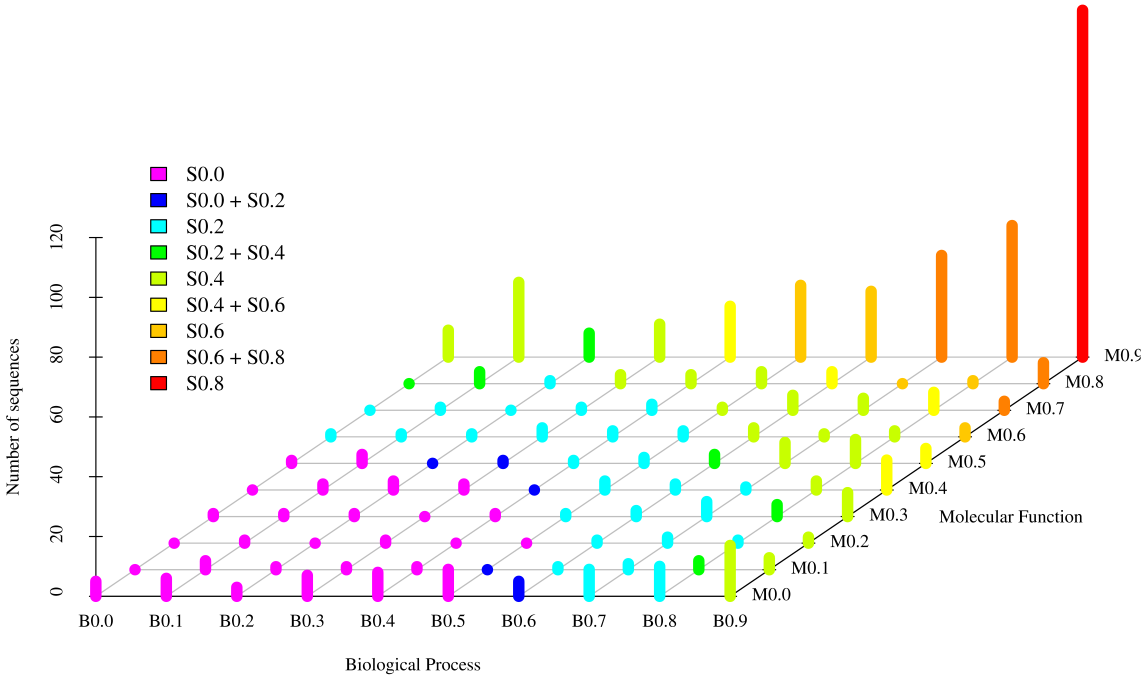

Figure 3.3: Distribution of $MFscore_{max}^{BMA}$ and $BPscore_{max}^{BMA}$ values for the IO dataset. The functional similarity bins correspond to the same intervals as in Figure 3.1. The bars are colored according to the *funSim* score of the protein pairs contained. If one bar contains protein pairs with *funSim* scores from two different bins, it has a color that is different from these two bins. Dark blue bars, for example, contain protein pairs with a *funSim* score in S0.0 or S0.2. From the plot, it can be seen that a considerable number of protein pairs have a high score (higher than 0.8) in one of the ontologies and a low score (lower than 0.2) in the other ontology, which corresponds to the upper left and the lower right corners of the plot. The *funSim* score of these protein pairs lies between 0.4 and 0.6.

3.5 Comparison of Average and Best-Match Average Approaches

We compared our measure of functional similarity between gene products to the approach previously proposed by Lord *et al.* (2003). Several challenges complicate such a comparison. First, there are no objective validation sets available. Second, Lord's measure has no upper limit and can be arbitrarily large. Third, there is no established cutoff for significant similarity for functional similarity measures. However, a partial comparison of the two approaches is still possible regarding the combination of semantic similarity scores. We compared our $MFscore_{max}^{BMA}$ and $BPscore_{max}^{BMA}$ to the corresponding $MFscore_{avg}$ and $BPscore_{avg}$, which rely on the average semantic similarity between the GO terms as proposed by Lord and colleagues. In order to obtain scores that range within predefined intervals using Lord's measure, we used sim_{Rel} to estimate the semantic similarity between

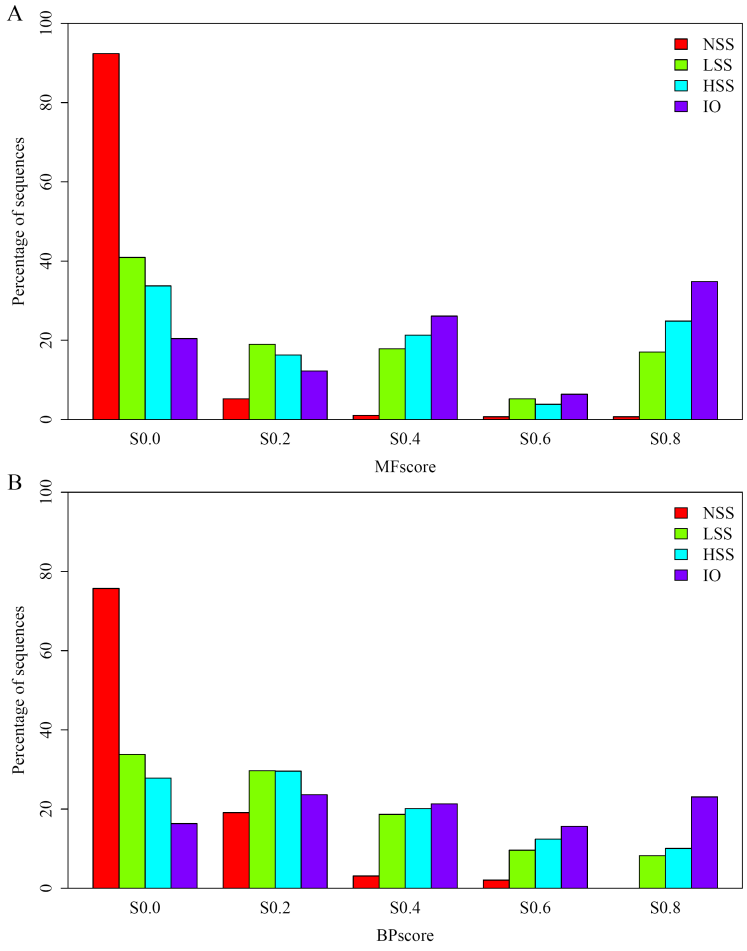

Figure 3.4: Distribution of (A) $MFscore_{avg}$ and (B) $BPscore_{avg}$ for different sets of protein pairs. The bins correspond to the same intervals as in Figure 3.1. Percentages were calculated according to the total number of protein pairs in the different sets. The different sets contain the following numbers of protein pairs: NSS 288, LSS 364, HSS 338, and IO 563. GO annotations using the evidence codes IEA (inferred from electronic annotation) and ISS (inferred from sequence or structural similarity) were removed before calculating functional similarity values.

GO terms. The $MFscore_{avg}$ and $BPscore_{avg}$ distributions were calculated for the NSS, LSS, HSS, and IO sets. Most protein pairs in the NSS set are not functionally related and therefore should obtain low similarity scores whereas pairs in the IO set generally have similar functions. The NSS set, however, contains protein pairs with no significant sequence similarity despite being functionally related. Although this prevents a fully objective performance assessment, comparing the shapes of the *GOscore* distributions for the NSS and the IO sets provides an indication of the discriminative power of the two approaches.

From Figure 3.4 one can observe that the shapes of the distributions of $MFscore_{avg}$ and $BPscore_{avg}$ differ from that of the corresponding distributions of $MFscore_{max}^{BMA}$ and $BPscore_{max}^{BMA}$ (Figure 3.1). There is a substantially lower percentage of protein pairs with $MFscore_{avg}$ above 0.8 than with $MFscore_{max}^{BMA}$ but a higher percentage of pairs with similarity between 0.2 and 0.6. The $MFscore_{avg}$ distribution of the IO set has two peaks, one at S0.4 and one at S0.8. From this, one can conclude that $MFscore_{avg}$ does not discriminate as clearly between non-homologous and homologous, and in particular orthologous, proteins as $MFscore_{max}^{BMA}$ does. The NSS results for $MFscore_{avg}$ closely resemble the results with $MFscore_{max}^{BMA}$. In case of the $BPscore_{avg}$, the IO, HSS, and LSS distributions are more uniform without pronounced peaks compared to the $BPscore_{max}^{BMA}$. The NSS distribution is very similar to the distribution obtained with $BPscore_{max}^{BMA}$.

In summary, these results confirm that functionally related proteins tend to have higher sequence similarity. This is even more evident for the $MFscore_{max}^{BMA}$. Nevertheless, a considerable percentage of orthologous protein pairs that have a high sequence similarity show no functional similarity. The comparison with the average approach of combining semantic similarity scores introduced by Lord *et al.* shows significantly different results. In particular, our proposed approach provides a better discrimination between non-homologous and orthologous proteins than the approach proposed by Lord and colleagues.

3.6 Finding Functionally Related Proteins

Sequence similarity is commonly applied for inferring functional relationships between orthologous sequences. Taking advantage of available functional annotation, however, represents a more direct approach for searching for functional relationships. We used the *funSim* score for identifying functionally related proteins in yeast and human. For every yeast protein, this comparison yielded a list of functionally related human proteins sorted by *funSim* score. In total, we compared 7,356 yeast proteins each with 70,447 human proteins from UniProtKB. The overall distribution of the highest *funSim* score for each yeast protein is depicted in Figure 3.5. The distribution shows that there are only about 30 yeast proteins with a maximal score below 0.4, which indicates that there is no functionally related protein in human. There is a functionally very similar protein in human with a score above 0.8 for almost 2,200 (30 %) yeast proteins. Out of these pairs, more than 1,600 have no significant sequence similarity with human proteins (NoSeqSim) and almost 1,400 share no Pfam (Finn *et al.*, 2006) families with human proteins. These functionally related protein pairs are either non-homologous and evolved independently to a similar function or are remote homologs that cannot be identified by standard sequence-based methods.

In the following, we analyze some exemplary protein pairs that were chosen to represent the different ranges of *funSim* values. The Glutaredoxin1 from yeast (UniProtKB accession: P25373) has a very high *funSim* score of about 1 to two human proteins

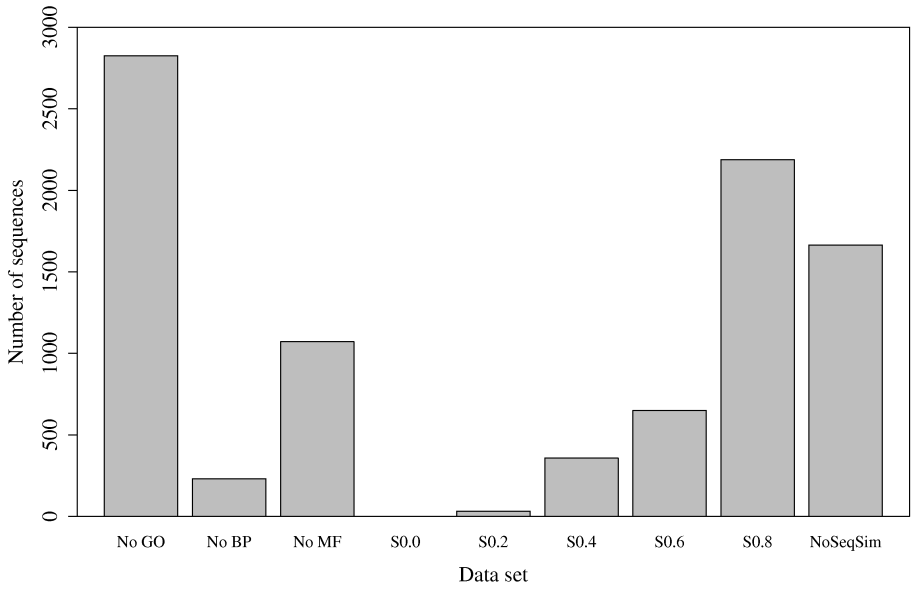

Figure 3.5: Functional comparison of yeast proteins with human proteins. Only the best hit (highest *funSim* score) for each yeast protein was taken into account for the score distribution. The 'No GO' bin contains all yeast proteins without BP and MF annotation; the 'No BP' and 'No MF' bins contain all yeast proteins without BP or MF annotation, respectively. The other bins correspond to the following intervals of maximum *funSim* values to any human protein: S0.0: $[0.0, 0.2[$; S0.2: $[0.2, 0.4[$; S0.4: $[0.4, 0.6[$; S0.6: $[0.6, 0.8[$; S0.8: $[0.8, 1.0]$. The 'NoSeqSim' bin contains all yeast proteins from S0.8 that show no significant sequence similarity to the functionally most similar human protein. The y-axis shows the number of yeast proteins in the corresponding bins.

(UniProtKB accessions: Q6NXQ3, Q5T501). All three proteins have glutathione peroxidase activity as response to oxidative stress. The three proteins belong to the same superfamily (thioredoxin-like) according to SCOP (Andreeva *et al.*, 2004) although the two human proteins have no significant sequence similarity with the yeast protein. This is reflected by their Pfam classification, which shows that the human proteins share the same family, but the yeast protein belongs to a different family.

The phosphoacetylglucosamine mutase from yeast (UniProtKB accession: P38628) matches one human protein (UniProtKB accession: O95394) with a high *funSim* score of 0.84. This human protein is also a phosphoacetylglucosamine mutase and performs exactly the same function on the same pathway, but it is annotated to a more generic BP term. The two proteins are reported as orthologs by Inparanoid. Specifically, they have a sequence identity of almost 46 % and share two Pfam families. These two proteins are functionally very similar.

Decarboxylating sterol-4-alpha-carboxylate 3-dehydrogenase (UniProtKB accession: P53199) from yeast is annotated with the MF "C-3 sterol dehydrogenase (C-4 sterol decar-

boxylase) activity" (GO:0000252) and the BP "ergosterol biosynthesis" (GO:0006696). The functionally most similar human protein is the sigma 1 isoform 1 variant Opioid receptor (UniProtKB accessions: Q53GN2, Q5T1J1) with a *funSim* score of 0.50. The human protein is annotated to the MF "C-8 sterol isomerase activity" (GO:0000247) and is involved in the same process as the yeast protein. The two proteins perform different functions, but they participate in the same processes, which is reflected by the low $MFscore_{max}^{BMA}$ (0.03) and the high $BPscore_{max}^{BMA}$ (1.0).

The serine/threonine-protein kinase ATG1 (UniProtKB accession: P53104) from yeast takes part in the "autophagy" (GO:0006914) process. The human protein with the highest *funSim* score (0.50) is phosphorylase b kinase gamma catalytic chain (UniProtKB accession: P15735) that is also annotated with serine/threonine protein kinase function but is involved in the "glycogen metabolism" (GO:0005977) process. Both proteins share the protein kinase domain from Pfam (Pfam accession: PF00069) and have a sequence similarity of 27 %. The proteins have the same molecular function ($MFscore_{max}^{BMA} = 1$) but take part in different processes ($BPscore_{max}^{BMA} = 0.16$). This is the type of functional relationship that tends to be predicted by homology-based methods.

The best hit for the nicotinamide riboside kinase 1 from yeast (UniProtKB accession: P53915) is the UMP-CMP kinase (UniProtKB accession: P30085) with a *funSim* = 0.30. The yeast protein catalyzes the synthesis of nicotinamide nucleotide from nicotinamide riboside, whereas the human protein catalyzes phosphoryl transfer from ATP to UMP and CMP. The two functions are not related, which is reflected by the low score.

3.7 Analysis of the Yeast Functional Space

We defined the distance score d_{MF} as a measure of functional distance with respect to the MF annotation. This distance is calculated as $d_{MF} = 1 - MFscore_{max}^{BMA}$. We computed d_{MF} scores for all pairwise combinations of yeast proteins. The underlying dataset consisted of all yeast proteins from UniProtKB with molecular function annotation, 3,459 proteins in total, resulting in 5,980,611 unique protein pairs. Approximately 5.3-million pairwise distances were larger than 0.8, indicating no functional similarity. Slightly more than 104,000 protein pairs had a distance below 0.2, suggesting high functional similarity. The d_{MF} scores were used as input for metric multidimensional scaling (MDS) and clustering in order to group the proteins according to their function. Previously, proteins have been grouped according to sequence or structure in a similar way (Hou *et al.*, 2003; Choi *et al.*, 2004; Kaplan *et al.*, 2005). The general aim of MDS is to represent points from a high dimensional space in a lower dimensional space while preserving the pairwise distances of the data points. As quality measure, normalized stress (NS, Equation 3.1) is applied for calculating how well the pairwise distances are preserved in the lower dimensional space. A low NS value indicates a good preservation of the original distances. For a dimension k, the change rate of normalized stress (CR, Equation 3.2) is defined as the difference $NS_k - NS_{k-1}$ divided by the difference $NS_{k+1} - NS_k$. The higher the CR value is, the

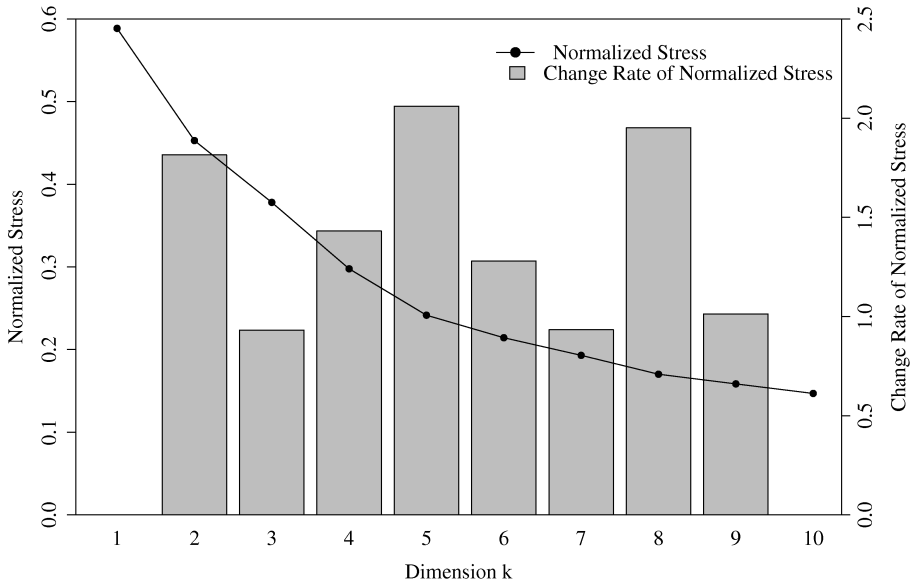

Figure 3.6: Scree-plot of the multidimensional scaling. The change rate indicates that a five-dimensional space would be optimal for representing the dataset. Furthermore, it shows that a three dimensional representation does not improve much over a two dimensional representation.

smaller is the benefit from adding one more dimension.

Figure 3.6 shows the plot with NS and the associated CR values. The normalized stress for the two-dimensional (2D) MDS of the dataset is 0.45, and the CR indicates that there is not much improvement in NS by using three dimensions instead of using two dimensions. Therefore, we chose the 2D MDS of the dataset for visualizing the map of the yeast functional space (Figure 3.7A). The contour plot in Figure 3.7B shows the regions corresponding to different functions, and the contour lines are colored to match certain child terms of "molecular function" and for some combinations of these high-level terms.

In general, proteins with the same function form clusters along axes and proteins annotated with two different functions are placed between the corresponding clusters. Proteins annotated with "catalytic activity" (1), for instance, are arranged along lines in the lower right part of the plot, and proteins with "binding" (2) annotation are located on an axis, approximately parallel to the x-axis to the left of the origin. The proteins that are annotated with both of these classes (6) are placed between these two clusters. Overall, the yeast proteins with different types of molecular functions are well separated in the MDS plot.

Further, we investigated how well the $MFscore_{max}^{BMA}$ discriminates between proteins with different types of "catalytic activity" (Figure 3.8). From this figure, it becomes evident that different functional subtypes are placed into distinct regions. By selecting six

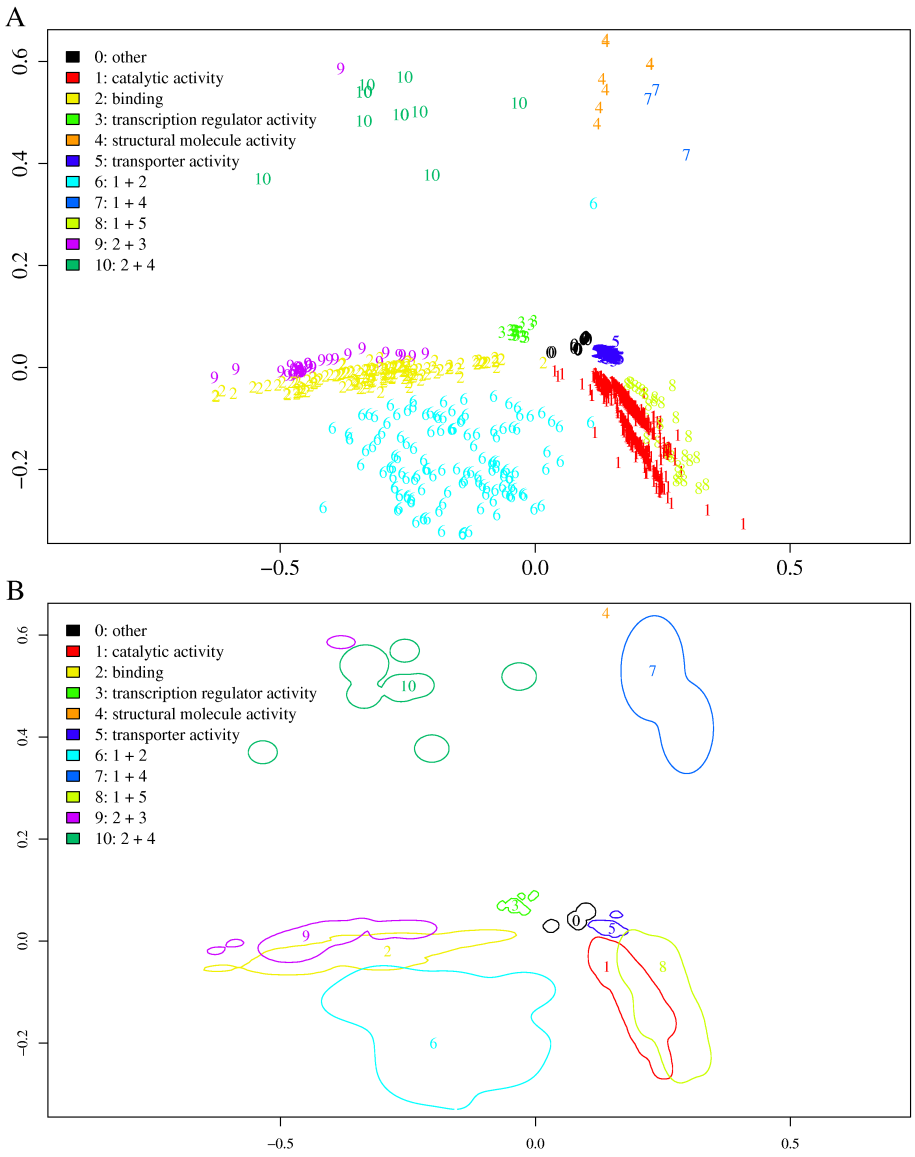

Figure 3.7: Yeast functional map. (A) The yeast functional map, obtained by 2D-multidimensional scaling of an all-against-all comparison of yeast proteins using d_{MF}. The proteins are represented by numbers in the plot and are colored according to their type of MF. The plot shows that proteins are clustered according to their diverging functions. Additionally, proteins annotated with two different functions are placed between the clusters corresponding to the single functions. (B) Contour plot of the MDS showing the density of the clusters. The contour lines were obtained from a 2-dimensional kernel density estimation using bivariate normal kernels. They were visually chosen to approximate the clusters.

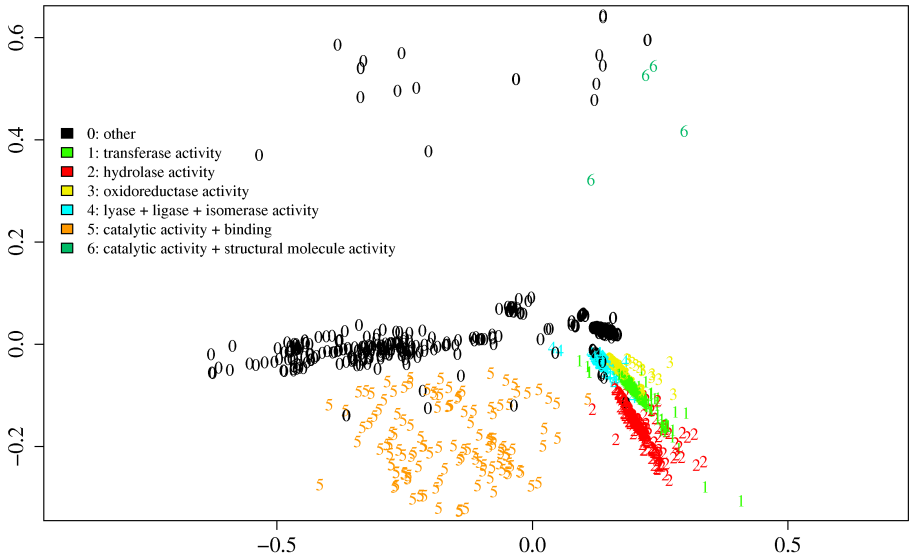

Figure 3.8: 2D-Multidimensional scaling plot with proteins colored according to the annotated type of "catalytic activity". One elongated region corresponds to "transferase activity" (1), another to "hydrolase activity" (2), and another region to "oxidoreductase activity" (3). Proteins annotated with "lyase activity", "ligase activity", or "isomerase activity" (4) are mostly located along the top of the whole "catalytic activity" region.

proteins annotated with a molecular function term descendant of "hydrolase activity", the arrangement of common functional subtypes was analyzed in further detail (Figure 3.9). This revealed that, in general, the lower the probability of occurrence of a term, the closer to the center of the plot proteins annotated with this term are located. This means that proteins positioned farther away from the origin are annotated with more generic and therefore less relevant GO terms. The same analysis with the $BPscore_{max}^{BMA}$ showed no clear separation of different high-level processes. This is possibly due to the increased density (number of edges) of the BP ontology as compared to the MF ontology.

The same distance matrix as used for MDS was utilized for performing a hierarchical clustering of yeast proteins according to their MF annotation. The resulting dendrogram is shown in Figure 3.10. From the clustering results, it can be seen that the five high-level functions form distinct clusters. The largest cluster "catalytic activity" is plotted in red. This cluster also contains proteins annotated with additional terms (labels 6 and 8 in Figure 3.7A). Proteins annotated with two different functional classes are placed into either one of the corresponding clusters. Generally, clustering with d_{MF} separates the yeast proteins according to their function, but the separation is not as clear as with multidimensional scaling.

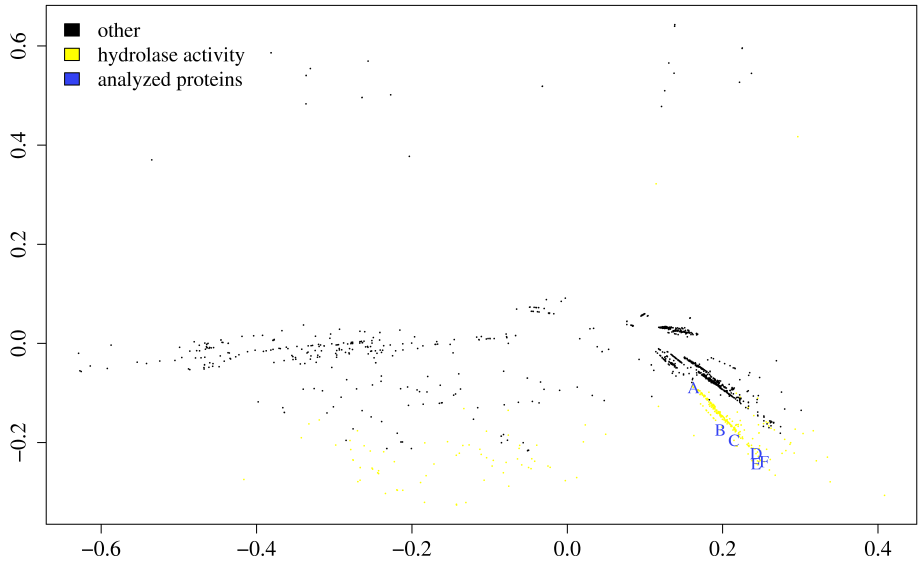

Figure 3.9: Detailed analysis of "hydrolase activity". Proteins annotated with a descendant of "hydrolase activity" are shown in yellow. The six marked proteins (A to F) are all annotated with a single molecular function as follows: Protein A (YBR177C), "serine hydrolase activity" ($p = 5.277 * 10^{-6}$); Protein B (DBP7), "ATP-dependent RNA helicase activity" ($p = 4.22 * 10^{-5}$); Protein C (YAL048C), "GTPase activity" ($p = 8.69 * 10^{-4}$); Protein D (Q36760), "endonuclease activity" ($p = 8.96 * 10^{-3}$); Protein E (YDL100C), "ATPase activity" ($p = 2.24 * 10^{-2}$); Protein F (IAH1), "hydrolase activity, acting on ester bonds" ($p = 2.71 * 10^{-2}$). The probability of the annotated term to occur increases moving on the line from A to F. This shows that proteins annotated with more general terms have a larger distance to all other proteins and thus are placed towards the edges of the plot.

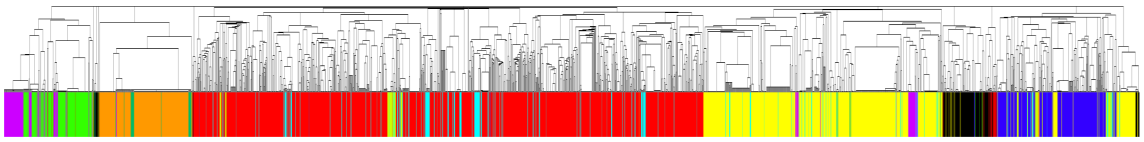

Figure 3.10: Hierarchical clustering of all yeast proteins using distances based on the *MFscore*. The color bar below the dendrogram is colored using the same scheme as in Figure 3.7 to indicate the molecular function annotation of the proteins. The dendrogram closely resembles the MDS of the yeast proteins and five clusters are apparent: "catalytic activity" in red, "binding" in pink, "transcription regulator activity" in light green, "structural molecule activity" in orange, and "transporter activity" in dark blue. The dendrogram was produced with the JavaTreeView software (http://jtreeview.sourceforge.net/).

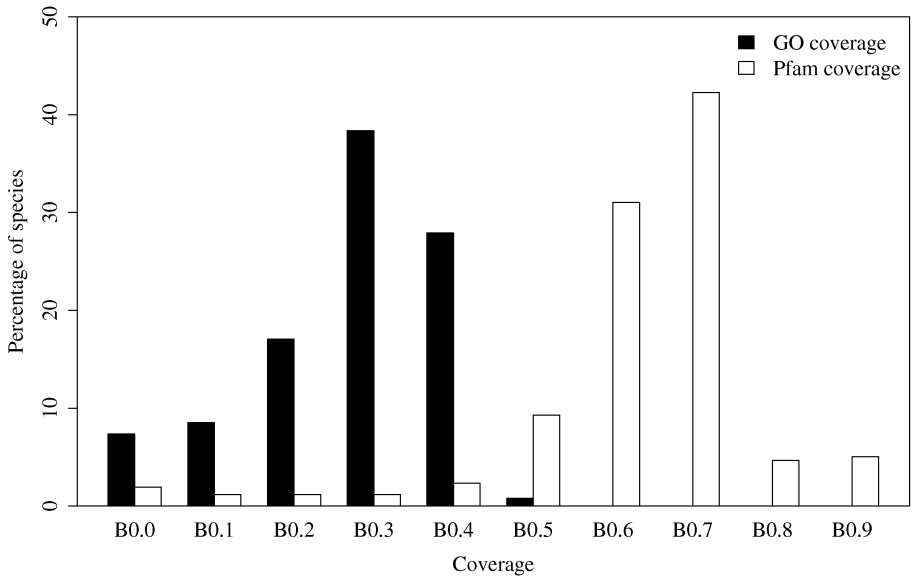

Figure 3.11: Distribution of coverage of proteins from completely sequenced genomes in UniProtKB annotated with GO terms and Pfam families. For calculating GO coverage, all proteins annotated with MFs and BPs were counted. The mean GO coverage for species in the database is 32 % and no species has a coverage above 60 %. The Pfam annotation is more complete with a mean of 67 %. The bins correspond to the following intervals of coverage: B0.0: $[0.0, 0.1[$; B0.1: $[0.1, 0.2[$; B0.2: $[0.2, 0.3[$; B0.3: $[0.3, 0.4[$; B0.4: $[0.4, 0.5[$; B0.5: $[0.5, 0.6[$; B0.6: $[0.6, 0.7[$; B0.7: $[0.7, 0.8[$; B0.8: $[0.8, 0.9[$; B0.9: $[0.9, 1.0]$.

3.8 Applying *funSim* to Pfam Families

Approximately half of the Pfam families are annotated with GO terms, and this annotation can be utilized for performing a functional comparison with the *funSim* measure. If a genome has low coverage with GO annotation but a rather high coverage with Pfam annotation, a *funSim* comparison based on Pfam families is actually preferable. Generally, for completely sequenced genomes, the Pfam coverage is higher than the GO coverage (Figure 3.11). One drawback of family-based functional comparisons is that Pfam families are largely annotated with more generic terms than proteins, because the functional annotation of a family has to match all its member proteins. As outlined before, the probability of a GO term can be used for quantifying how generic it is. Comparing the probabilities of GO terms annotated to human proteins to the probabilities of GO terms mapped to human protein families, it becomes evident that the annotation of families indeed is more generic than the protein annotation (Figure 3.12). Some genomes, however, have been annotated mostly using automated procedures based on sequence similarity, including Pfam searches with Hidden Markov Models. In such cases, the protein annotation will corre-

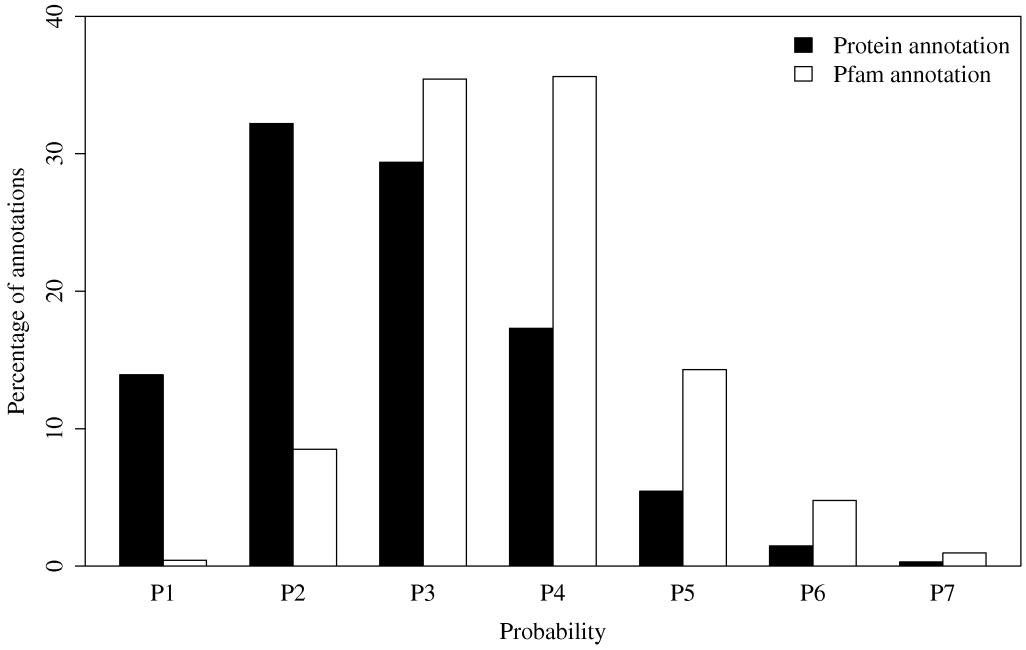

Figure 3.12: Distribution of probability values for GO terms annotated to human proteins or human Pfams. The bins correspond to the following intervals of GO term probability: P1: $[0.0, 10^{-7}[$; P2: $[10^{-7}, 10^{-6}[$; P3: $[10^{-6}, 10^{-5}[$; P4: $[10^{-5}, 10^{-4}[$; P5: $[10^{-4}, 10^{-3}[$; P6: $[10^{-3}, 10^{-2}[$; P7: $[10^{-2}, 10^{-1}[$.

spond to the terms shared by the different family members and more closely matches the Pfam annotation.

For Pfam families with molecular function annotation, we calculated all possible pairwise functional d_{MF} scores. The resulting distance matrix was used as input for a 2D MDS for obtaining a map of the Pfam functional space. In the graphical representation of the 2D MDS, the protein families are colored according to their MF (Figure 3.13). It can be seen that Pfams sharing the same function form well defined clusters, and overlapping clusters always contain families annotated with one common and possibly one additional function. Some clusters show a pronounced arrangement of protein families along axes where families annotated to more general GO terms locate towards the edges of the plot. Contour lines in the plot depict regions of constant density. They reveal a substantial overlap of clusters 2 and 9, which both contain families annotated with "binding". Additionally, cluster 2 is split into two large, distinct regions. An analysis of these two parts shows that the upper part consists of Pfams annotated to "protein binding" (GO:0005515) and the lower part contains Pfams annotated with other types of "binding". The main axes of the different clusters in the map of the Pfam functional space are shown in Figure 3.14.

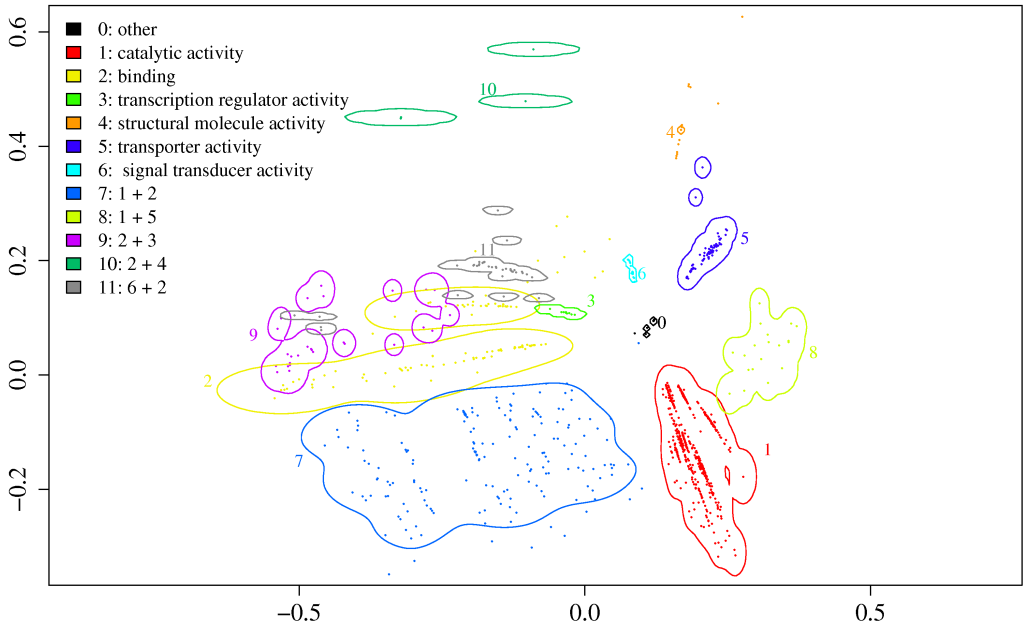

Figure 3.13: Functional map of Pfam families. Plot of the 2D MDS of all Pfam families with MF annotation. The colors were chosen to resemble the annotated functions.

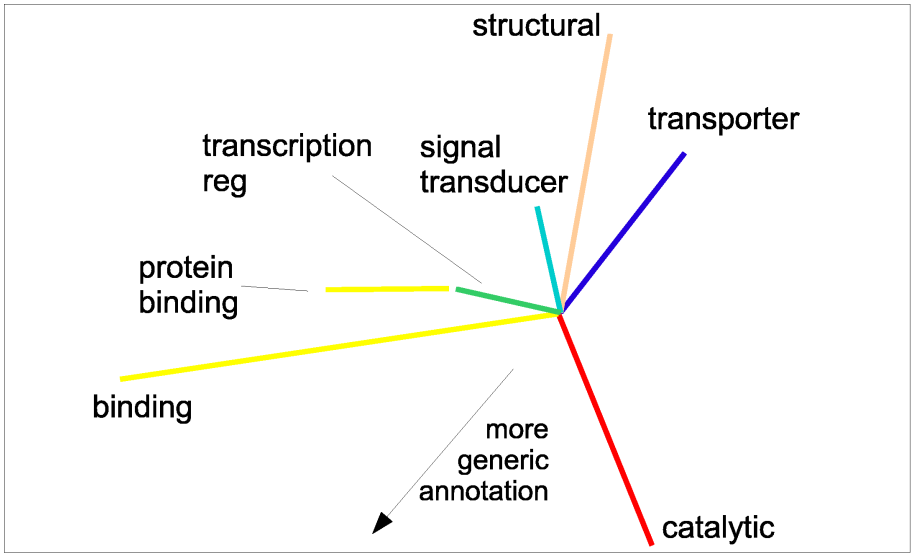

Figure 3.14: The plot shows the main axes of the largest clusters from Figure 3.13.

3.9 Conclusions

As a result of the genome annotation process, an increasing amount of functional information is being accumulated in a systematic and machine-readable fashion. New computational approaches that leverage this increasing knowledge promise to allow for more direct functional comparisons than traditional sequence comparison methods. While not aiming at replacing these sequence-based comparisons, the new approaches present an alternative for the objective comparison of the annotated gene products.

Here, we performed a detailed analysis of our semantic and functional similarity measures, sim_{Rel} and *funSim* (Schlicker, 2005). The sim_{Rel} score allows for quantifying the semantic similarity between two GO terms. Our evaluation shows that it combines the power of Resnik's (Resnik, 1995) and Lin's (Lin, 1998) measures in the sense that both the relevance of the most informative common ancestor (*MICA*) and the distance to the *MICA* are taken into account. The *funSim* score is based on sim_{Rel} for comparing the GO terms annotated to two entities; it also takes into account annotations with GO terms from different ontologies. The d_{MF} score is a variant of the $MFscore_{max}^{BMA}$ for measuring functional distances. Similar distance measures can be defined for the sim_{Rel} score, the $BPscore_{max}^{BMA}$, and the *funSim* score.

The $MFscore_{max}^{BMA}$, $BPscore_{max}^{BMA}$, and the *funSim* score allow for partial matches, therefore they are suitable for the comparison of multi-functional gene products. Moreover, these measures are applicable for the comparison of gene products for which only part of the functional knowledge is available as GO terms. This can be illustrated by the comparison of Glutaredoxin-1 from yeast with the GPX3 protein from human. The yeast and human proteins both have the peroxidase activity in common, but the yeast protein also exhibits transferase activity. Although the yeast protein is annotated with additional functions, both proteins clearly share similar functions, which is reflected by the high *funSim* score. Nevertheless, we also show that similarity measures based on annotation with ontologies are always limited by the availability and quality of this annotation and the underlying ontologies. This becomes evident in the missing "boron transport" annotation for the human protein Q8NBS3, which makes it impossible to find functionally related proteins in yeast.

The previously proposed measures of functional similarity rely on the maximum and average approaches. Consequently, these measures do not explicitly take into account partial matches, as they penalize all mismatches or consider only the best single match. These approaches are generally based either on Resnik's or Lin's similarity measures and consider either only the distance to the *MICA* or the relevance of the *MICA*. The lack of a gold standard of either truly functionally similar or dissimilar proteins limits the possibility of objectively comparing different functional similarity measures. This also constrains the comparison of our measures with Lord's average approach and restricts us to a comparison of the shapes of the score distributions. If the average approach of combining semantic similarity scores is used, the results differ significantly from the ones obtained with our BMA approach. The latter provides a better discrimination between

non-homologous and homologous proteins. Nevertheless, objective criteria for testing the performance of the different measures of functional similarity should be developed for fostering research in this area.

Our analysis provides examples for several medically relevant application scenarios for semantic and functional similarity measures. The sim_{Rel} score is used for finding terms that are shared between two sets of GO annotations or are unique to one of these sets. This is especially valuable for the comparison of the underlying molecular biology of different groups of organisms along the taxonomic tree. The comparison of the biological processes from fungi and mammals is one such example. In the second application scenario, functional relationships between proteins from different species are established using the *funSim* score. Two genomes are compared to find functionally similar gene products and to identify gene products unique to one of the species, respectively, as in the comparison between yeast and human proteins. Additionally, it is possible to compare all proteins from a single species and group them according to their functions. An example is the multidimensional scaling and the cluster analysis of the yeast proteins. A similar analysis can be performed on protein families in order to generate a map of the family functional space.

In summary, our analysis shows that semantic and functional similarity measures, which exploit ontological annotation, enable the comparison of the molecular functions and biological processes found in different groups of organisms. The latter present new tools for identifying functionally related gene products independent of homology. These comparisons can provide a better understanding of pathogenicity and aid in the identification of new drug targets. Established comparative genomics approaches for drug target discovery (Spaltmann *et al.*, 1999; White and Kell, 2004) can be extended with methods for functional comparison based on semantic similarity searches.

Although we have shown that functional similarity approaches are promising, we also provide evidence that the quality of the results is sensitive to the quality of the annotations. However, there is reason to be optimistic, since the situation is expected to improve as new GO terms are added and as more gene products are annotated. Most current approaches do not distinguish between the different relationships used in GO, for instance, "is a" and "part of". Further analysis is required for determining possible effects of the different relationships on the calculation of semantic and functional similarity. Another possible extension is to include annotations with terms from the cellular component ontology into the *funSim* score in order to completely assess the function and the cellular location of a gene product.

A future goal of methods for quantifying functional similarity is to identify functionally equivalent proteins from different species that perform the same molecular functions, take part in the same biological processes and are located in the same cellular components. This definition of functional equivalence is more generic than that of orthology as it does not depend on homology. The *funSim* score can be used as a basis for defining a new measure to identify the functionally equivalent gene products from different species.

Chapter 4

Software Applications for Functional Similarity Analysis

As outlined in Chapter 2, a multitude of semantic and functional similarity measures, including sim_{Rel} and *funSim*, have been developed. In order to take full advantage of ontological annotations in a wide variety of applications, devising robust similarity measures is only one part. In addition, it is crucial to implement software tools for fast calculation of functional similarity measures and make them available to a large user community. A further requirement for these tools is an effortless integration into existing and new workflows.

In this chapter, we describe two applications for performing functional similarity analyses, FSST and FunSimMat. The Functional Similarity Search Tool (FSST) was developed as part of the GOTax platform for comparative genomics to support flexible functional similarity analysis. FSST was described in a paper published in the journal Genome Biology (Schlicker *et al.*, 2007b). The Functional Similarity Matrix (FunSimMat) is a comprehensive database of precomputed functional similarity values. It provides interfaces for manually and programmatically accessing these values over the Internet. Papers describing FunSimMat were published in the Nucleic Acids Research database issues in 2008 (Schlicker and Albrecht, 2008) and 2010 (Schlicker and Albrecht, 2010).

4.1 Introduction

The complete sequencing and extensive annotation of genomes resulted in the creation of new opportunities for understanding biology at the molecular level (Camon *et al.*, 2004; Friedberg, 2006). Identifying the genes and gene products along with their corresponding molecular roles opens new possibilities for uncovering the agents and mechanisms taking part in the biology of different organisms. The comparison of two different genomes allows for identifying the common and unique characteristics of each of the genomes

and provides a way of transferring annotation from characterized to uncharacterized sequences.

As more and more genes and gene products from species across the whole taxonomic tree are functionally characterized and annotated, differences and similarities in the molecular biology between different taxonomic groups can be investigated in a systematic and objective way. The comparison of different sets of genomes allows for identifying the processes and functions unique to certain taxonomic groups or shared between taxonomic groups, for instance. A concrete application is the comparison between pathogenic and non-pathogenic bacteria, which can provide new insights into the mechanisms of pathogenicity. Another application is the comparison between human and pathogenic organisms in order to identify processes unique to pathogens, a first step in the discovery of new drug targets.

GO annotation and functional similarity is utilized in many different important applications. One such example is the analysis of gene expression data. Khatri and Draghici recently reviewed methods that apply GO for the analysis of microarray data (Khatri and Draghici, 2005), for example for identifying overrepresented GO terms in a list of differentially expressed genes (Draghici *et al.*, 2003; Alexa *et al.*, 2006). A number of approaches have also been developed for analyzing gene expression data considering functional similarity profiles (Speer *et al.*, 2004; Qu and Xu, 2004; Brameier and Wiuf, 2007; Yang *et al.*, 2008; Cho *et al.*, 2009). A different application of GO-based functional similarity is the prediction and validation of molecular interactions. For instance, functional similarity was found to be one of the best predictors for protein-protein interactions (Lin *et al.*, 2004; Lu *et al.*, 2005). In other work, it was utilized for the quality assessment of protein or domain interaction data (Schlicker *et al.*, 2007a; Ramírez *et al.*, 2007; Futschik *et al.*, 2007; Suthram *et al.*, 2006). Using functional similarity values, it is also possible to derive useful confidence thresholds for predicted domain-domain interactions (Chapter 5). Another application of functional similarity is the prioritization of putative disease genes (Chapter 6; Adie *et al.*, 2006; Franke *et al.*, 2006; Freudenberg and Propping, 2002; Perez-Iratxeta *et al.*, 2002; Rossi *et al.*, 2006; Chen *et al.*, 2009; Ortutay and Vihinen, 2009; Yilmaz *et al.*, 2009). Further uses of functional similarity include the identification of functional modules in interaction networks (Sen *et al.*, 2006; Pu *et al.*, 2007).

Despite the wide applicability of functional similarity measures, only few tools with limited functionality are readily available. The GOGraph tool by Lord *et al.* is available as a set of Perl scripts but requires additional packages and a database to be locally available (Lord *et al.*, 2003). DynGO is a downloadable application for performing functional similarity searches for gene products annotated with similar GO terms (Liu *et al.*, 2005a). In addition, some databases allow for functional similarity searches, but there is no comprehensive resource available providing access to a wide variety of precomputed functional similarity measures. The integrated bio-data warehouse BioDW at Fudan University integrates protein, protein family, and functional annotation databases (Cao *et al.*, 2004) and facilitates basic semantic similarity searches. However, such similarity searches are restricted to a single GO term and do not assess the overall functional similar-

ity of two proteins. GOTaxExplorer supports semantic and functional similarity searches, but it is restricted to data included in GOTaxDB (Schlicker, 2005). Furthermore, these searches are computationally expensive, and therefore, cannot be executed interactively. The query language employed by GOTaxExplorer permits easy selection of whole proteomes or proteins containing the same domains, but it was not designed for the selection of arbitrary sets of proteins. The Gene Functional Similarity Search Tool (GFSST) supports queries for functionally similar proteins, but restricts the user to either the human or the mouse proteome (Zhang *et al.*, 2006). The FuSSiMeG web service reports semantic similarities between GO terms annotated to two different proteins, but the results lack a combined score (Couto *et al.*, 2007).

In order to overcome the described limitations of available tools, we implemented the Functional Similarity Search Tool (FSST) and the Functional Similarity Matrix (FunSimMat). FSST forms part of the GOTax platform (http://gotax.bioinf.mpi-inf.mpg.de/) for comparative genomics. It is a stand-alone tool for performing functional similarity comparisons of user-defined sets and allows for including user-defined GO annotations (Schlicker *et al.*, 2007b). The database FunSimMat is a comprehensive resource providing direct access to several pre-computed semantic and functional similarity measures (Schlicker and Albrecht, 2008, 2010). It is accessible from a convenient online user front-end (http://www.funsimmat.de/) and through an XML-RPC interface. FunSimMat offers the semantic comparison of GO terms and several search options for functional similarity of proteins or protein families. The database contains precomputed functional similarity values for proteins and protein families from UniProtKB (Wu *et al.*, 2006a), Pfam (Finn *et al.*, 2006), and SMART (Letunic *et al.*, 2006). We implemented four different semantic similarity measures and apply them in the calculation of various functional similarity scores.

4.2 Extended Functional Similarity Scores

4.2.1 Introducing the *rfunSim* Score

In Section 2.3.3, we described our method for assessing the functional similarity of two gene products, the *funSim* score. It is based on the BMA approach, which calculates the similarity between two gene products A and B with sets of GO annotation GO^A and GO^B, respectively, as follows. For each term in GO^A, find the most similar term in set GO^B, and calculate the average of their similarities as $rowScore(A,B)$. Then, for each term in GO^B find the term with the highest similarity from set GO^A, and calculate the average as $columnScore(A,B)$. The $GOscore^{BMA}_{max}(A,B)$ is then defined as maximum of the $rowScore(A,B)$ and the $columnScore(A,B)$. The $GOscore^{BMA}_{max}(A,B)$ is calculated for the BP ontology (*BPscore*) and for the MF ontology (*MFscore*) (Section 2.3). The combined *funSim* score is calculated as follows:

$$funSim(A,B) = \frac{1}{2} \cdot \left[\left(\frac{BPscore}{\max(BPscore)} \right)^2 + \left(\frac{MFscore}{\max(MFscore)} \right)^2 \right], \tag{4.1}$$

where $max(BPscore)$ and $max(MFscore)$ denote the maximum possible scores for BP and MF, respectively. The *funSim* score ranges from 0 for completely unrelated gene products to 1 for gene products with identical functionality. In most cases, the *funSim* score is lower than the average of *BPscore* and *MFscore*. In order to obtain a more balanced score distribution, we define the *rfunSim* score for two gene products as (Schlicker *et al.*, 2007b):

$$rfunSim(A,B) = \sqrt{funSim(A,B)}, \tag{4.2}$$

which also ranges from 0 to 1, but its values are up to 25 % higher.

Despite being a simple transformation, the square root changes the performance of the score. In order to test how well the *funSim* and *rfunSim* scores differentiate between protein pairs without sequence similarity and orthologous protein pairs, we utilized the sets of Inparanoid orthologs (IO) and of protein pairs without significant sequence similarity (NSS) described in Section 3.2. For all protein pairs in both sets, the *funSim* and the *rfunSim* scores were computed and used for estimating the performance of predicting true positives (protein pairs in IO) and true negatives (protein pairs in NSS). The receiver operating characteristics (ROC) curve (Figure 4.1) was calculated and visualized using the ROCR package (Sing *et al.*, 2005) for the statistical computing environment R (http://www.r-project.org). It can be seen that the *rfunSim* score threshold is higher at given true positive and false positive rates.

The calibration error of a score measures how well the score coincides with the true class membership (Caruana and Niculescu-Mizil, 2004). Protein pairs with a score of 0.6 should belong to IO in 60% of the cases and to NSS in 40% of the cases, for example, and the calibration error measures the deviation form this ideal scenario. For calculating the calibration error, all protein pairs are ordered according to their score. Then, the pairs 1 - 100 are put into a bin and the percentage of true positives in this bin is calculated. Then, the mean prediction is calculated and the absolute frequency between observed true positive frequency and predicted positives gives the calibration error for this bin. This computation is repeated for protein pairs 2 - 101, 3 - 102 and so on. The final calibration error is the mean of the calibration errors of the single bins. For this test, the *funSim* and *rfunSim* scores are interpreted as probabilities of two proteins to be functionally similar. ROCR was used for calculating and plotting the calibration error of both scores (Figure 4.2). It becomes obvious that the *rfunSim* score has a smaller calibration error than the *funSim* score up to a value of approximately 0.75, and roughly equal thereafter. The results from the ROC curves and the calibration error analysis support the intuition that the *rfunSim* score gives better results.

For a more detailed analysis of the differences of the two scores, we performed a functional comparison between proteins from *Schizosaccharomyces pombe* (NCBI Tax-

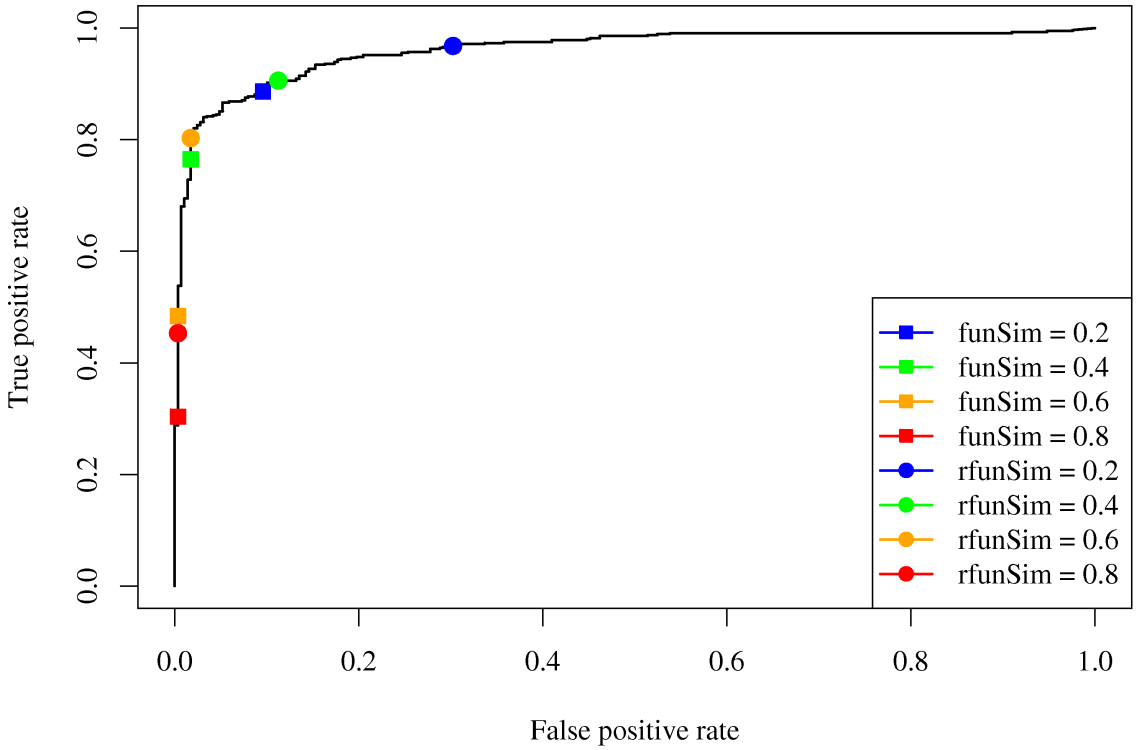

Figure 4.1: ROC curve for classifying protein pairs as belonging to the sets IO or NSS. The ROC curve shows the performance of the *funSim* score and the *rfunSim* score. The squares and circles mark *funSim* and *rfunSim* score thresholds, respectively. The symbols are colored according to the score threshold they represent: red corresponds to a threshold of 0.8, orange to a threshold of 0.6, green to a threshold of 0.4, and blue to a threshold of 0.2.

onomy id: 4986) and *Saccharomyces cerevisiae* (NCBI Taxonomy id: 4932) with the two scores. The proteins and their GO annotations were extracted from UniProtKB release 8.4. In the following, some examples for protein pairs with varying functional similarity illustrate the difference between the *funSim* and *rfunSim* scores. The stress response protein bis1 (UniProtKB accession: O59793) from *S. pombe* is annotated with the function "protein heterodimerization activity" (GO:0046982) and the process "response to stress" (GO:0006950). The high pH protein 2 (UniProtKB accession: P39734) from *S. cerevisiae* is involved in the same process but annotated with "protein binding" (GO:0005515) as function. The *funSim* score of these two proteins is 0.655 and the *rfunSim* score is 0.809. Since both proteins are involved in the same process and "protein heterodimerization activity" is a descendant of "protein binding" in the GO graph, the *rfunSim* score seems to more accurately reflect the true functional similarity.

The *S. pombe* protein glucan endo-1,3-alpha-glucosidase agn1 precursor (UniProtKB accession: O13716) is involved in "cell septum edging catabolism" (GO:0030995) and

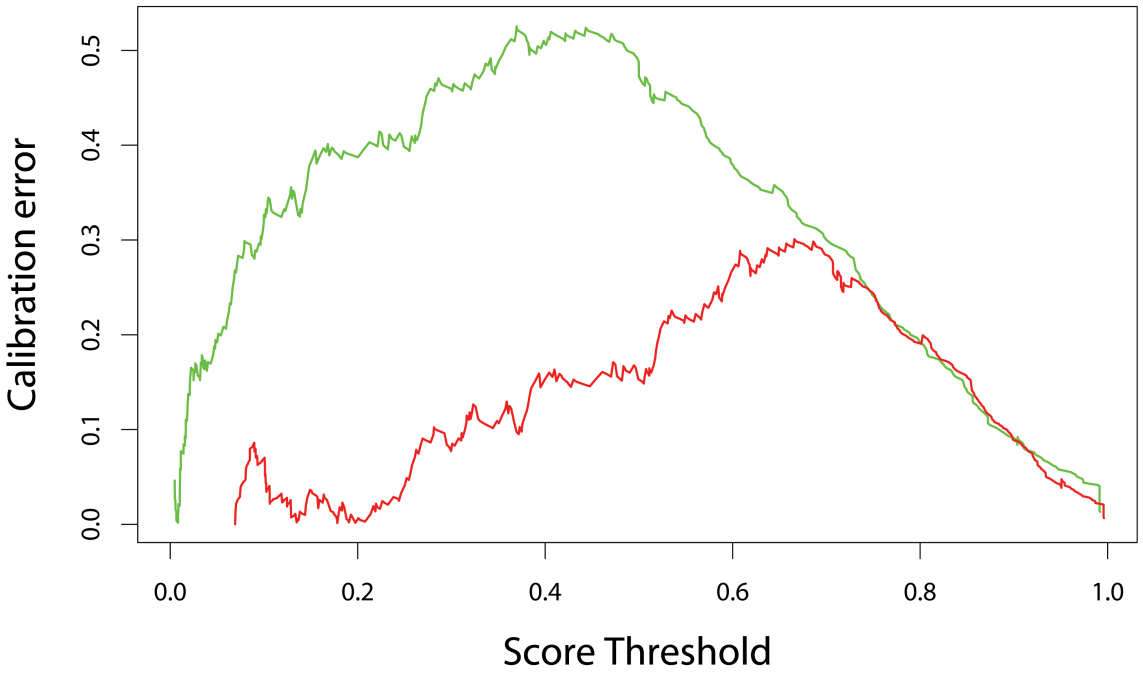

Figure 4.2: Calibration error for classifying protein pairs as belonging to the sets IO or NSS. The green curve shows the classification error for the *funSim* score and the red curve for the *rfunSim* score.

has "glucan endo-1,3-alpha-glucosidase activity" (GO:0051118). The protein EGT2 precursor (UniProtKB accession: P42835) from *S. cerevisiae* is annotated with the function "cellulase activity" (GO:0008810) and the process "cytokinesis" (GO:0000910). These two proteins have a *funSim* score of 0.364 and an *rfunSim* score of 0.603. Looking at the GO graph, it becomes evident that "cytokinesis" is an ancestor of "cell septum edging catabolism" and that the functions of the two proteins are related through the common ancestor "hydrolase activity, hydrolyzing O-glycosyl compounds" (GO:0004553). These close relationships between the MF terms and the BP terms annotated to the two proteins are more precisely captured by the *rfunSim* score.

Phosphatidylinositol-4-phosphate 5-kinase fab1 (UniProtKB accession: O59722) from *S. pombe* has "1-phosphatidylinositol-3-phosphate 5-kinase activity" (GO:0000285) in the process of "endocytosis" (GO:0006897). The 1-phosphatidylinositol-3-phosphate 5-kinase FAB1 (UniProtKB accession: P34756) from *S. cerevisiae* has the same function, but is annotated with three different processes, namely "phospholipid metabolism" (GO:0006644), "response to stress" (GO:0006950), and "vacuole organization and biogenesis" (GO:0007033). Assuming that the two proteins perform the same function, the *rfunSim* score of 0.711 seems more accurate than the *funSim* score of 0.505 although they are involved in completely unrelated processes.

4.2.2 Adding Cellular Component to the *funSim* Score

The *funSim* and the *rfunSim* scores calculate functional similarity based on BP and MF annotations. In order to assess the overall functional similarity of two annotated entities, however, it is also important to take into account in which cellular compartments they execute their specific functions. Therefore, we introduce the *funSimAll* and *rfunSimAll* scores that additionally integrate CC annotation. They are defined as follows:

$$\begin{aligned} funSimAll(A,B) &= \frac{1}{3}\cdot\Big[\Big(\frac{BPscore}{\max(BPscore)}\Big)^2+\Big(\frac{MFscore}{\max(MFscore)}\Big)^2+ \\ &\quad \Big(\frac{CCscore}{\max(CCscore)}\Big)^2\Big], \end{aligned} \tag{4.3}$$

$$rfunSimAll(A,B)=\sqrt{funSimAll(A,B)}. \tag{4.4}$$

Here, $max(BPscore)$, $max(MFscore)$, and $max(CCscore)$ denote the maximal score for biological process, molecular function, and cellular component, respectively. Both scores range between 0 for no similarity and 1 denoting maximum functional similarity.

4.3 Functional Similarity Search Tool (FSST)

The Functional Similarity Search Tool has been implemented for comparing user defined sets of annotated entities. FSST supports the computation of functional similarity scores based on an individual ontology, *BPscore*, *CCscore*, and *MFscore*, and of combined scores, *funSim*, *rfunSim*, *funSimAll*, and *rfunSimAll*. Its multi-threaded implementation takes advantage of symmetric multi-processing computers, decreasing runtime considerably. FSST is configurable using command line arguments and a configuration file.

As input to FSST, the user can provide a query file in plain text format containing the query entities with their GO annotations, and optionally, a database file with the same format defining the reference entities with their annotation. It is possible to either perform an all-against-all or an one-to-one comparison of query entities against reference entities. If no database file is given, the query entities are compared with each other. By default, the results are written to a text file containing all functional similarity scores for each pair consisting of one query and one reference entity. Additionally, FSST affords results of an all-against-all comparison in the form of a similarity matrix for one functional score. This permits to directly utilize the functional similarity results computed by FSST as input for further analysis tools, for instance, clustering programs. Since different applications might require distances rather than similarities, FSST is capable of transforming each score into a distance according to the formula:

$$dist_X(A,B)=1-X(A,B), \tag{4.5}$$

where X is one of the supported functional similarity scores.

FSST is distributed with an embedded Apache Derby database (http://db.apache.org/derby/) containing the sim_{Rel} values of all pairs of GO terms. However, it is possible to substitute the embedded database with any other database management system for which a JAVA JDBC compliant driver is available. The embedded version of Apache Derby has the advantage that it is administration free, and its deployment is completely hidden from the user. Moreover, FSST supports importing several semantic similarity measures into the database and for selecting one of the available measures for each program execution. Importantly, FSST is not restricted to GO, but can apply the implemented similarity scores to any ontology, provided that a database with semantic similarity values is available.

4.3.1 Functional Comparison of Proteins

With the help of FSST, we performed a functional comparison of all proteins from *Arabidopsis thaliana* (NCBI Taxonomy id: 3702) and *Saccharomyces cerevisiae* (NCBI Taxonomy id: 4932). UniProtKB release 8.4 contains 47,498 proteins from *A. thaliana*; out of these, 20,261 and 15,470 are annotated with MFs and BPs, respectively. From the 7,498 *S. cerevisiae* proteins in this UniProtKB release, 4,070 and 4,467 are annotated with MF and BP terms, respectively. Figure 4.3 shows the distributions of *BPscore*, *MFscore*, *funSim* score, and *rfunSim* score for the best hits of *A. thaliana* proteins. The NA column contains proteins for which the corresponding score could not be computed because of the lack of BP or MF annotations. More than half of *A. thaliana* proteins either have no BP or MF annotations.

Although most of the annotated proteins have a high functional similarity with an *S. cerevisiae* protein, some proteins have an *rfunSim* score between 0.4 and 0.6, which indicates only distant similarity. One such example is the cytokinin dehydrogenase 6 precursor (UniProtKB accession: Q9LY71) from *A. thaliana*. It is annotated with the process "stomatal complex morphogenesis" (GO:0010103) and the function "cytokinin dehydrogenase activity" (GO:0019139). The most similar protein from yeast is the dihydrofolate reductase (UniProtKB accession: P07807), which is annotated with the process "folic acid and derivative metabolism" (GO:0006760), and the functions "dihydrofolate reductase activity" (GO:0004146) and "protein binding" (GO:0005515). The two proteins have an *rfunSim* score of 0.47. The common oxidoreductase activity translates into an *MFscore* of 0.664, but they are part of completely unrelated processes (*BPscore* = 0.0).

4.4 Functional Similarity Matrix (FunSimMat)

FunSimMat has been implemented as application with three layers. The top-most layer consists of three user interfaces: a web front-end, an XML-RPC interface, and an REST-like interface. These interfaces have been implemented as a set of PHP scripts and run on

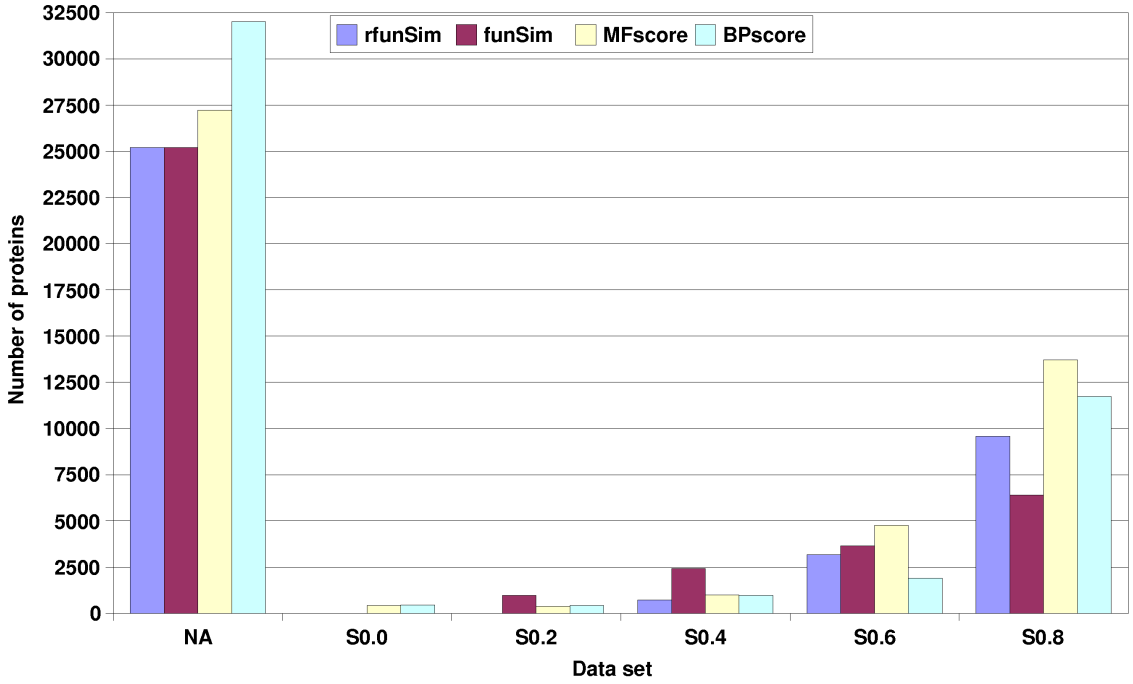

Figure 4.3: Functional comparison of *Arabidopsis thaliana* proteins with *Saccharomyces cerevisiae* proteins. Only the best hit (highest *rfunSim* score) for each yeast protein was taken into account for the score distributions. The NA column contains proteins for which the corresponding score could not be computed because of the lack of BP or MF annotation. The bins correspond to the following intervals of similarity values: S0.0: $[0.0, 0.2[$; S0.2: $[0.2, 0.4[$; S0.4: $[0.4, 0.6[$; S0.6: $[0.6, 0.8[$; S0.8: $[0.8, 1.0]$.

an Apache web server. The middle layer is a back-end server that has been implemented in Java 1.5. The lowest layer comprises a MySQL database containing all pre-calculated semantic and functional similarity values. Each user request sent to one of the interfaces is forwarded to the back-end server using XML-RPC. This back-end server queries the database and prepares the result set. The results are sent back to the calling interface, which subsequently processes and displays them.

4.4.1 Materials and Methods

Annotation classes

The initial release of FunSimMat contained more than 4.6-million proteins and protein families. Since functional similarity measures are symmetric, roughly ten trillion computations of functional similarity values would have been required for a complete all-against-all comparison. However, not every protein or protein family is annotated with a unique combination of GO terms. Therefore, we define an annotation class to be a spe-

cific, lexically sorted list of GO terms from one ontology. An annotation class can be identified by a unique accession number that remains stable between different database releases.

Each protein or protein family is assigned to three annotation classes that correspond to the annotated GO terms, one class for biological process (BPclass), one for molecular function (MFclass), and one for cellular component (CCclass). For example, the terms "mitochondrion inheritance" (GO:0000001) and "actin cortical patch assembly" (GO:0000147) constitute one BPclass. If c_1 and c_2 are two annotation classes, then all pairs of proteins A and B that belong to c_1 and c_2, respectively, have the same functional similarity value. This decreases the amount of computations required by several orders of magnitude. In addition to BPclasses, MFclasses, and CCclasses, we also define GO annotation classes (GOclasses). Each GOclass consists of one BPclass, one MFclass, and one CCclass. Theoretically, more than a trillion different GOclasses could be derived from the available BPclasses, MFclasses, and CCclasses. However, only a fraction of these occur in practice, which reduces the search space considerably when comparing one protein or protein family against the complete database. The definition of annotation classes as well as the mapping of proteins and protein families to annotation classes are available for download on the FunSimMat website (http://www.funsimmat.de).

As of release 3.1, the FunSimMat database contains GO annotations from UniProtKB and the Gene Ontology Anntotation (GOA) project (Barrell *et al.*, 2009). The addition of annotations from GOA almost doubled the number of available mappings between proteins and GO terms and thereby provides a significantly higher coverage as well as an improved functional characterization of proteins sharing similar functions. This is signified by the number of annotation classes in release 3.1, which is four times higher than in the previous release: 47,538 BPclasses, 59,814 MFclasses, 18,753 CCclasses, and 151,151 GOclasses. However, many of these classes differ by a single term only, which results in a very high functional similarity between them.

In order to exploit this relatedness, we introduce hierarchically structured networks of annotation classes for BP, MF, and CC. In these networks, nodes represent annotation classes and two classes, c_1 and c_2, are connected by an edge if the following two conditions are satisfied: (i) all terms from c_1 are contained in c_2, and (ii) c_2 contains exactly one additional term. Annotation classes consisting of solely one term constitute the source nodes in these networks. Nodes without descendants represent the most specific classes and are defined as annotation superclasses. The newly established hierarchies of annotation classes enable refining comparisons of a specific protein or protein family with a list of proteins or families. The user can restrict the query to superclasses, and thereby, concentrate on the largest functional differences. By including all annotation classes in a subsequent query, it is possible to obtain a comprehensive overview for identifying smaller differences in functional similarity.

Data Sets

The FunSimMat release 3.1 contains almost 8.4-million proteins from UniProtKB (release 15.3) and approximately 26.9-million GO annotations of proteins extracted from UniProtKB and from GOA (release of May 2009). Additionally, FunSimMat includes over 10,000 Pfam families (release 23) and 720 SMART families (from InterPro release 20). The annotations of protein families with GO terms were derived from the pfam2go and smart2go mapping files (both downloaded in April 2009 from http://www.geneontology.org/). The database also contains 19,481 entries from OMIM (downloaded on 10 June 2009). In total, release 3.1 of the FunSimMat database is 326 GB in size, which is almost four times the size of the previous release. The number of annotation classes is small in contrast to the number of proteins in the database. Therefore, we anticipate that our approach will scale well with the growing number of proteins and annotations that can be expected in the upcoming years. We also intend to update the databases every three months, which takes about two days.

Semantic Similarity Measures

In FunSimMat, we implemented four different semantic similarity measures. These measures are based on the *IC* of a GO term (Section 2.2.2). The more specific a GO term is, the smaller is its probability and the higher its information content. The probability of a GO term is defined as its relative frequency in UniProtKB (Equation 2.2). Based on the calculated *IC* values, we compute for each GO term pair Resnik's measure (Equation 2.4), Lin's measure (Equation 2.7), the sim_{Rel} score (Equation 2.8), and a similarity based on Jiang and Conrath's distance measure (Equation 2.6).

Functional Similarity Measures

We implemented several functional similarity measures for proteins and protein families that are based on the DAG structure of GO or on the semantic similarity measures. First, we implemented the following four different groupwise functional similarity approaches (Section 2.3.2): *simUI*, *simGIC*, *TO*, and *NTO*. Additionally, several pairwise functional similarity scores were implemented that are based on the semantic similarity scores (Section 2.3.1): $GOscore_{avg}$ (Equation 2.12), $GOscore_{max}$ (Equation 2.13), $GOscore^{BMA}_{avg}$ (Equation 2.16), and $GOscore^{BMA}_{max}$ (Equation 2.17). The latter four scores can be computed using either of the four semantic similarity measures. For all measures, the lowest similarity value is 0, and the maximum similarity is 1, except for scores calculated with Resnik's measure, which has no upper bound. For each pair of proteins or protein families, three different functional measures can be computed: one for biological process (*BPscore*), one for molecular function (*MFscore*), and one for cellular component (*CCscore*). In order to allow for assessing the full functional similarity of entity pairs, we

also implemented the *funSim* (Equation 2.21), *rfunSim* (Equation 4.2), *funSimAll* (Equation 4.3), and *rfunSimAll* (Equation 4.4) scores.

4.4.2 Query Options

FunSimMat offers several query options. The first option is the semantic all-against-all comparison of GO terms contained in an input list provided by the user. The GO terms have to be entered using their accession numbers, for example GO:0000001, and the results table contains the computed semantic similarity values. The second option is the comparison of an individual query entity with a list of proteins or protein families. The query entity can be a protein from UniProtKB, a protein family from Pfam or SMART, or an OMIM disease. FunSimMat supports several alternatives of how to compile the list of proteins or protein families. The simplest one is to enter the corresponding accession numbers into the query form of the website. It is also possible to upload a text file containing the accession numbers. Moreover, the query entity can be compared to all proteins associated with an OMIM entry by entering the accession number of the disease. Alternatively, users may select all proteins and protein families from a certain taxon by entering the corresponding NCBI Taxonomy identifier. The query form also contains a drop-down box to quickly choose from common taxa. It is also possible to compare the query entity to the whole database. The computation results contain the functional similarity scores between the query entity with every protein or protein family from the list. If the user selected a taxon or the whole database, the results table contains the annotation classes corresponding to the selected proteins or protein families. By clicking on one annotation class, the user can obtain a list of selected proteins or protein families belonging to that class.

The third query option is the definition of a functional profile. A functional profile consists of a list of GO terms from one of the three ontologies, BP, MF, or CC. This functional profile is treated as an annotation class and compared to a list of proteins or protein families. The user can either choose a taxon, as in the case above, or compare the profile with the whole database. This helps finding proteins and protein families that are similar to a prototype protein the user is interested in. Similar to the second query option, the results table contains the comparison between the functional profile and the annotation classes. The list of selected proteins or protein families belonging to an individual class can be accessed by clicking on the class identifier.

As fourth query option, we implemented our new method for prioritizing disease gene candidates based on functional similarity (Chapter 6). The MedSim approach exploits GO annotations of genes or proteins known to be involved in a disease of interest and ranks candidate genes or proteins by functional similarity in a two step process. First, GO terms are automatically transferred from proteins cross-referenced to OMIM diseases by UniProtKB to the corresponding OMIM entry. Second, the list of candidates is ranked by functional similarity between the candidate proteins and the disease of interest. The

higher the functional similarity is, the more likely is this candidate involved in the disease of interest. This method was implemented in FunSimMat by mapping each disease to the annotation classes matching the transferred GO terms, and calculating all functional similarity values between human proteins and the diseases. This approach makes it possible to utilize FunSimMat for the fast prioritization of a list of candidates by entering the OMIM accession number of the disease of interest and the list of UniProtKB accessions of the candidate proteins.

4.4.3 Web Front-End

The web front-end offers HTML forms for all query options offered by FunSimMat. The results are displayed in a table (Figure 4.4), and can be downloaded as tab-delimited text file or printed. Many options are available for customizing the results table. It can be sorted according to anyone column by clicking on the corresponding column header. Initially, the score values are colored with a gradient from white for low similarity to blue for high similarity. This gradient can be changed to red or green. Additionally, it is possible to hide and show specific columns or groups of columns, for example, all biological process scores at once. However, it is important to note that the features for changing the color gradient and for hiding columns are only available if JavaScript is enabled. The first two columns of the table contain the GO, UniProtKB, Pfam, or SMART accessions linked to the respective source database, or the annotation class accessions. The annotation class accessions are linked to a page listing all proteins and protein families from the input list that belong to this annotation class along with their complete GO annotation. Tooltips containing the GO annotation of proteins or protein families are shown when the mouse hovers over an accession.

The results of each query are stored for two days allowing researchers to continue with an analysis. Additionally, the results page contains a link that allows for modifying a previous query. After clicking on this link, the query form is loaded with all the information that was previously entered for performing the query. This link does not expire and can be bookmarked or shared with colleagues. This allows for either modifying earlier queries or for re-running them, for instance, after a database update. More details can be found in the help section on the website.

4.4.4 XML-RPC Interface

The extensible markup language remote procedure call (XML-RPC) protocol provides a means for accessing remote services and programs over a network. The XML-RPC interface allows for automatically querying FunSimMat over the Internet and for processing the results. This interface has been implemented using PHP and is available at http://funsimmat.bioinf.mpi-inf.mpg.de/xmlrpc.php. It provides the same query options as the web front-end. For instance, in order to semantically compare a list of GO terms,

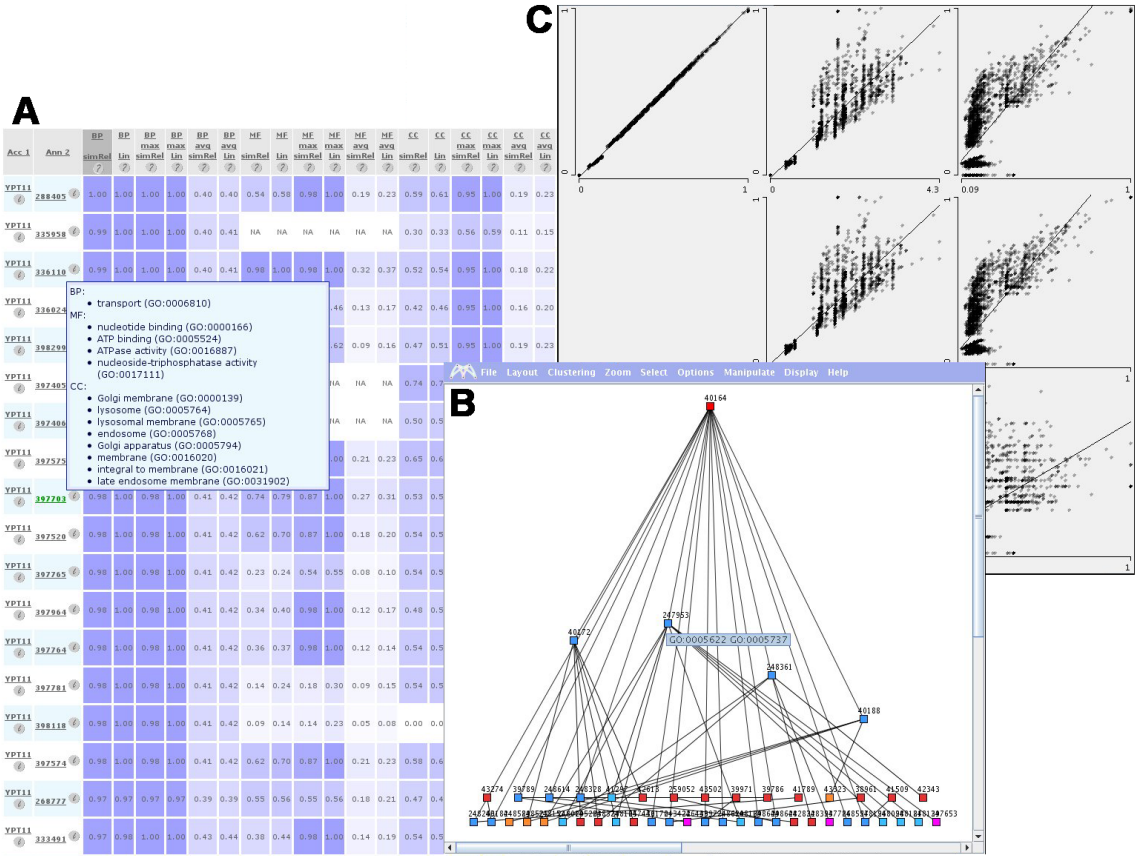

Figure 4.4: Different visualization options for a result set provided by FunSimMat. The figure shows some of the results obtained by the functional comparison of GTP-binding protein YPT11 (UniProtKB accession: P48559) with GO annotation superclasses of human proteins. (A) The results table lists all functional similarity scores of the query protein with different GOclasses. Each table cell is colored by a gradient; white color represents no similarity and blue color high similarity. The popup box gives all GO terms for the GOclass 397703. (B) Visualization of some CCclasses contained in the results using the Medusa program. The classes were clustered using the k-means algorithm with k set to 20 and placed by applying a hierarchical layout. The nodes are colored according to cluster membership. (C) Scatter plots that compare biological process similarities obtained by different semantic similarity measures obtained with the Mondrian program. The three plots in the first row show, for example, that the results obtained with sim_{Rel} (Equation 2.8) are strongly correlated with Lin's similarity (left, Equation 2.7), less correlated with Resnik's similarity (center, Equation 2.4), and only weakly correlated with scores computed using Jiang and Conrath's similarity (right, Equation 2.6). The straight lines in the scatter plots are least-squares regression calculated by Mondrian.

the function *Semantic.getSemSims*() can be called, which accepts the accession numbers of the GO terms as comma-delimited list. The returned response contains all rows of the

results table in the form of an array. The detailed documentation of this interface can be found in the help pages on the FunSimMat website.

4.4.5 RESTlike Interface

The web front-end allows for easily performing manual queries and the XML-RPC interface permits to integrate FunSimMat into automated analysis pipelines and workflows. The additional RESTlike interface provides the same query options as the other two front-ends, but all query parameters are specified inside an URL. This way, web links for querying FunSimMat can easily be added to other web sites and services. A detailed description of the available URL parameters is provided in the online documentation of FunSimMat.

4.4.6 Tools for Visualizing Result Sets

The basic type of visualization for result sets employed by FunSimMat is to provide a summary using a table (Figure 4.4). This table provides special means for easily investigating the similarity between a pair of GO terms, proteins, or protein families in detail. However, if the query result set is large, a visual analysis may prove beneficial for quickly obtaining an overview. Therefore, we offer two additional options for displaying and analyzing FunSimMat results (Figure 4.4).

First, the tool Mondrian facilitates a comprehensive statistical analysis of the result set (Theus, 2002). In particular, it has the functionality of drawing different types of plots, for instance, scatter plots, bar charts, box plots, and histograms. Various plots can be opened simultaneously and be compared directly. For instance, they can be used to investigate the correlation between different functional similarity scores in a specific result set. Data points selected in one plot are highlighted instantly in all other plots, which aids in studying an interesting subset of results from various perspectives.

Second, the tool Medusa visualizes the hierarchical relationships between the annotation classes contained in the result set from a functional comparison (Hooper and Bork, 2005). Users can apply different layout and clustering algorithms for discovering relationships between annotation classes in the result set. Furthermore, it is possible to search for all classes that contain specific GO terms.

The original implementations of both tools were modified to enable their deployment using Java Web Start. Both are started by clicking on the corresponding link on the results page, and then, the result set is loaded automatically. Plots generated by both tools can be saved in various bitmap and vector image formats.

4.5 Conclusions

Functional similarity measures are used in many different applications like gene clustering, protein-protein interaction prediction and validation, and disease gene prioritization. However, existing tools have several limitations. They usually support only one functional similarity measure and calculation of similarities is rather time-consuming. This makes it difficult to test different scores for an application. Moreover, there is no comprehensive resource of functional similarity values available. In consequence, users have to compute these values themselves using one of the existing tools or their own implementation.

In order to remedy these problems, we have implemented a tool for functional similarity searches, FSST, and a database of precomputed functional similarity values, FunSimMat. The stand-alone tool FSST allows for comparing user-defined sets of entities and for integrating custom GO annotations. Its multi-threaded implementation considerably decreases runtime by taking advantage of modern multi-processor computers. While the embedded database simplifies deployment significantly, the option to use a different database management system provides maximal configurability. Notably, the database Pfam deploys FSST for calculating functional similarities between protein families. Moreover, Merkl and Wiezer used FSST for pairwise functional comparisons of prokaryotic genomes (Merkl and Wiezer, 2009).

The database FunSimMat contains functional similarities between more than 8.4-million proteins and protein families and integrates GO annotation from UniProtKB and GOA. The web front-end provides several query options for flexible, simple and fast retrieval of the similarity values. All query results are accessible online for two days and may be downloaded in tab-delimited files, which facilitates their use in many applications. The additional XML-RPC and RESTlike interfaces make it possible to automatically query and link to FunSimMat. This greatly supports the integration of FunSimMat and the use of functional similarity in many existing and new data analysis pipelines and tools. The visual analysis tools afford new innovative possibilities to analyze increasingly large sets of functional similarity results. Additionally, FunSimMat provides a new way of performing rapid functional similarity searches within large protein databases. Since the original publication of FunSimMat, almost 2-million queries have been submitted by about 300 users to the FunSimMat server.

Chapter 5

Functional Evaluation of Domain-Domain Interactions and Human Protein Interaction Networks

Large amounts of protein and domain interaction data are being produced by various experimental high-throughput techniques and computational approaches. These methods are not free of errors and produce a considerable number of false positive interactions. Therefore, it is necessary to assess the confidence of the interactions and perform a quality ranking.

In this chapter, we present the application of our functional similarity measures to validate domain and protein interactions. We derive useful confidence score thresholds for dividing predicted domain interactions into subsets of low and high confidence. This work was first presented at the German Conference on Bioinformatics (GCB) in 2006 (Schlicker *et al.*, 2006b), and published subsequently in the journal Bioinformatics (Schlicker *et al.*, 2007a).

5.1 Introduction

Large amounts of protein-protein interaction (PPI) data for different species have been generated with the help of experimental high-throughput techniques (Sharan and Ideker, 2006). These data are now mined for new information on the functions and relationships of proteins (Bork *et al.*, 2004). In particular, the large-scale prediction of human protein interaction networks was supported by different bioinformatics methods that are mainly based on the homology of protein sequences (Huang *et al.*, 2004; Lehner and Fraser, 2004; Brown and Jurisica, 2005; McDermott *et al.*, 2005; Persico *et al.*, 2005; Rhodes *et al.*, 2005). Recently, manually curated literature data and four large-scale yeast two-hybrid interaction maps have become available, which greatly increased available data of the human interactome (Goehler *et al.*, 2004; Rual *et al.*, 2005; Stelzl *et al.*, 2005;

Lim *et al.*, 2006; Mishra *et al.*, 2006). Nevertheless, experimental coverage of the human interactome is still low in contrast to predicted data. Domain-domain interactions (DDI) are a commonly utilized tool for predicting protein interaction networks (Wojcik and Schachter, 2001; Deng *et al.*, 2002; Liu *et al.*, 2005b; Rhodes *et al.*, 2005). For this purpose, bioinformatics methods have been devised for predicting sets of DDIs (Ng *et al.*, 2003; Liu *et al.*, 2005b; Riley *et al.*, 2005) that supplement experimental DDI sets derived from 3D structure data (Finn *et al.*, 2005; Stein *et al.*, 2005).

As mentioned in the previous chapters, the functional vocabulary provided by the GO consortium (Ashburner *et al.*, 2000) is commonly used to annotate proteins and protein domains with processes and functions. This annotation allows for assessing the functional similarity of proteins and domains. In two studies conducted by Lin *et al.* and Lu *et al.*, the usefulness of different features, which ranged from expression profiles to functional relationships between genes, for predicting PPIs was evaluated (Lin *et al.*, 2004; Lu *et al.*, 2005). Both groups concluded that functional similarity based on GO annotation leads to high accuracy in predicting PPIs. Wu *et al.* introduced new similarity measures between GO terms and proteins, and applied these measures to create a predicted PPI network and to evaluate genome-scale datasets (Wu *et al.*, 2006b). Later, Guo *et al.* assessed the applicability of GO-based similarity measures to human regulatory pathways (Guo *et al.*, 2006). They showed that the functional similarity between two proteins decreases as their distance within the same regulatory pathway increases.

One problem with previously applied GO-based similarity measures is that they do not account for the frequent annotation of proteins or protein domains with multiple GO terms or that they simply average over all annotations. To address this issue, we used our functional similarity measure that explicitly deals with this functional diversity ($GOscore^{BMA}_{max}$, Section 2.3.1). The measure is applied to rank interaction networks and the corresponding prediction methods based on the overall functional similarity of the interacting proteins or domains. We further compared sets of experimentally derived DDIs with sets of predicted DDIs using our GO similarity measure and subsequently derive confidence score thresholds separating low- and high-confidence subsets of predicted DDIs. In addition, we utilized our measures to analyze experimental and predicted networks of human protein interactions.

5.2 Materials and Methods

5.2.1 Domain Interaction Networks

Two sets of experimental interactions between Pfam-A domains (Finn *et al.*, 2006), taken from iPfam (Finn *et al.*, 2005) and the database of 3D interacting domains (3did, Stein *et al.*, 2005) were compared to three sets of predicted DDIs. The first predicted set was taken from InterDom, a database of putatively interacting domains compiled from different data sources (Ng *et al.*, 2003). The other two sets were extracted from the publications

Table 5.1: Number of Pfam-A domains and their interactions in the different DDI datasets. The columns for biological process ('BP') and molecular function ('MF') contain the percentage of interactions whose interacting domains are both annotated with GO and could be used for calculating the *BPscore* and *MFscore* (see Section 5.2.3), respectively.

Dataset	Number of domains	Number of interactions	BP (%)	MF (%)
iPfam	2,145	3,046	52.07	56.30
3did	2,247	3,034	49.51	54.19
InterDom	3,535	29,957	27.07	37.64
LLZ	1,980	5,806	28.01	30.99
DPEA	1,026	1,812	37.14	40.12

by Liu *et al*. (LLZ, Liu *et al.*, 2005b) and by Riley *et al*. (DPEA, Riley *et al.*, 2005). Their bioinformatics approaches are both based on the expectation-maximization algorithm for predicting domain interactions devised by Deng *et al.* (2002). More than 50 % of all DDIs in iPfam and 3did are homodimeric self-interactions. InterDom and LLZ do not predict self-interactions between Pfam-A domains in contrast to DPEA.

The three DDI prediction methods assign a confidence score (CS) to each interaction and rank predicted DDIs according to this score. InterDom infers DDIs through integrating different data sources and calculates the CS of an interaction based on its support from these sources. LLZ and DPEA derive the CS from maximum-likelihood estimates, and we utilize the probability λ and the log-odds score E provided by LLZ and DPEA, respectively, as CS. For our analysis, all interactions between Pfam-A domains contained in the datasets were taken into account, including self-interactions and intra-chain interactions from iPfam and 3did.

The pfam2go file from the GO web site contains annotations of Pfam-A domains with GO terms. This mapping is automatically derived from the manually curated annotations of InterPro entries with GO terms (Camon *et al.*, 2003). An InterPro entry is annotated with a GO term if the term matches the function or the process of this entry and all proteins containing this domain are annotated with this GO term. All available GO terms including all evidence codes were used for annotating Pfam-A domains using the pfam2go mapping (downloaded on 7 July 2005). The number of domains and DDIs in each dataset is summarized in Table 5.1. We additionally mapped the annotations from the pfam2go file to more generic functions and processes from the generic GO-slim set (http://www.geneontology.org/GO.slims.shtml). This allows for identifying GO terms that occur more frequently than others in the dataset. In particular, we were interested in determining whether domains annotated with "protein binding" are enriched in the experimental datasets compared to the predicted ones.

5.2.2 Protein Interaction Networks

In addition to the DDI sets, we analyzed six predicted sets of human PPIs named Bioverse (McDermott *et al.*, 2005), HiMAP (Rhodes *et al.*, 2005), HomoMINT (Persico *et al.*, 2005), Sanger (Lehner and Fraser, 2004), OPHID (Brown and Jurisica, 2005), and POINT (Huang *et al.*, 2004). For Bioverse, HiMAP and Sanger, we derived subsets of core interactions with high confidence. The Bioverse-core set contains very reliable interactions based on a sequence similarity threshold of at least 80 % between human and the homolog of the source species (Yu *et al.*, 2004). Interactions in HiMAP-core have a large likelihood ratio (Rhodes *et al.*, 2005), and Sanger-core comprises only predictions with the highest experimental support (Lehner and Fraser, 2004). Furthermore, we assembled five consensus sets named ConSet*n* that consist of protein interactions contained in at least *n* predicted datasets, with *n* ranging from 2 to 6.

As experimental PPI datasets, we downloaded the manually curated human protein reference database (HPRD, Mishra *et al.*, 2006), release of 13 September 2005, and two yeast two-hybrid (Y2H) maps, named "Vidal" (Rual *et al.*, 2005) and "Wanker" (Stelzl *et al.*, 2005). They became available after the six predicted human networks had been published. By merging the two Y2H sets, a combined dataset, Vidal & Wanker, was created. Further experimental PPIs were extracted from the published networks of direct and indirect interaction partners for ataxins (ATX, Lim *et al.*, 2006) and huntingtin (HTT, Goehler *et al.*, 2004). The HTT and ATX networks consist of Y2H and literature-derived interactions, and the latter additionally contains interologs-based PPIs. Therefore, we split the two datasets according to the source of the interactions into the sets ATX-/HTT-Y2H, ATX-/HTT-literature, and ATX-interologs. Since the interactions in the ATX-interologs set have been derived by mapping interologs, we regard it as another set of predicted PPIs. The diverse gene and protein accession numbers used in the various PPI datasets were mapped to NCBI Entrez gene identifiers for enabling a comparison (Maglott *et al.*, 2005). The mapping of Entrez gene identifiers to GO annotations was obtained from NCBI (ftp://ftp.ncbi.nih.gov/gene/DATA/).

An additional set of PPIs was compiled using the proteins that underlie iPfam DDIs. This set was annotated using two different sources, that is, with the GO annotations of proteins in the UniProtKB (Wu *et al.*, 2006a) release 5.4 (IUP-set) and with the GO annotations of the protein domains in the pfam2go file (IPG-set) using Pfam release 17.0 (Finn *et al.*, 2006). In case of the IPG-set, only the annotations of the interacting domains were taken into account. Self-interactions were excluded from both the IUP-set and the IPG-set, whereas they were not removed from the other PPI datasets.

5.2.3 Functional Similarity Measures

For calculating the semantic similarity between two single GO terms, we utilized the sim_{Rel} measure as defined in Equation 2.8 (Section 2.2.2). We defined the probability of a term as its relative frequency of occurrence in a set of annotated gene products (Equa-

tion 2.3). We used the GO annotations of all proteins in UniProtKB release 5.4 for the calculation of term frequencies. Furthermore, we applied the $GOscore_{max}^{BMA}$ (Equation 2.17) that is based on the BMA approach for calculating the functional similarity of two domains or proteins with respect to an individual ontology. We refer to this $GOscore_{max}^{BMA}$ simply as $MFscore_{max}^{BMA}$ or $BPscore_{max}^{BMA}$ in case of MF or BP, respectively. One major aspect of this score is that it allows for comparing gene products with multiple functions. This property is especially important when comparing GO annotations of domains because they occur in diverse proteins involved in different processes.

Since this functional similarity measure requires that either both interacting proteins or both interacting domains are annotated with GO terms, the functional similarity analysis considers only interactions for which GO annotations are available for both interacting partners (see columns 4 and 5 in Table 5.1 for DDI sets). Therefore, the analyzed interaction datasets differ concerning the $BPscore_{max}^{BMA}$ and the $MFscore_{max}^{BMA}$. Self-interactions do not necessarily receive a high $GOscore_{max}^{BMA}$ because the definition of the semantic similarity measure takes into account how generic the GO annotation term is. For instance, a DDI of a self-interacting domain that is annotated with the general term "binding" will receive a low $MFscore_{max}^{BMA}$.

5.3 Results and Discussion

5.3.1 Comparing Confidence Scores for Domain Interactions

InterDom, LLZ, and DPEA utilize different bioinformatics methods that exploit diverse data sources for predicting DDIs. In order to gain insight into the similarity and the quality of these predictions, we compared the three sets of DDIs with each other and with the experimentally derived sets iPfam and 3did. Table 5.2 summarizes the overlap between the three predicted DDI sets with respect to the contained Pfam-A domains as well as regarding the predicted interactions. LLZ and DPEA share many Pfam-A domains and predicted DDIs with the much larger InterDom dataset. In contrast, the overlap between LLZ and DPEA is much smaller. Only 51.9 % of the domains in DPEA are also contained in LLZ and only 26.9 % of the domains in LLZ appear in an interaction predicted by DPEA. Regarding domain interactions, the DPEA set has 32.9 % of its interactions in common with the LLZ set, but only 10.6 % of the LLZ interactions are contained in the DPEA set. Considering only interactions with both interacting domains contained in each dataset reveals that 72.7 % of the interactions predicted by LLZ and 89.3 % of the DPEA interactions are present in the InterDom set. Overall, DPEA shows a better agreement with InterDom interactions than LLZ, although the complete DPEA set is much smaller than the complete LLZ set.

Figure 5.1 gives an overview of the overlap of the experimental interactions contained in iPfam and 3did and the three sets of predicted interactions InterDom, LLZ, and DPEA. 11.9 % of the DDIs predicted by DPEA are confirmed by iPfam or 3did, whereas only

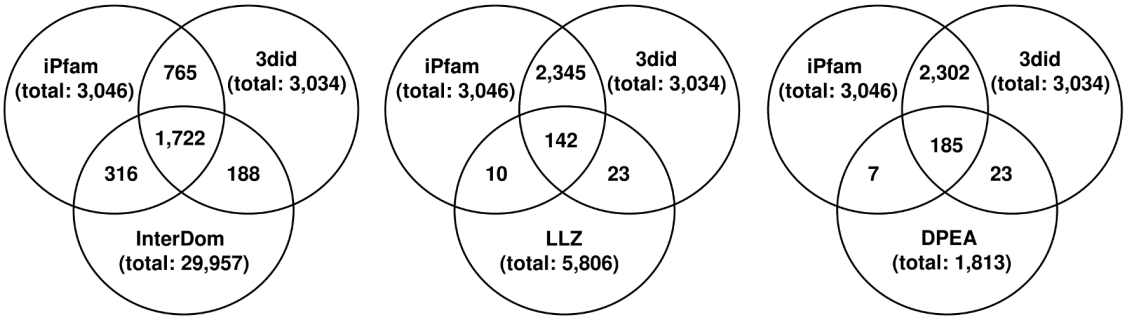

Figure 5.1: Overlap of the predicted DDI datasets with the two sets containing experimental Pfam-A domain interactions.

Table 5.2: Pairwise overlap of the datasets InterDom, LLZ and DPEA with regard to Pfam-A domains and their predicted interactions. Each number refers to the percentage of domains or interactions in the row dataset that are also contained in the respective column dataset. Percentages in parentheses give the number of DDIs shared between two datasets in relation to the overall number of DDIs with interacting domains contained in both datasets.

	Pfam-A domains (%)			Domain-domain interactions (%)		
Dataset	InterDom	LLZ	DPEA	InterDom	LLZ	DPEA
InterDom	100.0	44.4	25.1	100.0 (100.0)	11.4 (19.3)	4.8 (23.2)
LLZ	79.3	100.0	26.9	58.8 (72.7)	100.0 (100.0)	10.6 (60.8)
DPEA	86.5	51.9	100.0	78.9 (89.3)	32.9 (62.2)	100.0 (100.0)

7.4 % and 3.0 % of the DDIs predicted by InterDom and LLZ, respectively, are in common with iPfam or 3did. Thus, DPEA appears to be the best of the three prediction methods.

The CS (Figure 5.2) and the rank assigned to experimentally observed domain interactions can serve as additional criteria for prediction quality. Although DDIs contained in iPfam and 3did are assigned top ranks by all three prediction methods, further analyses showed only weak correlations between ranks of different prediction methods. Nevertheless, detailed results indicated that if a DDI from iPfam is predicted by two different computational methods, it is assigned a good rank by at least one of the prediction methods. This suggests that all methods are able to detect correct domain interactions.

5.3.2 Background Distributions and Randomized Domain Networks

Background distributions for $BPscore_{max}^{BMA}$ and $MFscore_{max}^{BMA}$ were obtained by computing the functional similarity of all possible Pfam-A domain pairs. Most of the domain pairs

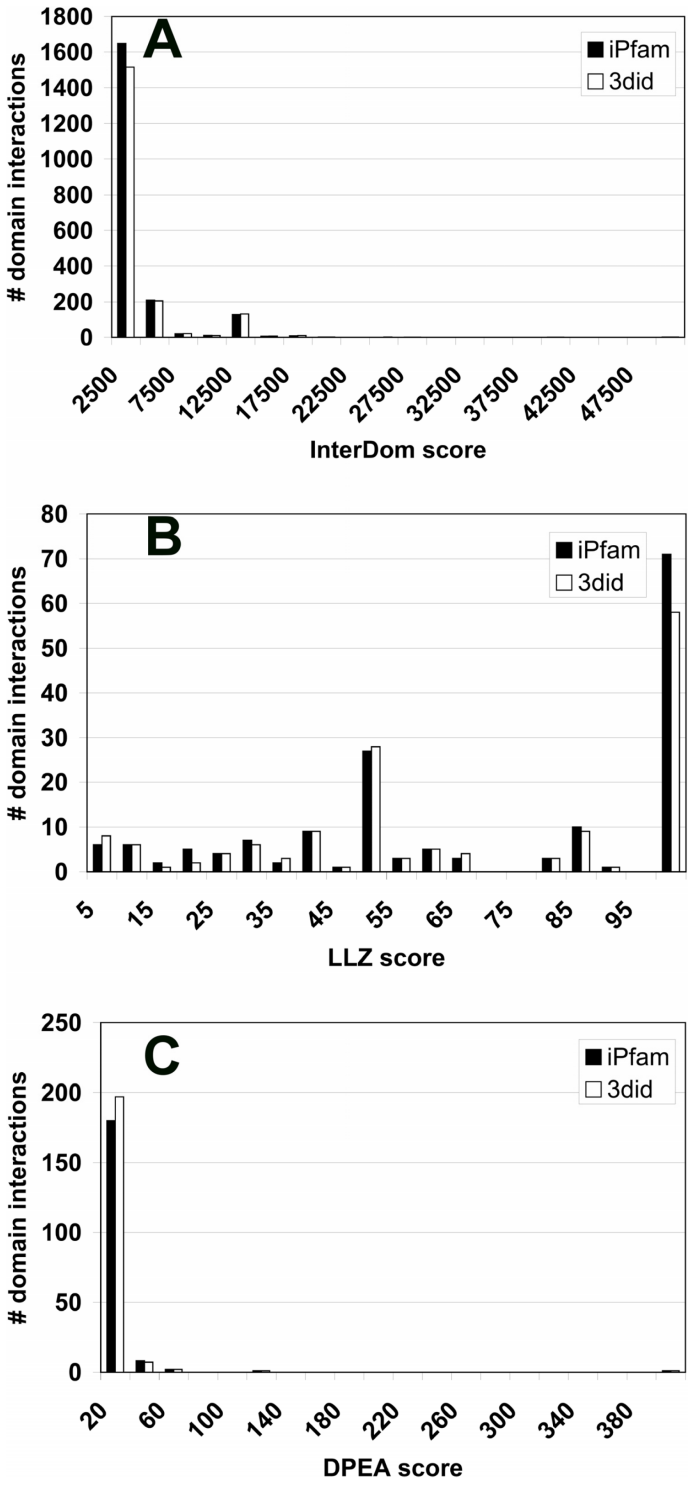

Figure 5.2: Distributions of CSs assigned to experimentally verified domain-domain interactions by (A) InterDom, (B) LLZ and (C) DPEA. All methods assign a CS to their predictions. These scores are derived differently by each method, but In all cases, a higher score translates into higher confidence.

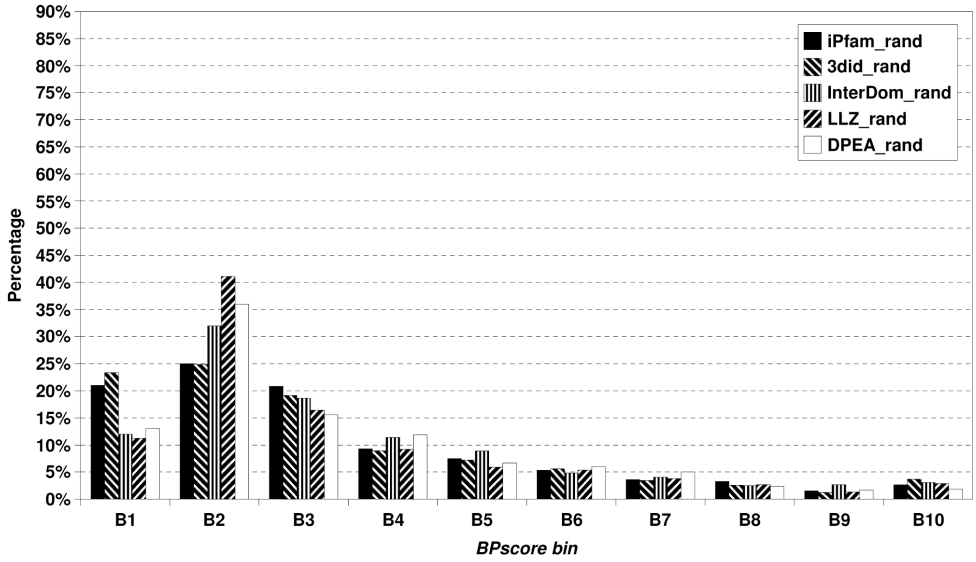

Figure 5.3: Distributions of the $BPscore_{max}^{BMA}$ values for the randomized DDI datasets (iPfam, 3did, InterDom, LLZ and DPEA). The $BPscore_{max}^{BMA}$ bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

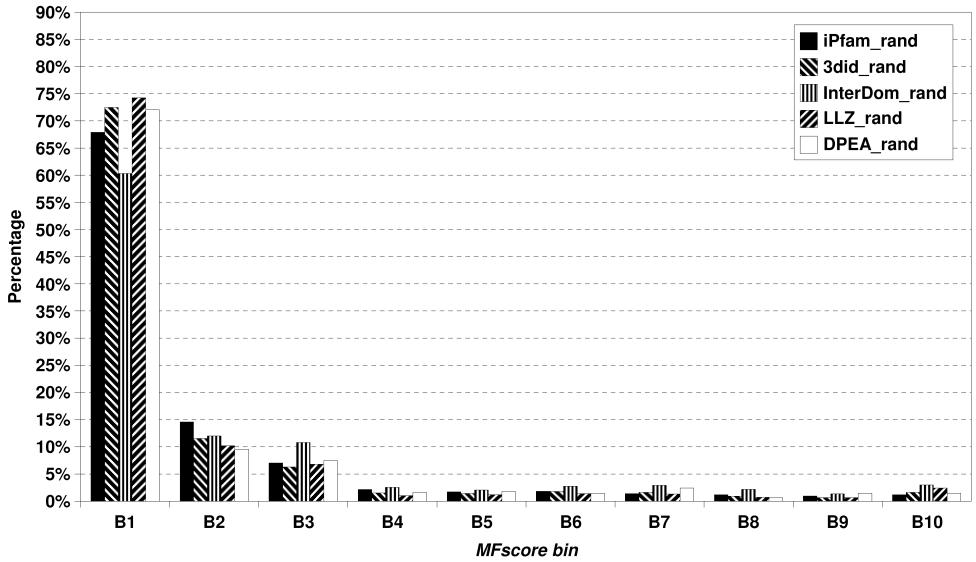

Figure 5.4: Distributions of the $MFscore_{max}^{BMA}$ values for the randomized DDI datasets (iPfam, 3did, InterDom, LLZ and DPEA). The *MFscore* bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

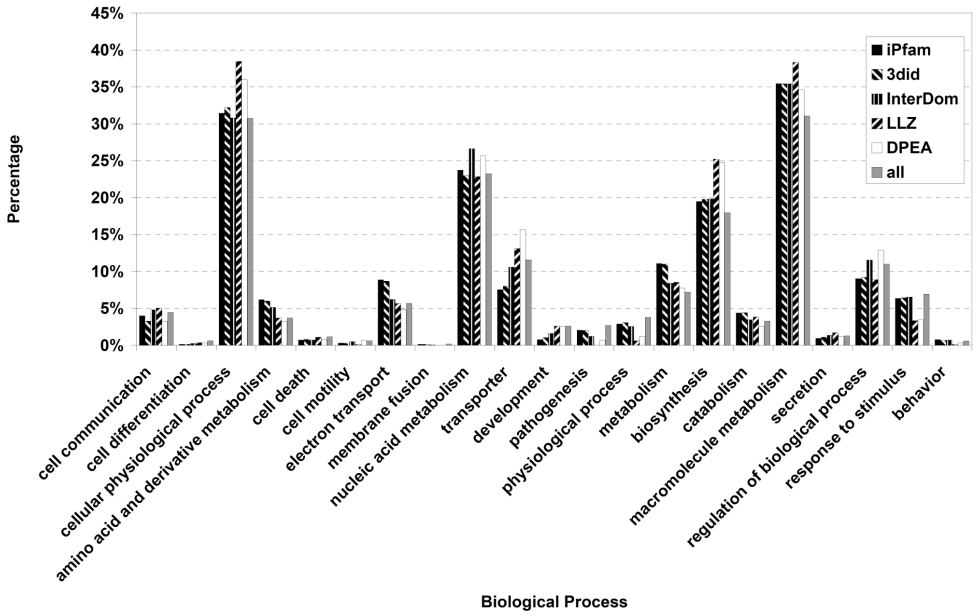

Figure 5.5: Distributions of GO-slim BP terms in the various DDI datasets. The single BP terms are shown on the x-axis, and the y-axis gives the percentage of domains in the different DDI sets annotated with the respective BPs.

have dissimilar molecular functions, resulting in a low $MFscore_{max}^{BMA}$, which is also reflected by a low mean and median of about 0.1 and 0, respectively. For $BPscore_{max}^{BMA}$, the distribution is similar, although, in comparison with the $MFscore_{max}^{BMA}$, more domain pairs have a similarity between 0.1 and 0.2 and fewer pairs a score below 0.1. This finding is also reflected by the increased $BPscore_{max}^{BMA}$ mean and median of 0.23 and 0.17, respectively. These results indicate that the $BPscore_{max}^{BMA}$ should generally be higher than the $MFscore_{max}^{BMA}$.

In order to test for a possible bias towards specific processes or functions, we additionally randomized all DDI networks. This was accomplished by keeping one of the two nodes of each interaction edge fixed while randomly shuffling the other nodes of the edges. Figures 5.3 and 5.4 depict the distributions of the $BPscore_{max}^{BMA}$ and the $MFscore_{max}^{BMA}$, respectively, for the randomized networks. All of the resulting distributions are very similar and closely resemble the background distribution for BP and MF. In contrast to the predicted sets, the distributions of the randomized experimental iPfam and 3did networks contain more DDIs with $BPscore_{max}^{BMA}$ below 0.1, but fewer with $BPscore_{max}^{BMA}$ between 0.1 and 0.2. Figures 5.5 and 5.6 depict the results of the analysis using the GO-slim sets. The analyzed GO categories, including "protein binding", are very similarly distributed between the different datasets suggesting no bias towards specific processes or functions for any of the tested domain interaction networks.

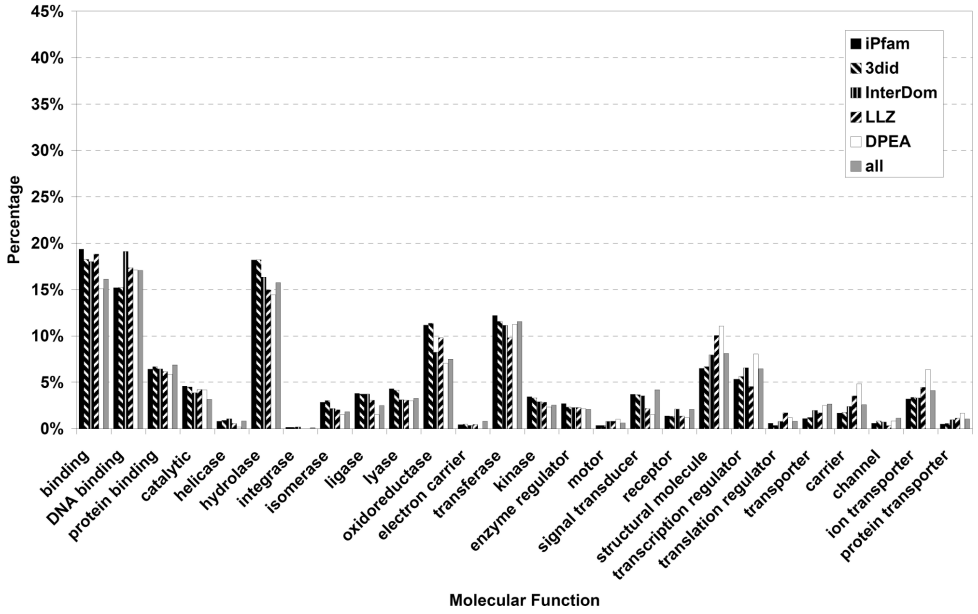

Figure 5.6: Distributions of GO-slim MF terms in the various DDI datasets. The single MF terms are shown on the x-axis, and the y-axis gives the percentage of domains in the different DDI sets annotated with the respective MFs.

5.3.3 Computing and Analyzing $GOscore_{max}^{BMA}$ Distributions

The $BPscore_{max}^{BMA}$ distributions for iPfam and 3did (Figure 5.7) show that most experimental DDIs have a very high similarity score exceeding 0.8, which means that the corresponding interacting domains are part of the same process or closely related processes. The high means of about 0.9 and medians of almost 1 further support this interpretation. For the predicted sets InterDom and DPEA, the distributions look alike. Interestingly, only one third of the predicted interactions in both sets have a $BPscore_{max}^{BMA}$ above 0.8. Furthermore, both datasets include a large fraction of interactions with $BPscore_{max}^{BMA}$ below 0.4, which indicates almost no functional similarity between the domains. The mean is 0.51 for both datasets and the median is 0.39 and 0.41 for InterDom and DPEA, respectively. DDIs predicted by LLZ contain substantially fewer interactions with high $BPscore_{max}^{BMA}$, and many more interactions with very low $BPscore_{max}^{BMA}$. This is also reflected by the relatively low mean of 0.35 and the median of 0.2.

More than 50 % of all interactions in iPfam and 3did are self-interactions. The predicted sets InterDom and LLZ contain no self-interactions between Pfam-A domains with GO annotation in contrast to DPEA. Therefore, we calculated the $BPscore_{max}^{BMA}$ distributions after removing self-interacting domains (Figure 5.8). The resultant distributions are very similar to the distributions obtained using all available domain interactions. The $BPscore_{max}^{BMA}$ mean and median values are 0.001 to 0.130 lower. Particularly, the medians for iPfam and 3did are only decreased by 0.003 and 0.001, respectively. In summary,

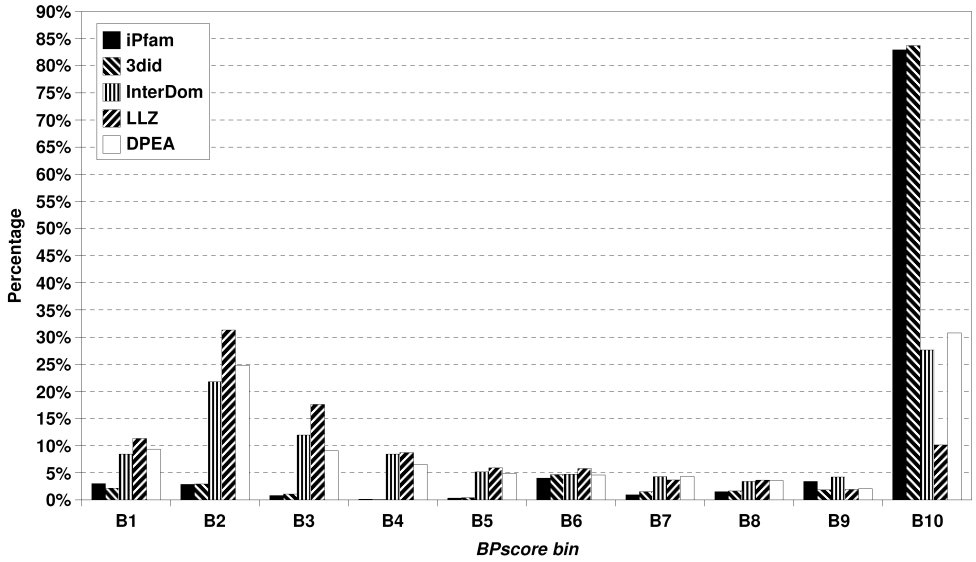

Figure 5.7: Distributions of the $BPscore_{max}^{BMA}$ values for the different datasets of experimental DDIs (iPfam and 3did) and predicted DDIs (InterDom, LLZ and DPEA). The *BPscore* bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

InterDom performs slightly better than DPEA, and both show better performance than LLZ.

Interestingly, the $MFscore_{max}^{BMA}$ distributions for iPfam and 3did are quite distinct from the distributions obtained for the predicted datasets (Figure 5.9). Almost 80 % of the domain interactions in iPfam or 3did have an $MFscore_{max}^{BMA}$ above 0.8 indicating that interacting domains are annotated with related molecular functions. In both datasets, only few domain interactions have a very low $MFscore_{max}^{BMA}$. The means of over 0.8 and the medians of almost 1 corroborate this interpretation. The predictions made by InterDom and DPEA show similar distributions, but rather low means and medians. As in case of the $BPscore_{max}^{BMA}$ distribution described above, predictions made by LLZ show a lower $MFscore_{max}^{BMA}$. Again, excluding self-interactions does not alter the obtained distributions markedly (Figure 5.10); the mean and median $MFscore_{max}^{BMA}$ are 0.02 to 0.18 lower. Overall, InterDom has better performance than DPEA, and both perform better than LLZ.

5.3.4 Deriving Confidence Score Thresholds

The three analyzed prediction methods InterDom, LLZ, and DPEA provide CSs for their predictions. In order to utilize sets of predicted interactions in practice, however, it is important to derive reasonable thresholds for low- and high-confidence sets of DDIs. It can

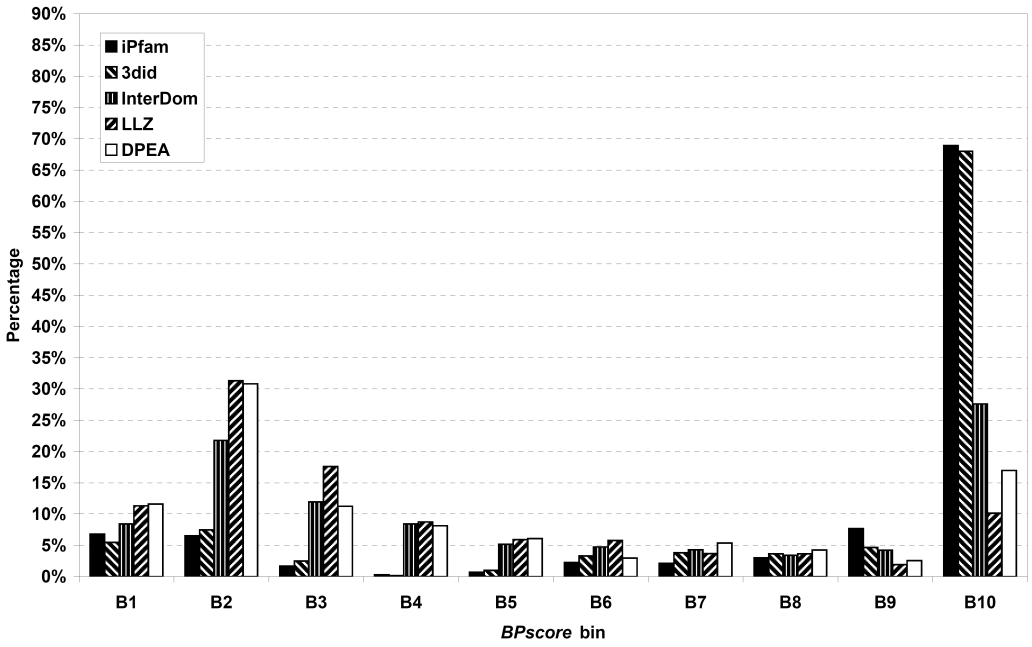

Figure 5.8: Distributions of the $BPscore_{max}^{BMA}$ values for the different datasets of experimental and predicted DDIs excluding self-interactions between domain. The *BPscore* bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

be expected that with rising CS also the functional similarity of the putatively interacting domains increases. To verify this expectation, we applied different CS thresholds and calculated the $GOscore_{max}^{BMA}$ means and medians for all interactions exceeding the respective CS threshold. Additionally, we calculated the overlap of these interactions with iPfam and 3did.

For the DPEA set, the change in $BPscore_{max}^{BMA}$ mean and median, and the change in dataset size with varying CS threshold are depicted in Figure 5.11A. In this case, when raising the CS threshold from 3 to 6, the $BPscore_{max}^{BMA}$ median increases from slightly over 0.4 to almost 1 and the mean rises from 0.51 to approximately 0.7. The $MFscore_{max}^{BMA}$ median and the overlap with iPfam and 3did also show a steep increase in this CS range (Figure 5.11B and C). Consequently, we suggest assigning predictions with a CS between 3 and 6 to a DPEA subset of low-confidence DDIs, and interactions with a CS above 6 to a high-confidence subset.

The analysis of the InterDom set reveals that the $BPscore_{max}^{BMA}$ median reaches 0.98 with a CS threshold of 30 (Figure 5.12A). The $BPscore_{max}^{BMA}$ mean is 0.68 at this point and continues to increase with higher thresholds. The same score development is observed for $MFscore_{max}^{BMA}$, but the score is slightly shifted towards higher thresholds (Figure 5.12B).

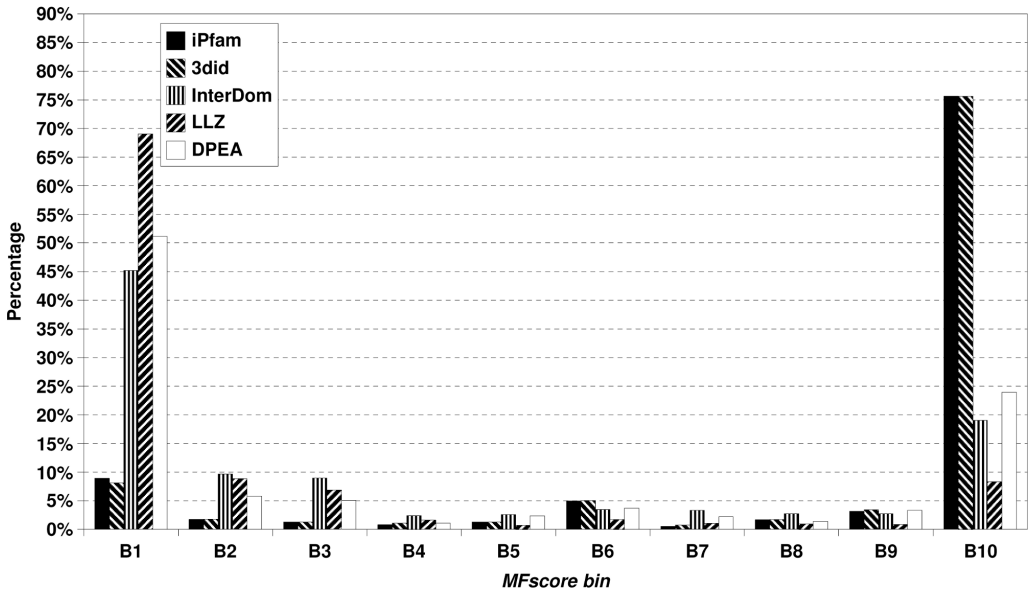

Figure 5.9: Distributions of the $MFscore_{max}^{BMA}$ values for the different datasets of experimental and predicted DDIs. The *MFscore* bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

At a threshold of 60, 1,888 interactions remain in the dataset and the median increase diminishes. With rising InterDom score, the overlap with iPfam and 3did increases and is about 27 % for a threshold of 60 (Figure 5.12C). Altogether, these results suggest a threshold of 60 for InterDom predictions with high confidence.

The analysis of LLZ predictions reveals that neither the $BPscore_{max}^{BMA}$ mean and median (Figure 5.13A) nor $MFscore_{max}^{BMA}$ mean and median (Figure 5.13B) reach significant values over the whole CS range. The same can be observed for the overlap with iPfam and 3did (Figure 5.13C). Consequently, these results do not allow for deriving any reasonable CS threshold for creating subsets of DDIs predicted by LLZ.

5.3.5 Comparing Human Protein Interaction Networks

We calculated the $BPscore_{max}^{BMA}$ for all PPI datasets described in Section 5.2.2. The results are summarized in Table 5.3, which is ranked by the mean $BPscore_{max}^{BMA}$. It ranges from 0.82 for Bioverse-core to 0.37 for the Wanker PPI set. In contrast to the experimental Y2H datasets, which have rather low mean $BPscore_{max}^{BMA}$, predicted datasets receive high mean scores, for instance, both HiMAP sets and Bioverse-core as well as the manually curated sets HPRD and HTT-literature. The diverging results for the HTT and ATX networks also indicate that literature-curated PPIs reach a higher $BPscore_{max}^{BMA}$ than PPIs derived by high-throughput experiments.

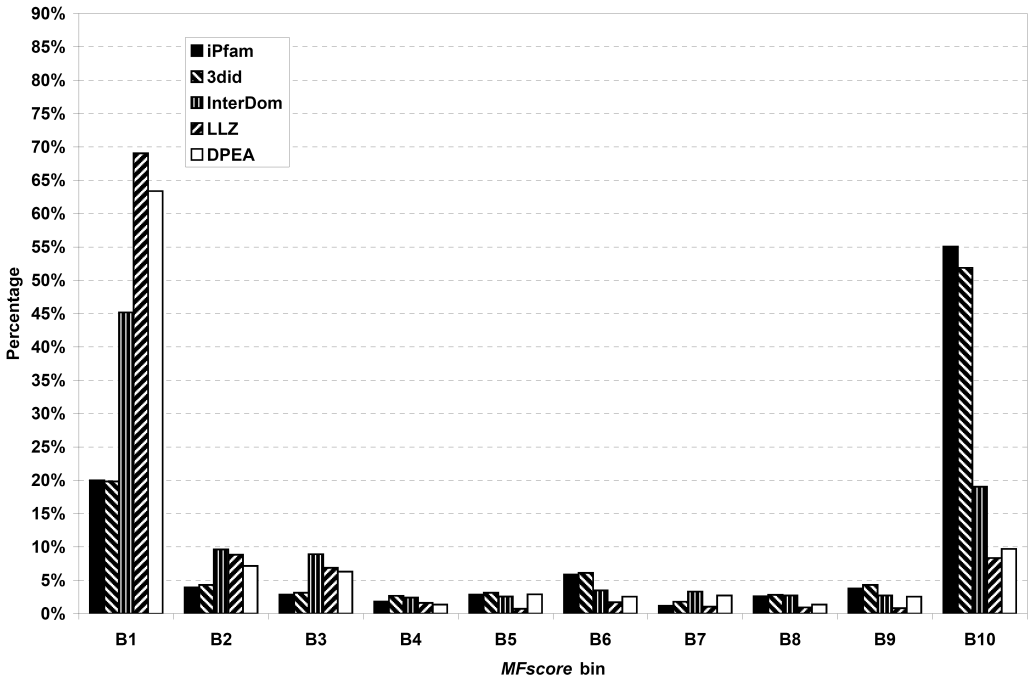

Figure 5.10: Distributions of the $MFscore_{max}^{BMA}$ values for the different datasets of experimental and predicted DDIs excluding self-interactions between domain. The *MFscore* bins correspond to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

Both the IUP- and IPG-sets have been derived from iPfam and contain the same PPIs but distinct GO annotation sources were utilized. Their $BPscore_{max}^{BMA}$ means are 0.76 and 0.81, respectively. These two values are lower than the mean of the corresponding DDIs in iPfam, which in part, may be due to the fact that we excluded self-interactions in the two PPI sets. Additionally, the annotation of the IUP-set is not restricted to interacting domains as in the case of the IPG-set but includes all available GO protein annotations. The score distributions for the IUP- and IPG-sets show that using the GO annotation of proteins or Pfam-A domains leads to different results (Figure 5.14). One possible explanation is that Pfam-A domain annotations do not fully describe all protein functions. If the same domain occurs in both interacting proteins and is responsible for the interaction, the calculated $BPscore_{max}^{BMA}$ will be higher for the IGP-set than for the IUP-set. In comparison, the manually curated HPRD set has a mean $BPscore_{max}^{BMA}$ of 0.66, and the distribution shows that more than 50 % of the interactions have a score above 0.7, but also 10 % of the interactions have a score between 0.1 and 0.2. While the first three consensus PPI sets (ConSet2-4) show a similar mean $BPscore_{max}^{BMA}$, the ConSet5 and ConSet6 score higher, but contain only few interactions.

Especially on the lower ranks, the $BPscore_{max}^{BMA}$ ranking of the datasets is similar to

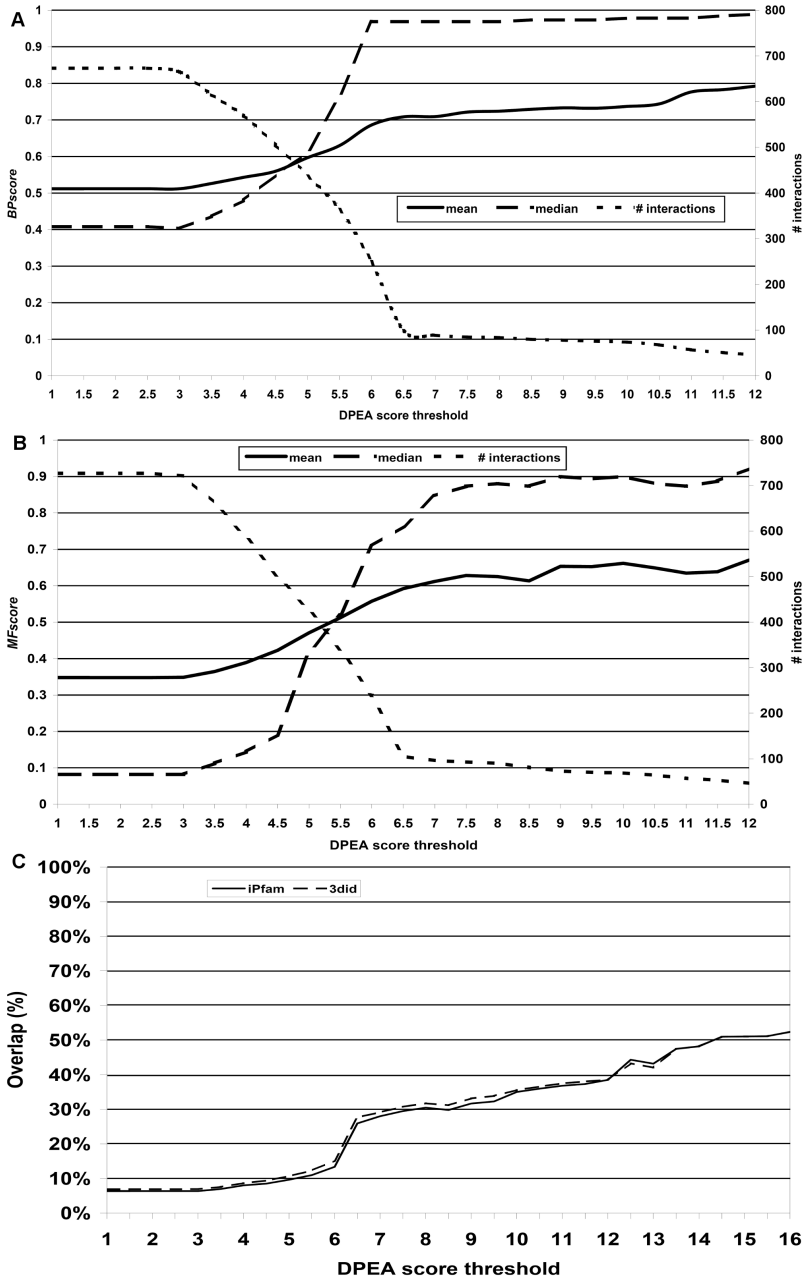

Figure 5.11: The plot depicts changes in the three validation measures for the DPEA dataset with varying confidence score threshold. (A) Change in $BPscore_{max}^{BMA}$ mean and median, and in dataset size. (B) Change in $MFscore_{max}^{BMA}$ mean and median, and in dataset size. (C) Change in overlap with iPfam and 3did. The size in panels (A) and (B) refers to the number of DDIs with confidence score exceeding the threshold. The changes in the validation measures suggest a threshold of 6 for high confidence predictions.

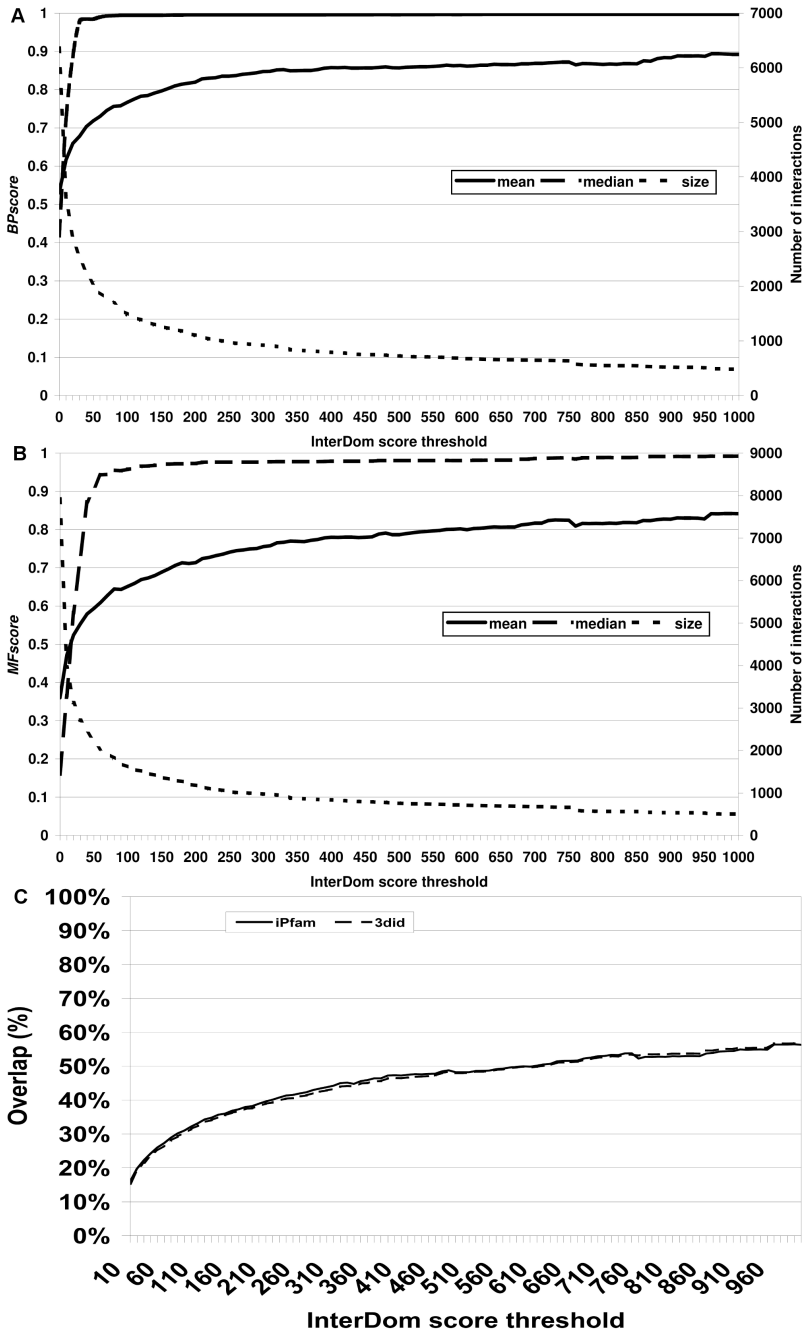

Figure 5.12: The plot depicts changes in the three validation measures for the InterDom dataset with varying confidence score threshold. (A) Change in $BPscore_{max}^{BMA}$ mean and median, and in dataset size. (B) Change in $MFscore_{max}^{BMA}$ mean and median, and in dataset size. (C) Change in overlap with iPfam and 3did. The size in panels (A) and (B) refers to the number of DDIs with confidence score exceeding the threshold. The changes in the validation measures suggest a threshold of 60 for high confidence predictions.

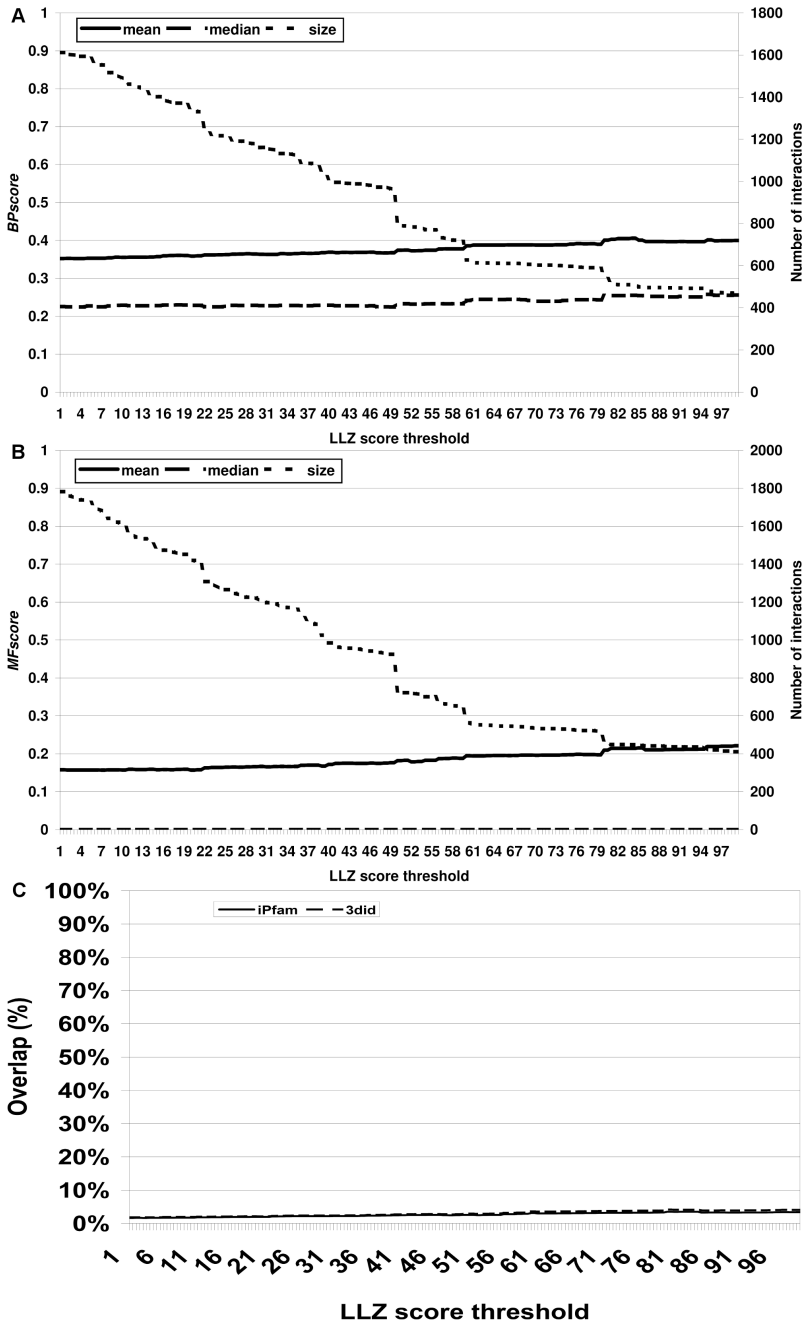

Figure 5.13: The plot depicts changes in the three validation measures for the LLZ dataset with varying confidence score threshold. (A) Change in $BPscore_{max}^{BMA}$ mean and median, and in dataset size. (B) Change in $MFscore_{max}^{BMA}$ mean and median, and in dataset size. (C) Change in overlap with iPfam and 3did. The size in panels (A) and (B) refers to the number of DDIs with confidence score exceeding the threshold. It can be seen the three validation measures remain almost constant over the whole confidence score range.

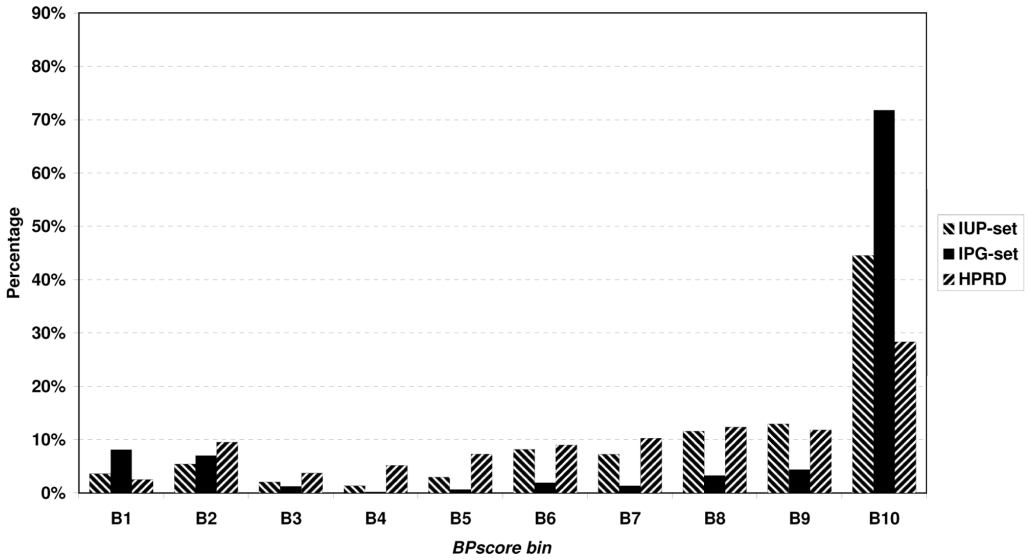

Figure 5.14: Distributions of the $BPscore_{max}^{BMA}$ values for the IUP, IGP, and HPRD PPI datasets. The *BPscore* bin corresponds to the following intervals: B1: [0.0, 0.1[; B2: [0.1, 0.2[; B3: [0.2, 0.3[; B4: [0.3, 0.4[; B5: [0.4, 0.5[; B6: [0.5, 0.6[; B7: [0.6, 0.7[; B8: [0.7, 0.8[; B9: [0.8, 0.9[; B10: [0.9, 1.0].

rankings computed from the HPRD or Y2H verification rate (Table 5.3). The predicted Bioverse-core set and the consensus sets have the best verification rates with respect to HPRD. The published validation rates of 78 % and 62-66 % for the Vidal and Wanker sets, respectively, agree well with the slightly higher mean $BPscore_{max}^{BMA}$ 0.47 of Vidal in contrast to the mean 0.36 of Wanker.

5.4 Conclusions

The basic assumption underlying our analysis of interaction networks is that interacting domains or proteins should have highly similar BP annotations and, to a smaller degree, similar MF annotations. In light of this hypothesis, we evaluated the functional similarity of three predicted and two experimental domain-domain interaction (DDI) networks as well as several predicted and experimental human protein-protein interaction (PPI) networks. Furthermore, we investigated to which extent predicted DDIs or PPIs overlap with experimentally derived interactions.

We demonstrated that the application of functional similarity measures is not restricted to the validation of PPIs (Guo *et al.*, 2006), but also useful for DDIs. Our analysis of DDIs revealed that the BP similarity of interacting domains is generally higher than their MF similarity. This observed difference between BP and MF similarity agrees well with findings by Guo and colleagues for PPIs using other GO similarity measures (Guo *et al.*,

Table 5.3: Ranking of predicted and experimental protein interaction networks based on the $BPscore^{BMA}_{max}$. The column "Scored" contains the fraction of PPIs with an assigned $BPscore^{BMA}_{max}$. The two rightmost columns give the percentages of PPIs contained in HPRD or the combined Y2H set Vidal & Wanker.

Dataset	Interactions	Scored (%)	mean BPscore	HPRD (%)	Vidal & Wanker (%)
Bioverse-core	3,266	83.2	0.823	28.9	1.1
IPG-set	1,931	45.9	0.815	15.9	0.7
HiMAP-core	8,832	84.6	0.813	9.1	0.6
HiMAP	38,378	89.4	0.799	3.8	0.2
IUP-set	1,931	22.8	0.764	15.9	0.7
ConSet6	484	77.5	0.709	21.3	1.2
HPRD	20,121	86.1	0.662	100.0	0.6
HTT-literature	428	97.4	0.643	90.2	0.2
ConSet5	1,565	73.2	0.642	16.1	1.3
Bioverse	233,941	81.4	0.572	1.5	0.1
ConSet3	10,844	66.5	0.561	9.2	0.8
ConSet4	4,747	67.1	0.559	10.2	0.9
ConSet2	38,258	69.3	0.556	6.0	0.4
Sanger-core	11,131	65.3	0.551	4.5	0.6
ATX-literature	4,796	67.5	0.537	46.9	39.1
HomoMINT	10,870	57.5	0.510	5.6	0.7
OPHID	28,255	62.6	0.499	4.4	0.2
Vidal	2,754	40.2	0.471	3.5	100.0
HTT-Y2H	164	62.2	0.456	3.8	5.1
POINT	98,528	56.9	0.451	2.6	0.2
Sanger	67,518	62.3	0.427	1.3	0.1
ATX-interologs	1,527	62.0	0.418	6.8	1.2
ATX-Y2H	770	39.9	0.394	1.4	1.0
Wanker	2,033	54.8	0.370	1.2	100.0

2006). The difference may be partly explained by the fact that interacting domains or proteins participate in similar processes but perform different functions. Another reason may be that GO terms are more densely connected in the top levels of the BP ontology than in the MF ontology.

The iPfam-derived IUP- and IPG-sets consist of the same PPIs, but the IUP-set is annotated with the GO terms of the proteins in UniProtKB and the IPG-set with the GO terms of the interacting Pfam-A domains. The direct comparison of these two sets re-

vealed that the utilized annotation source has an effect on the $BPscore_{max}^{BMA}$ results. This provides evidence that the choice of the annotation source contributes to the differing findings for DDIs and PPIs. Moreover, a larger number of proteins annotated with diverse BPs may decrease the mean $BPscore_{max}^{BMA}$ of protein networks in contrast to sets of DDIs annotated with more generic GO terms.

In agreement with our results on human PPI networks, Reguly *et al*. observed for yeast interaction datasets that the GO annotation of literature-curated PPI sets is more coherent than the GO annotation of high-throughput PPI sets (Reguly *et al.*, 2006). Since datasets of manually curated PPIs taken from scientific literature have a higher mean $BPscore_{max}^{BMA}$ than most predicted and high-throughput datasets, the latter sets may contain a significant number of false interactions or a large amount of proteins involved in novel processes. This can lead to a considerable decrease in observed $BPscore_{max}^{BMA}$. Furthermore, proteins described in the literature may be annotated particularly well using GO, which is in contrast to most other proteins that were automatically annotated by less reliable methods (Camon *et al.*, 2003). Therefore, a more thorough analysis of the PPI results using alternative measures will be required to explain differences between predicted and experimental datasets.

Our functional similarity analysis in conjunction with an evaluation of the overlap between experimentally derived and predicted DDIs facilitated the definition of confidence score thresholds for DDI predictions. These thresholds are useful for improving the utilization of DDIs for predicting PPIs as well as for assessing the confidence of PPIs derived by high-throughput experiments. In the future, incorporating other similarity criteria besides GO may further improve the confidence assessment of predicted interactions. As the coverage and quality of GO annotations improves, the importance of approaches that use functional similarity for the validation and prediction of PPIs and DDIs will increase.

Chapter 6

Improving Disease Gene Prioritization using the Semantic Similarity of Gene Ontology Terms

Many hereditary human diseases are caused by a combination of effects resulting from sequence alterations in multiple genes. Experimental methods commonly applied for identifying disease-related genes often yield lists of several hundred candidate genes. Therefore, computational methods have been devised for prioritizing the candidates for further validation. In this chapter, we provide evidence that a method based on functional similarity measures achieves a comparable or even better performance than previously developed more complex methods.

After reviewing previously developed methods for prioritizing candidate disease genes, our new MedSim method is presented. MedSim automatically annotates diseases with GO terms and ranks candidate genes using our functional similarity measures. Using leave-one-out cross validation analysis with three different benchmark sets, we demonstrate the performance achieved by our method. Examples for exemplary diseases are used to illustrate the benchmarking results. Finally, the performance achieved by MedSim is compared to five previously published methods. Part of this work was presented as poster at ISMB/ECCB 2009 and received one of three Outstanding Poster Awards for over 700 posters.

6.1 Introduction

More than 1,800 hereditary disorders in human are known to be caused by mutations in a single gene (O'Connor and Crystal, 2006), but these monogenic diseases are mostly very rare. In contrast, many diseases of major importance to public health, like cancer, diabetes, and cardiovascular disorders, are influenced by simultaneous alterations in several

genes (Gibson, 2009). In order to identify genes involved in such multigenic diseases, genomic linkage and association studies are performed (Altshuler *et al.*, 2008; Cordell and Clayton, 2005; Teare and Barrett, 2005; Marchini *et al.*, 2005; Plomin *et al.*, 2009). The genomic regions identified in these studies may comprise up to several hundred candidate disease genes. While most of these candidates are not related to the disease of interest, experimentally testing the complete list of candidate genes is not feasible because of the time and cost involved with such an extensive procedure. Therefore, computational techniques are increasingly used for ranking putative disease genes according to their potential of being involved in the disease under investigation (Ideker and Sharan, 2008; Kann, 2007; Oti and Brunner, 2007; van Driel and Brunner, 2006; Yu *et al.*, 2008; van Driel *et al.*, 2006).

Several studies have examined the specific properties of genes and their products known to be associated with human genetic disorders and have explored networks linking diseases based on the involved genes (van Driel *et al.*, 2006; Feldman *et al.*, 2008; Goh *et al.*, 2007; Jimenez-Sanchez *et al.*, 2001; Lee *et al.*, 2008). Feldman *et al*. investigated genes contributing to inherited diseases in the context of protein interaction networks (Feldman *et al.*, 2008). They provided evidence for the fact that disease genes are less likely to be essential for cell survival than other genes. Furthermore, proteins encoded by genes related to polygenic diseases interact with a significantly larger variety of proteins than other disease proteins. Jimenez-Sanchez *et al*. found a substantial correlation between functional categories of disease proteins, like receptors or enzymes, and observed disease features, for instance, age of onset or mode of inheritance (Jimenez-Sanchez *et al.*, 2001). Van Driel *et al*. performed a text-mining analysis of phenotypes (van Driel *et al.*, 2006) taken from the Online Mendelian Inheritance in Man database (OMIM, Amberger *et al.*, 2009). They defined a similarity measure for phenotypes, the MimMiner score, that is based on mapping OMIM disease descriptions to the "anatomy" ("A") and "diseases" ("C") sections of the Medical Subjects Headings (MeSH, Lowe and Barnett, 1994) vocabulary. This similarity score correlates positively with the relatedness of functional gene annotations. Goh *et al*. studied the human disease network and the disease gene network (Goh *et al.*, 2007). In the human disease network, nodes represent diseases and are connected by edges if they share at least one disease gene. In contrast, in the disease gene network, nodes are disease genes and are linked if they are associated with the same disease. Goh *et al*. manually grouped the investigated diseases into 22 classes, such as cancer, developmental or skeletal, and revealed that many genes contribute to disorders belonging to the same disease class. Additionally, if genes are involved in the same disease, their functions are significantly more similar than those of randomly selected genes. Lee *et al*. observed similar results for metabolic diseases (Lee *et al.*, 2008). In the metabolic disease network (MDN) created in this study, diseases are linked by an edge if they are related through similar reactions. The authors concluded that diseases exhibited elevated comorbidity if they were connected in the MDN. Furthermore, disorders that were highly connected in the MDN displayed larger prevalence in the population than others.

The discovered relationships between properties of genes and gene products as well as their involvement in genetic disorders have been exploited by a number of bioinformatics approaches for prioritizing disease gene candidates. Many of these methods are discussed in recent review articles (Ideker and Sharan, 2008; Kann, 2007; Oti and Brunner, 2007; van Driel and Brunner, 2006; Yu *et al.*, 2008; van Driel *et al.*, 2006). One outcome of studies of known disease genes and proteins is that measures of phenotype similarity are helpful in reproducing biological relationships and are suited for finding new disease gene associations. However, in order to improve current disease gene prioritization methods, it is necessary to further develop structured vocabularies for describing phenotypes as well as for annotating genes and their products (Oti and Brunner, 2007; van Driel and Brunner, 2006; Yu *et al.*, 2008; van Driel *et al.*, 2006).

Most computational approaches rely on the integration of several sources of heterogeneous data such as sequence features, gene expression data, and protein-protein interactions (PPIs). The PROSPECTR method introduced by Adie *et al*. is a sequence-based approach that applies decision trees trained on features such as gene and protein length, and the number of exons (Adie *et al.*, 2005). Later, Aerts and colleagues devised Endeavour, which is based on the integration of biological evidence from many different types of data, including PPIs, pathways, gene expression, and sequence similarity (Aerts *et al.*, 2006). The characteristics of known disease genes were extracted from each data source separately to rank candidate genes; the resultant ranking lists were then combined to a final overall ranking. Recently, several new methods have been published that solely build on interaction networks and GO annotation (Chen *et al.*, 2009; Ortutay and Vihinen, 2009; Ozgür *et al.*, 2008; Shriner *et al.*, 2008). Chen *et al*. applied different algorithms from the analysis of social and web networks to disease gene prioritization (Chen *et al.*, 2009). They concluded that although network data provide valuable information, methods exploiting functional annotation are generally better than network-based methods. Ortutay and Vihinen integrated GO annotation and protein interactions for identifying genes involved in immunodeficiencies (Ortutay and Vihinen, 2009). Three different network topology parameters were computed pertaining to an interaction network of genes known to be related to the immune system. For each of these parameters, a set of genes was selected from this gene network and then subjected to GO enrichment analysis. If a gene was annotated with enriched terms and achieved some significant network parameter value, it was given higher priority.

Cross-references to structured vocabularies are leveraged by another class of methods for disease gene prioritization that uses phenotype similarity measures (Chen *et al.*, 2007; Freudenberg and Propping, 2002; Lage *et al.*, 2007; Perez-Iratxeta *et al.*, 2002; Robinson *et al.*, 2008; Tiffin *et al.*, 2005; Wu *et al.*, 2008; Yilmaz *et al.*, 2009). The ACGR method by Yilmaz *et al*. is based on manual annotation of diseases with GO terms (Yilmaz *et al.*, 2009). Candidate genes are selected based on the number of annotated GO terms shared with the disease in question, and subsequently, the annotation similarity between each candidate and the input disease is calculated. Validation of this approach was done using three rare syndromes (AICARDI syndrome, CHARGE syndrome, and focal dermal

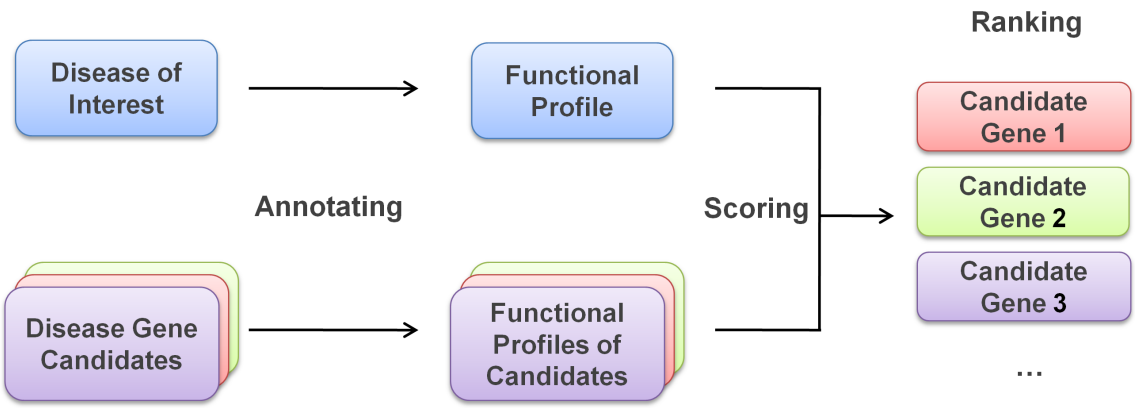

Figure 6.1: Flow chart of the MedSim method. First, the functional profiles of the disease of interest and the disease gene candidates are created using one of the annotation strategies. Afterwards, the functional profile of the disease is scored against each functional profile of a candidate, and the candidates are ranked according to this functional similarity score.

hypoplasia) as case studies. A major hurdle for the large-scale application of the ACGR method is the required manual annotation of the disease with GO terms.

In the following, we present MedSim, a novel approach to disease gene prioritization that exploits the similarity between functional annotations of diseases and of candidate genes (Figure 6.1). In particular, MedSim automatically derives a functional profile for a disease phenotype based on the genes and proteins that are already known to be related to this phenotype. Potential candidate disease gene products are compared to this profile using our functional similarity measures (Chapter 2.3). Using different sets of proteins encoded by known disease genes, we demonstrate that our novel method allows for assigning known disease genes specifically to the correct phenotype. Above all, we show that MedSim is able to significantly outperform previous, more complex methods that rely on more diverse and voluminous, and therefore, harder accessible data. We further explore the effect of different semantic similarity measures on prediction performance. MedSim also enables the distinction of phenotypes with identical disease-associated genes from unrelated phenotypes. Finally, we implemented the best method in our FunSimMat (Chapter 4) web server (http://www.funsimmat.de), making it easily accessible to biological and medical users.

6.2 Materials and Methods

6.2.1 Data Sources

OMIM is a database of human genes and genetic disorders. Entries in OMIM describe either a single gene involved in some genetic disease or a phenotype with known or putative, but unknown, genetic basis. For this study, we focused on phenotype entries (accession numbers starting with "#" or "%") from OMIM (downloaded on 10 October 2007). The mapping of proteins encoded by human disease genes to the OMIM phenotypes was obtained from UniProtKB (release 12.3, The UniProt Consortium, 2009). The annotations of proteins with GO terms from all three ontologies were also extracted from this UniProtKB release. The majority of human GO annotations (approximately 62 % in our dataset) were derived by purely automatic methods (evidence code IEA).

A set of human PPIs was compiled from the Human Protein Reference Database (HPRD, version 7, Prasad *et al.*, 2009), IntAct (downloaded on 16 May 2008, Kerrien *et al.*, 2007), the Molecular Interactions Database (MINT, downloaded on 7 April 2008, Chatr-Aryamontri *et al.*, 2007), the Database of Interacting Proteins (DIP, downloaded on 14 February 2008, Salwinski *et al.*, 2004), protein complexes extracted from SIFTS (downloaded on 4 March 2008, Velankar *et al.*, 2005), and the CORUM database (downloaded on 19 May 2008, Ruepp *et al.*, 2008). All protein and gene identifiers used by these sources were mapped to UniProtKB accession numbers. Members of the same protein complex possibly affect the same diseases. Therefore, we chose the matrix model (Bader and Hogue, 2002) for decomposing protein complexes into pairwise PPIs. Using this model, all proteins in a complex are connected with each other resulting in a full interaction graph for this complex. A set of random PPIs was created by keeping one partner of each interaction fixed and randomly shuffling the interacting partners.

Mouse orthologs of human proteins were obtained from InParanoid (version 6.1, Berglund *et al.*, 2007). Each Inparanoid cluster is seeded with one protein from the two species, which are reciprocally best matches. These two proteins are called main orthologs and receive an inparalog score of 1.0. Proteins from both species are added to the cluster if their sequence similarity to the main ortholog of the same species is higher than the sequence similarity between the two main orthologs. These added proteins are called in-paralogs and have an inparalog score that is smaller than 1.0. From all Inparanoid clusters, we extracted the main orthologs. MGI (Blake *et al.*, 2009) and Ensembl (Hubbard *et al.*, 2009) accessions used by InParanoid were converted to UniProtKB accessions using data from Ensembl BioMart (downloaded on 14 May 2008). Additionally, the chromosomal location of human genes and the cross-references to UniProtKB proteins were obtained via BioMart on 21 October 2008.

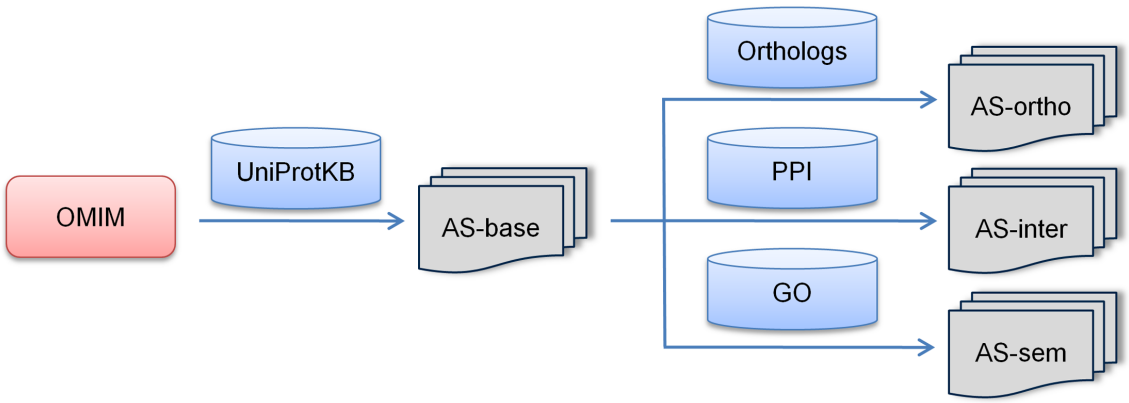

Figure 6.2: Flow chart of the automatic annotation strategies. In AS-base, OMIM phenotypes are annotated with GO terms from known disease-related proteins from UniProtKB. The other three annotation strategies (AS-ortho, AS-inter, and AS-sem) add GO annotations that were derived from mouse orthologs, interaction partners, or semantically similar terms, respectively.

6.2.2 Functional Profiles for Phenotypes

Descriptions of human diseases are usually written in natural language, and genes or proteins are annotated with the respective diseases in which they are involved. However, diseases are not directly annotated with ontologies like GO preventing the application of functional similarity measures. Therefore, we developed several new strategies for automatically annotating OMIM entries with GO terms (Table 6.1 and Figure 6.2). We define the functional profile as set of all GO terms annotated to a gene product or a disease. The first annotation strategy (AS-base) transfers all GO terms annotated to proteins annotated to a disease in UniProtKB to the corresponding OMIM entry. Since genes and proteins are often annotated with terms from different levels of the GO hierarchy, functional profiles may also contain ancestral terms. These ancestral terms are redundant according to the true path rule because annotation with a term already implies annotation with all its predecessors (Section 2.1.3). Therefore, a term is removed from a functional profile if one of its descendants from the GO hierarchy is also contained in this functional profile.

If the known disease genes and proteins lack any GO annotation, OMIM entries cannot be annotated by applying AS-base. Furthermore, not all functions and processes involved in the respective disease may be covered by the genes and proteins annotated with this disease. Therefore, we explored several possibilities for automatically extending the available annotation. The second annotation strategy (AS-ortho) adds GO terms from mouse orthologs of human disease proteins to the functional profile, and the third annotation strategy (AS-inter) augments the profile with GO terms from direct interaction partners of disease proteins. In both strategies, redundant terms are removed after adding the GO terms to the profile. A fourth strategy for expanding the functional profiles

Table 6.1: Summary of annotation strategies for creating functional profiles for diseases and candidates. The table lists sources of GO annotation used by the different annotation strategies. Term filtering can be applied to functional profiles created by any of these strategies. In all cases, terms are removed from a functional profile if one of their descendants from the GO hierarchy is also contained in this functional profile.

Annotation strategy	GO annotation sources
AS-base	known disease genes/proteins
AS-ortho	known disease genes/proteins orthologs of known disease genes/proteins
AS-inter	known disease genes/proteins interaction partners of known disease genes/proteins
AS-sem	known disease genes/proteins semantically similar terms

(AS-sem) is based solely on GO. The sim_{Rel} measure (Equation 2.8) is used to identify terms that are highly related to at least one other term in the profile. Two different sim_{Rel} cut-offs, 0.90 and 0.95, are applied for selecting and adding related terms to a profile.

Annotating diseases with one of the described automatic strategies might lead to a diffuse functional profile containing diverging processes, functions, or components. If a protein has many interaction partners with diverse functions or the dataset contains false positive interactions, such a profile might result from applying an automatic annotation strategy. Therefore, a term filtering step is introduced that removes unrelated terms from functional profiles. In this step, terms are retained only if they have a sim_{Rel} score exceeding a predefined threshold with at least one other term in the profile. Given a functional profile consisting of four GO terms, if two of these terms are similar and the other two terms are not related to any term in the profile, the latter two are deleted. In contrast, if the latter two terms are similar to each other, all four terms are retained in the profile. We tested the two sim_{Rel} thresholds 0.60 and 0.80. A threshold of 0.60 removes only terms that have a very low similarity to all other terms in the profile. The threshold of 0.80 is more restrictive and removes all terms with a low or medium similarity to all other terms. The term filtering step is only applied to profiles consisting of at least three GO terms. If no term pair in the profile has a similarity above the sim_{Rel} threshold, the respective disorder is ignored.

6.2.3 Benchmark Set 1

Several prioritization methods assess the general probability of proteins to be associated with a disease, but are unspecific with respect to the disease. In order to test the ability

of MedSim to assign known disease proteins to exactly the phenotype they are known to be involved in, we conducted leave-one-out cross validation. For this benchmark, only OMIM phenotypes with at least three known disease proteins were selected, resulting in a preliminary set of 99 phenotypes. For each of these phenotypes, one disease protein was randomly selected and removed. Subsequently, the functional profiles of the 99 phenotypes were derived using strategies AS-base, AS-ortho, or AS-inter based on the remaining known disease proteins. Phenotypes were discarded if either the phenotype or the randomly selected protein was not annotated with terms from all three GO ontologies. Benchmark set 1 consists of 78 phenotypes with 78 randomly selected known disease proteins. Five of these selected proteins contribute to two diseases in the test set and were coincidentally chosen for both phenotypes.

6.2.4 Benchmark Set 2

Genomic linkage and association studies are used to associate segments of a chromosome with a particular quantitative trait. These quantitative trait loci (QTL) can consist of up to several hundred candidate genes. Our second benchmark set simulates genomic experiments, which result in QTLs that provide lists of candidate disease genes. For each disease protein associated with one of the 99 phenotypes with at least three known disease proteins (Section 6.2.3), leave-one-out cross validation was performed for classifying the candidates according to their disease relatedness. After removing a protein p from the list of known proteins for one disease, the remaining associated proteins for this disease were used for deriving the functional profile. An artificial quantitative trait locus (aQTL) of size 10 Mbp was centered at the genomic start position of the gene encoding p. All proteins translated from any gene in this aQTL were added to the list of putative disease proteins. Benchmark set 2 contains 519 different aQTLs for 99 phenotypes. A complete list of the 99 phenotypes and the 519 known associated proteins is given in Appendix A. All four annotation strategies were applied to annotate this benchmark set. Additionally, term filtering with both thresholds 0.60 and 0.80 was used in conjunction with AS-base and AS-inter, and term filtering using threshold 0.80 with AS-sem. As control, random PPIs were used for AS-inter.

6.2.5 Benchmark Set 3

Some previously published approaches were benchmarked using random artificial quantitative trait loci (rQTL), which are sets of random genes supplemented with one known disease gene. To facilitate a performance comparison between MedSim and these methods (Section 6.3.8), we created a third benchmark set. This set differs from benchmark set 2 in the method for creating the rQTLs. Here, leave-one-out cross validation was performed for each disease protein associated with one of the 99 phenotypes from benchmark set 2 that was annotated with terms from all three ontologies. In an rQTL, one left-out

protein was complemented with 99 proteins randomly drawn from the set of all human proteins annotated with terms from all three ontologies. Benchmark set 3 consists of 287 distinct rQTLs for the 99 different phenotypes. This set was annotated using AS-base without and with term filtering (threshold 0.80) as well as AS-sem (cut-off 0.95) with term filtering (threshold 0.80).

6.2.6 Functional Similarity Measures

MedSim calculates the similarity between functional profiles of diseases with the profiles of candidate genes and ranks the candidates according to these scores. We used FSST version 1.3.1 (Section 4.3) for calculating the functional similarity scores. The computed similarity scores are based on semantic similarity between GO terms and apply the best-match average approach (Section 2.3.1). The sim_{Rel} score (Equation 2.8), which assesses differences and commonalities between GO terms, was used to determine the semantic similarity of GO terms. The level of detail of the annotated terms is a further factor influencing this score. In order to find out whether the performance of MedSim depends on the choice of the semantic similarity measure, Lin's measure (Equation 2.7) was used as well. This similarity score measures commonalities and differences between two GO terms, but in contrast to sim_{Rel}, is not affected by the degree of specificity of some term. To compare two functional profiles, several similarity scores are evaluated: *BPscore* (Equation 2.17) for biological process, *CCscore* (Equation 2.17) for cellular component, *MFscore* (Equation 2.17) for molecular function, *rfunSim* (Equation 4.2) combining *BPscore* and *MFscore*, and *rfunSimAll* (Equation 4.4) combining *BPscore*, *CCscore* and *MFscore*.

6.2.7 MedSim Implementation

We implemented the MedSim approach in our FunSimMat web service (http://www.funsimmat.de, Section 4.4). The functional profiles for all OMIM entries and human proteins in UniProtKB were derived using strategy AS-base without and with term filtering (threshold 0.80); all functional scores between OMIM diseases and human proteins were precomputed. The FunSimMat web page offers a simple HTML form for prioritizing a list of candidates by entering the OMIM accession of a specific disease and the UniProtKB accessions of the corresponding candidate disease proteins. The results table contains the candidates ranked by the functional similarity score. An alternative for providing a candidate list is the possibility of scoring all human proteins against the disease of interest. Additionally, programmatic access to the data is possible through the XML-RPC and RESTlike interfaces.

6.3 Results and Discussion

6.3.1 Functional Similarity of Diseases

The functional profiles of diseases provide a possibility for calculating the similarity between different phenotypes. To test this hypothesis, we created a disease network from the phenotype entries in OMIM. In this network, two diseases are linked by an edge if they share at least one disease-associated protein. In total, this network contains 1,205 nodes and 1,821 edges. Additionally, we created a randomized disease network by keeping one node of each edge fixed while randomly shuffling the other nodes of the edges, and we used a background network containing all possible edges between diseases to obtain a background distribution of similarity values. The random network contains 23 edges between phenotypes that actually share disease gene products.

The distributions of BP similarities and MimMiner scores for the disease and random networks are illustrated in Figure 6.3. The distributions for the background network resemble the ones for the random network (not shown). From Figure 6.3, it can be seen that most disease pairs sharing disease proteins are assigned very high scores by MedSim. In contrast, the similarity scores of random disease pairs vary over the whole score range, and only few pairs receive a high score. MimMiner score distributions of random pairs and protein-sharing disease pairs have a higher overlap. Most random phenotype pairs receive a low similarity, but many pairs sharing proteins also have a low similarity.

The scores of the 23 disease pairs that appear in both the random and the true disease networks confirm this result. Twelve out of these 23 pairs have a *BPscore* above 0.9, another 5 have a score above 0.5, and the remaining 6 pairs lack BP annotation completely. In contrast, all these pairs have MimMiner scores between 0.03 and 0.53 rendering them indistinguishable from true random pairs. The distributions for the other functional similarity scores, *CCscore* and *MFscore*, are similar to the distribution of the *BPscore*. In conclusion, MedSim performs better than the text-mining method MimMiner for determining diseases with a common functional basis and reliably detects pairs of diseases with similar functional profiles.

6.3.2 Performance of Different Annotation Strategies

The purpose of benchmark set 1 was to assess whether MedSim selectively associates phenotypes with known associated disease proteins. This set contains 78 pairs of phenotype and disease protein, which were annotated using strategies AS-base, AS-ortho, and AS-inter. The all-against-all comparison of phenotypes and disease proteins was performed using FSST. For each phenotype, the proteins were ranked in descending order of their functional similarity to this disease. For each disease phenotype, the protein that was randomly removed from this phenotype was treated as positive and all other proteins were treated as negatives. The goal of the prioritization is to rank the positive protein on top of

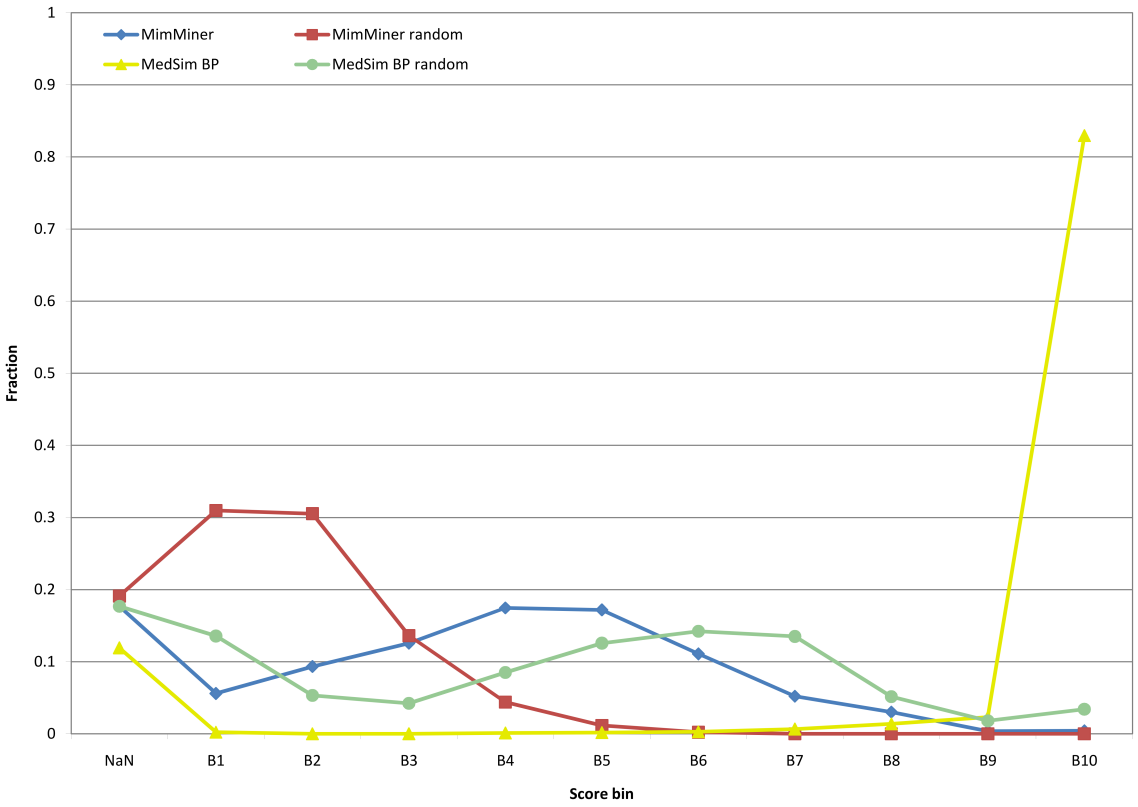

Figure 6.3: Distributions of the *BPscore* and the MimMiner score for disease pairs associated with identical proteins as well as for random disease pairs. The x-axis is divided into ten bins of size 0.1; scores in B1 fall into [0; 0.1[, scores in B2 fall into [0.1; 0.2[, and so on. The bin labeled "NaN" contains disease pairs for which no MimMiner score was available, or no *BPscore* score could be computed because of missing GO annotation. The y-axis shows the fraction of disease pairs belonging to the respective bins.

the list. We applied receiver operating characteristic (ROC) analysis and determined the area under the ROC curve (AUC) for testing the ability of MedSim to detect the correct protein for each disease. Additionally, we calculated the sensitivity and specificity of the predictions. Sensitivity is the percentage of correctly identified disease proteins ranked above a preset rank cut-off. Specificity is the percentage of proteins not involved in the disease ranked below this cut-off. When stating sensitivity values, we will always refer to a specificity threshold of 90 %. The performance values presented in this chapter are conservative estimates due to the following two reasons. First, the ranked list of proteins may contain several proteins associated with a disorder, but solely the randomly left-out protein is considered as true positive. Second, proteins labeled as negative might, in fact, be as yet unknown true positives.

Figure 6.4 gives an overview of AUC values achieved by MedSim on benchmark set

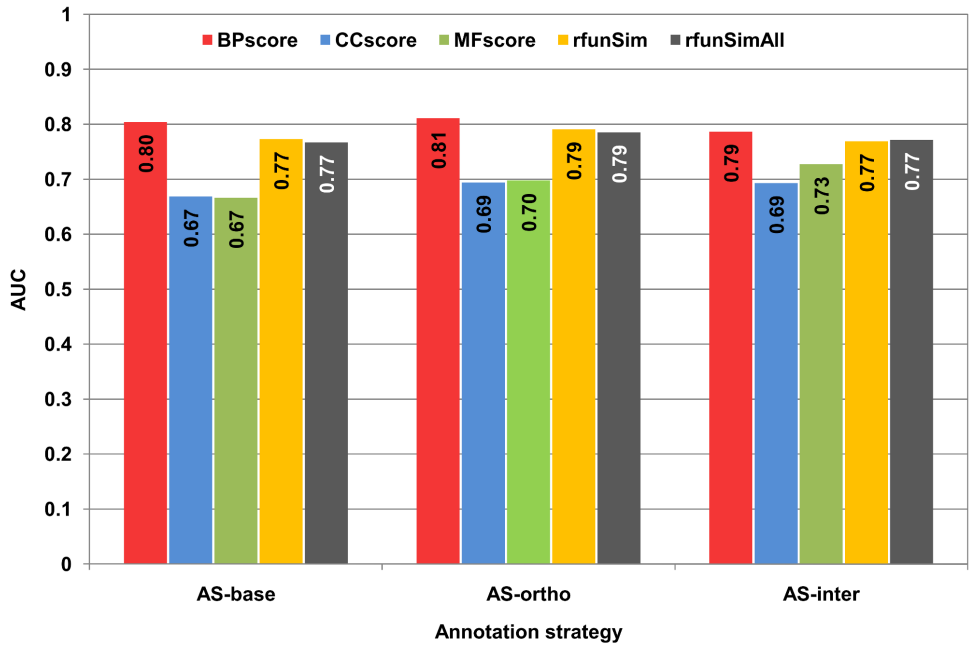

Figure 6.4: AUC values for prioritization of benchmark set 1 with different annotation strategies. The bars depict the AUC values achieved by MedSim utilizing the different functional similarity scores when the functional profiles were derived using AS-base, AS-ortho, or AS-inter.

1 using the different annotation strategies and functional similarity scores. Figure 6.5 depicts the ROC plot for the different scores obtained with strategy AS-base on this benchmark set. In this case, the *BPscore* achieves the best performance with an AUC of about 0.80 and a sensitivity of 56 %, followed by the scores *rfunSim* and *rfunSimAll*. When adding annotations derived from orthologs (AS-ortho), the prediction performance in terms of AUC remains almost constant. However, the sensitivity of the *BPscore* increases to 59 %. Applying AS-inter has an inconclusive effect on the AUC and sensitivity values. The AUC of *BPscore* is slightly reduced, while the *MFscore* performance increases to 0.73. The sensitivity of the *BPscore* predictions falls to 53 %, but sensitivity using *MFscore* rises to 38 %. Conducting the analysis with Lin's similarity score yields similar results as obtained with sim_{Rel}; AS-base performs slightly better, but AS-ortho and AS-inter are worse than with sim_{Rel}. This indicates that the performance of MedSim is robust with respect to the selected semantic similarity measure.

The results obtained on benchmark set 1 confirm that MedSim effectively assigns top ranks to the correct protein in a list of known disease proteins. Together with the ability to detect diseases with similar functional profiles, this indicates that MedSim is able to identify commonalities between diseases and their associated proteins. However, MedSim achieves only a rather low with sensitivity of up to 59 % on this benchmark set.

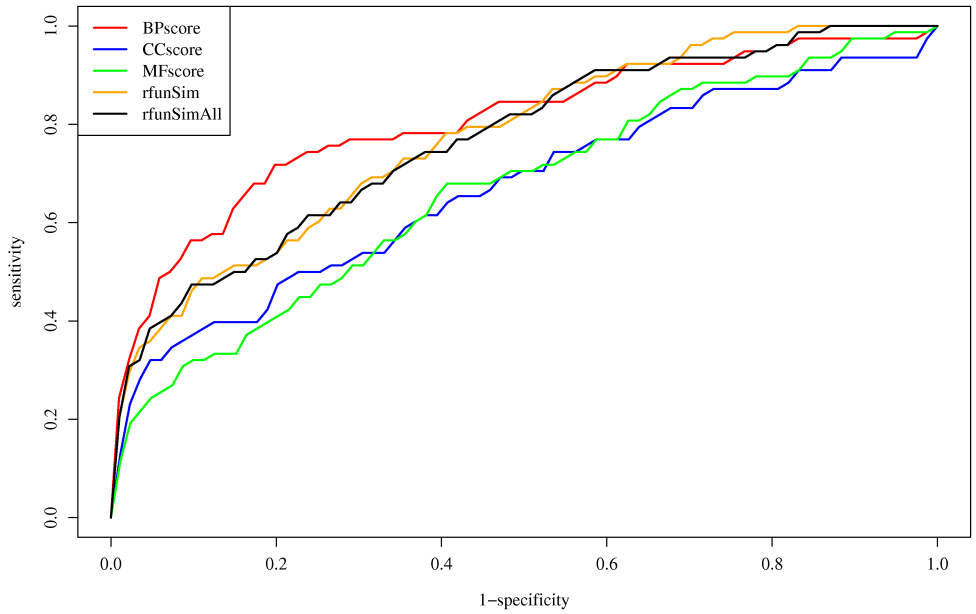

Figure 6.5: ROC plot for benchmark sets 1 annotated with annotation strategy AS-base. The ROC curves show the results of predicting the correct disease gene using the functional similarity scores *BPscore*, *CCscore*, and *MFscore* as well as *rfunSim* and *rfunSimAll*.

Table 6.2: Number of functional profiles in benchmark set 2 that contain terms from a given ontology (BP/MF/CC) using different annotation strategies. The columns for leave-one-out (LOO) give the number of cases in which the disease and the randomly selected disease protein are both annotated with the respective ontology. The total number of cross validations is 519. The aQTL columns show the number of annotated proteins averaged over all rQTLs that could be ranked using the respective annotation strategy.

Annotation strategy	BP		MF		CC	
	LOO	aQTL	LOO	aQTL	LOO	aQTL
AS-base	408.0	167.3	395.0	173.4	334.0	159.7
AS-ortho	426.0	178.3	409.0	184.3	357.0	174.0
AS-inter	483.0	175.7	473.0	184.6	470.0	165.8

Benchmark set 2 was designed to resemble the most common application scenario for disease gene prioritization methods. Given a list of putative disease genes or proteins the task is to rank as high as possible the truly disease-associated proteins. Benchmark set 2 contains 519 aQTLs of size 10 Mbp, which encompass 312 proteins on average, including

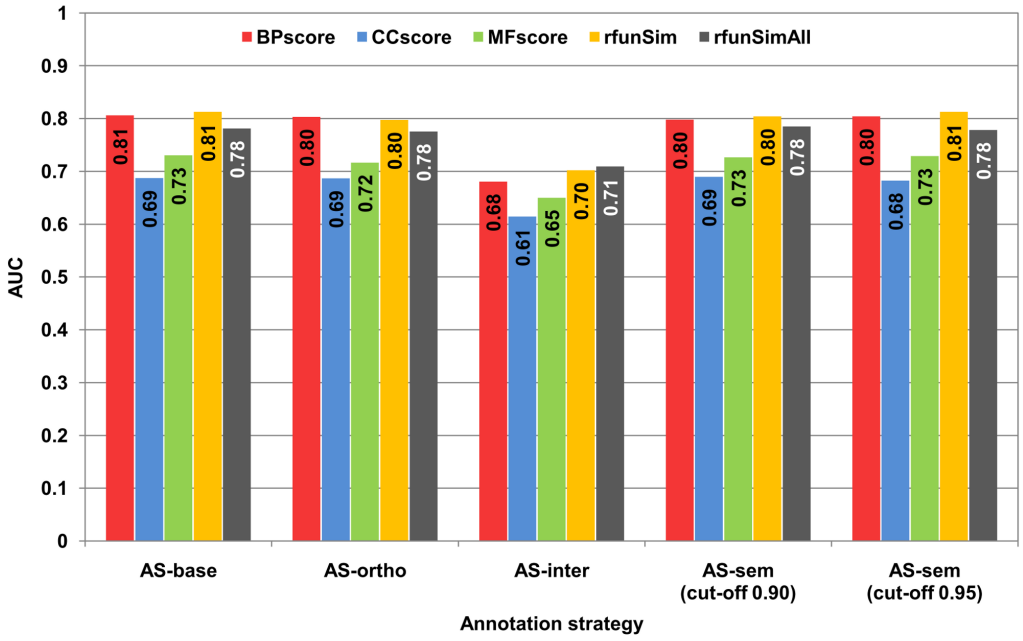

Figure 6.6: AUC values for prioritization of benchmark set 2 with different annotation strategies. The bars depict the AUC values achieved by MedSim utilizing the different functional similarity scores when the functional profiles were derived using AS-base, AS-ortho, AS-inter, or AS-sem (cut-off 0.90 or 0.95).

one known disease protein. Again, FSST was used to calculate the different functional similarities between phenotypes and the proteins in the corresponding aQTLs. Table 6.2 compares the number of diseases and proteins in the aQTLs that are annotated with GO terms by the different annotation strategies. The AUC values achieved by MedSim on benchmark set 2 using the different functional similarity measures are summarized in Figure 6.6. Using strategy AS-base, the best prediction AUC of 0.81 is achieved by the *BPscore* and the *rfunSim* score with a sensitivity of 0.51 and 0.50, respectively (Figure 6.7). AUC and sensitivity values remain virtually unchanged if annotation from orthologs is added. However, prediction performance using *MFscore* drops slightly, which also affects the results obtained with the *rfunSim* score. AS-inter performs slightly worse, the best AUC being 0.71 for the *rfunSimAll* score. Sensitivity, however, is only slightly decreased by adding protein interaction data; the highest sensitivity is 0.50, reached by the *BPscore*. From Table 6.2, it becomes evident that AS-ortho improves availability of GO annotation while it preserves the performance. AS-inter increases the coverage with functional annotation even more but has a negative effect on the prediction performance.

Despite the negative impact AS-inter has on the overall performance, the increased coverage potentially allows for accurately ranking candidate disease genes and proteins that are not amenable to analysis using AS-base due to the lack of direct GO annotation. We thus studied the results with the *rfunSim* and *rfunSimAll* scores for aQTLs to which we

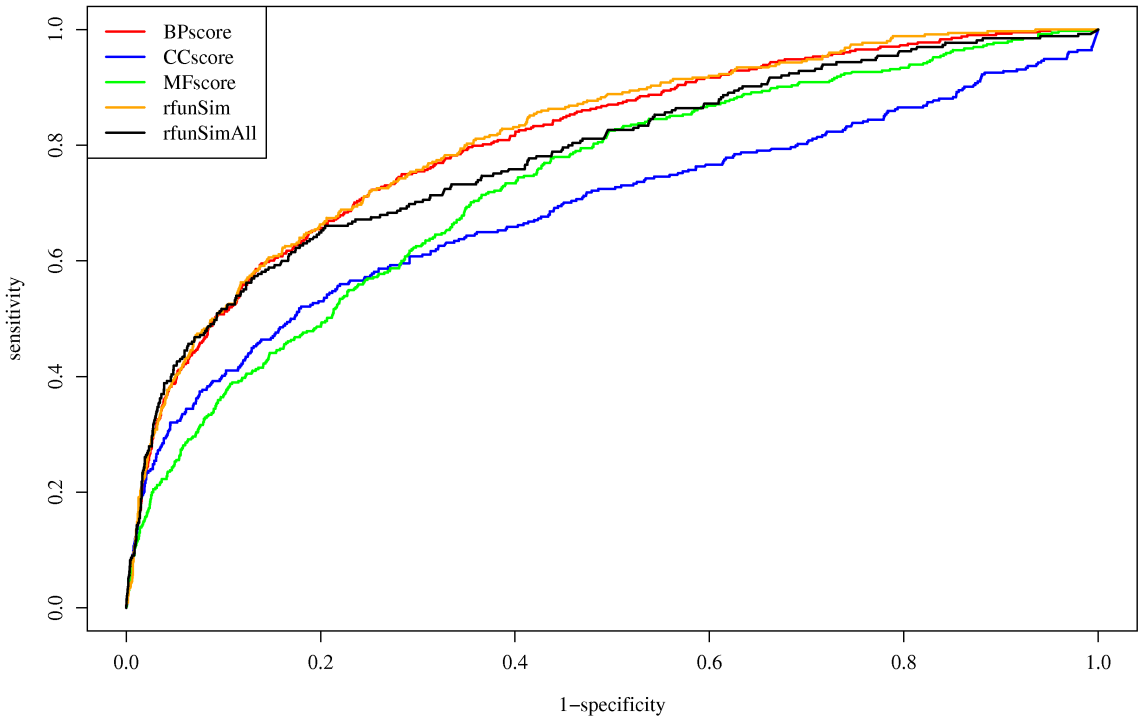

Figure 6.7: ROC plots for benchmark sets 2 annotated with annotation strategy AS-base. The ROC curves show the results of predicting the correct disease gene using the functional similarity scores *BPscore*, *CCscore*, and *MFscore* as well as *rfunSim* and *rfunSimAll*.

could not apply the strategy AS-base because of missing GO annotation. The sensitivity of MedSim using AS-ortho for these cases is 33 % and 46 % with *rfunSim* and *rfunSimAll*, respectively. Using AS-inter, MedSim has a sensitivity of about 25 % for both scores. These results indicate that both annotation strategies help ranking candidates if known human disease genes and proteins are not yet annotated with GO terms.

Annotation strategy AS-sem adds terms with high semantic similarity to terms already contained in the functional profile, preventing the functional profile from becoming too diverse. Although, coverage cannot be improved using AS-sem, this strategy reduces possible negative effects that the inclusion of additional data could have on prediction performance. We applied AS-sem to benchmark set 2 using two different sim_{Rel} cut-offs, 0.90 or 0.95, for adding terms. In both cases, the AUC and sensitivity values are similar to the performance obtained with AS-base.

6.3.3 Detailed Analysis of Performance Using AS-inter

The results described in the previous section suggest that strategies AS-base and AS-ortho achieve similar performance. They also annotate a similar number of GO terms to diseases (about 10 BP terms, 5 CC terms, and 5 MF terms) and proteins (about 1.2 terms on average from all three ontologies). In contrast, strategy AS-inter not only performs worse, but leads to approximately 10 times and 3 times as many annotations for phenotypes and proteins, respectively. The resulting large number of annotated GO terms may have a negative effect on prediction performance. However, there is only a low correlation between the rank of the correct disease protein and the number of GO terms in a functional profile (Spearman's correlation coefficient smaller than 0.3).

A second possibility is that strategy AS-inter creates more diverse functional profiles, which prevents MedSim from identifying the key similarities between diseases and proteins. As measure of the functional diversity of a profile, the average pairwise semantic similarity can be used, with low values indicating diverse profiles. Therefore, we examined the average semantic similarity of GO terms annotated to the diseases by strategies AS-base and AS-inter. The annotated BPs have a mean similarity of almost 0.41 (variance 0.13 and standard deviation 0.34) and 0.20 (variance 0.07 and standard deviation 0.25) when applying AS-base and AS-inter, respectively. Compared to BP, the drop in average semantic similarity is even larger for CC and MF. In both cases, the average decreases about 0.25 points (from 0.57 to 0.30 for CC and from 0.40 to 0.16 for MF). As for BP, the variances and standard deviations are also smaller for strategy AS-inter. This indicates that the functional profiles are consistently more diverse using AS-inter. Interestingly, the *MFscore* achieved not only the worst sensitivity, but also the smallest average semantic similarity using PPI information. However, we could not find a pronounced correlation between prediction rank of the correct disease protein and mean semantic similarity. The Spearman correlation coefficient is around -0.2 for BP and CC as well as 0.0 for MF in case of AS-base, and -0.27, -0.18 and -0.1 for BP, CC and MF, respectively, in case of AS-inter.

To further investigate the performance drop caused by incorporating PPI data, we randomized the PPI dataset by keeping one partner of each interaction fixed and randomly shuffling the interacting partners (Section 6.2.1). Figure 6.8 depicts the ROC curves obtained from benchmark set 2 after applying AS-inter to the set of random PPIs. It becomes obvious that all scores perform worse than using real PPI data (AUC values between 0.47 and 0.55). Additionally, we created three datasets by randomly exchanging 25 %, 50 %, or 75 % of the prioritizations with real PPIs with prioritizations with random PPIs. The higher the proportion of prioritizations with randomized PPIs is in these sets, the lower are the AUC values obtained. Figure 6.9 compares the AUC values achieved by MedSim with the varying fractions of prioritizations with randomized PPIs. These results show that PPI data can provide valid information for the prioritization of disease gene candidates although performance is not increased in general.

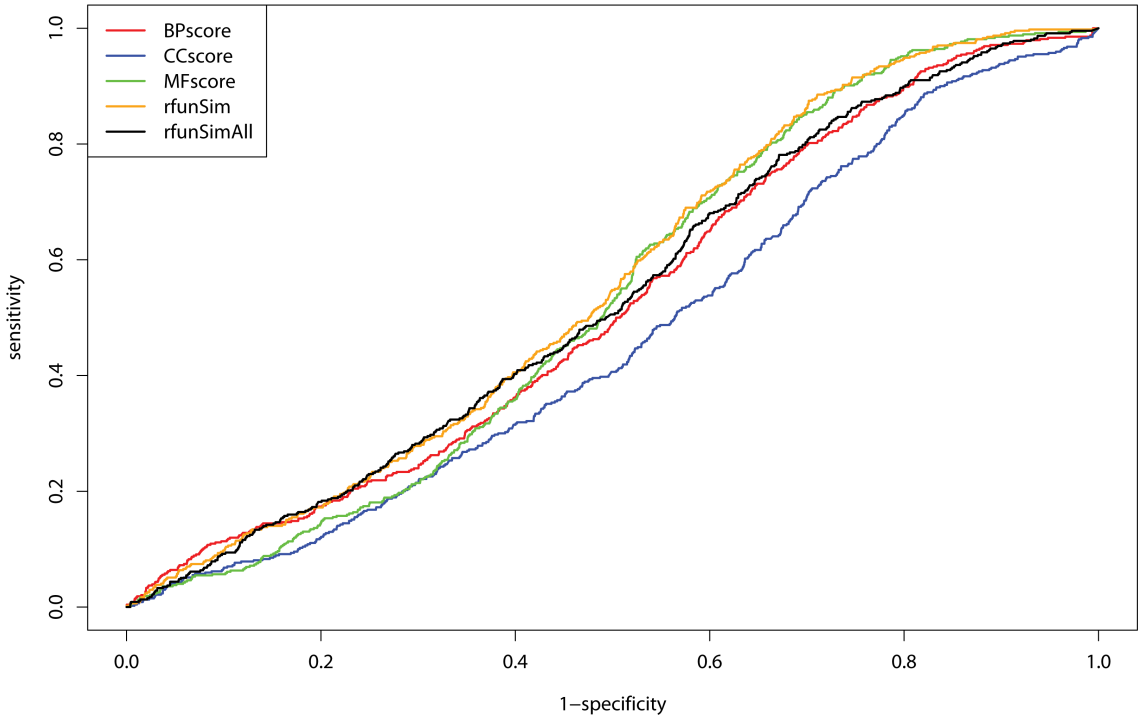

Figure 6.8: ROC plots for benchmark sets 2 annotated with annotation strategy AS-inter and a randomize set of PPIs. The ROC curves show the results of predicting the correct disease gene using the functional similarity scores *BPscore*, *CCscore*, and *MFscore* as well as *rfunSim* and *rfunSimAll*.

6.3.4 Improving Prediction Performance with Term Filtering

The previous results suggest that semantically unrelated terms negatively influence prediction performance. Therefore, we applied a semantic similarity term filter to functional profiles created by AS-base and AS-inter for benchmark set 2. This term filter removes all terms that do not have a sim_{Rel} score greater than a specific threshold (0.60 or 0.80) to any other term in the profile.

Figure 6.10 summarizes the AUC values achieved by MedSim when term filtering is combined with different annotation strategies. With respect to AUC, the results are inconclusive for AS-base. The AUC drops slightly for BP and MF using term filtering with both sim_{Rel} thresholds, but the AUC of CC and of the combined scores are larger than without term filtering. The best AUC is achieved using the *rfunSim* score (AUC 0.85) with AS-base and term filtering by the threshold 0.80. If the functional profiles are complemented by PPIs in AS-inter, term filtering improves the AUC in most cases. The *rfunSimAll* score achieves an AUC of 0.82 using AS-inter and term filtering (threshold 0.80), which is even better than the best performance of AS-base without term filtering. The sensitivity values show the same trend, and the maximum is reached at 65 % using AS-base with term

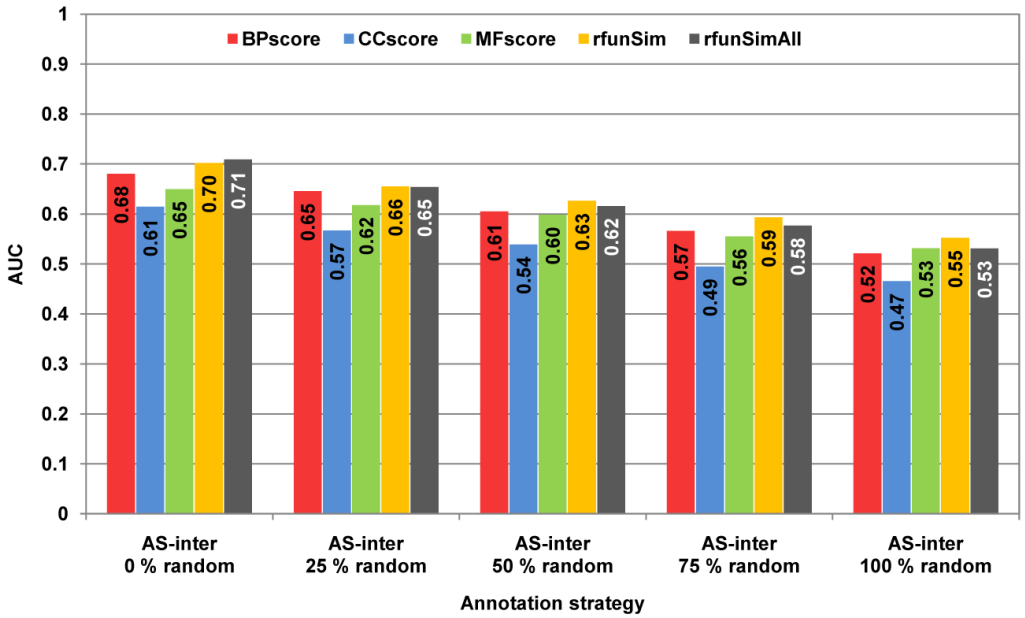

Figure 6.9: AUC values for prioritization of benchmark set 2 with AS-inter using increasing fractions of prioritizations with random PPIs. The bars depict the AUC values achieved by MedSim utilizing AS-inter when 0 %, 25 %, 50 %, 75 %, or 100 % of the predictions were made using random PPIs.

filtering (threshold 0.80). The annotation coverage of proteins in aQTLs is lower after applying the term filtering procedure, but when using AS-inter, it is about as high as with AS-base without term filtering. In case of the combined scores, *rfunSim* and *rfunSimAll*, however, the coverage is significantly lower using term filtering. Applying term filtering (threshold 0.80) before adding terms based on high semantic similarity (AS-sem) does not improve the results compared to AS-base with term filtering.

Above, we have already shown that adding terms from protein interaction partners helps ranking candidates in aQTLs that are not amenable to analysis with AS-base. When considering only cases in which the disease or the left-out protein could not be annotated using AS-base with term filtering (threshold 0.80), AS-inter with term filtering achieves a sensitivity of 31 % with *rfunSim* and 36 % with *rfunSimAll*. This further confirms that PPIs aid in identifying disease-related proteins if known human disease proteins are not annotated with GO terms.

6.3.5 Performance Increases with Randomized QTLs

Disease gene prioritization methods are commonly benchmarked using sets consisting of one protein known to be associated with a disease supplemented with random proteins. In order to facilitate a comparison to these methods (Section 6.3.8), we designed benchmark

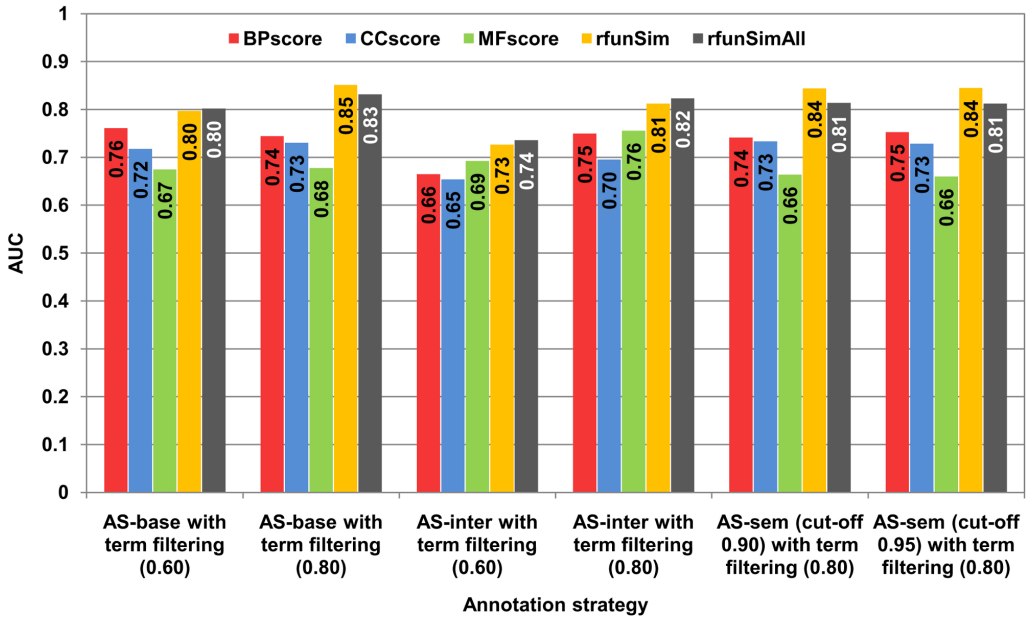

Figure 6.10: AUC values for prioritization of benchmark set 2 with different annotation strategies and term filtering. The bars depict the AUC values achieved by MedSim utilizing the different functional similarity scores when the functional profiles were derived using AS-base with term filtering (threshold 0.60 and 0.80), AS-inter (thresholds 0.60 and 0.80), or AS-sem (cut-off 0.90 or 0.95) and term filtering (threshold 0.80).

set 3. It consists of 287 rQTLs where each contains one known disease protein and 99 random proteins. When creating this benchmark set, we included only proteins that were annotated with terms from all three GO ontologies. The functional profiles were derived using AS-base without and with term filtering (threshold 0.80), and AS-sem (cut-off 0.95) with term filtering (threshold 0.80).

Using AS-base (Figure 6.11), the best performance is achieved with the combined scores, *rfunSim* (AUC 0.85) and *rfunSimAll* (AUC 0.84). The sensitivity of the combined scores (57 %) also improves over the sensitivity of any other score (42 % to 53 %). Applying term filtering deteriorates the AUC of the *BPscore* and the *MFscore* but increases the sensitivity of the *CCscore* from 42 % to 57 % and of the *MFscore* from 47 % to 51 %. In case of the combined scores, both performance measures improved upon AS-base without term filtering. The *rfunSimAll* score reaches the maximal AUC of 0.90 and a sensitivity of 73 %. When applying term filtering to AS-sem, no difference in performance can be observed. Figure 6.12 depicts the AUC values achieved by MedSim on this benchmark set. From this figure, it can be easily seen that the overall maximal AUC of 0.90 is achieved using AS-base with term filtering (threshold 0.90) using the *rfunSimAll* score and AS-sem (cut-off 0.95) with term filtering (threshold 0.90) using the *rfunSim* and *rfunSimAll* scores.

The impact that the removal of unrelated GO terms from functional profiles has on

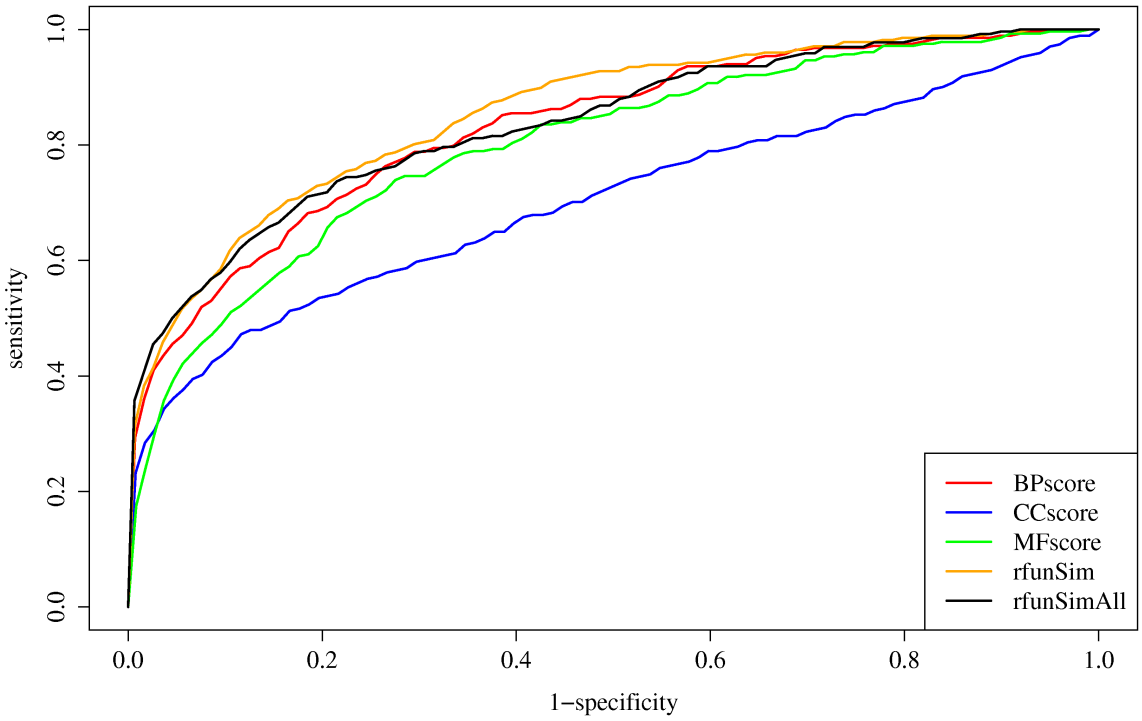

Figure 6.11: ROC plots for benchmark sets 3 annotated with annotation strategy AS-base. The ROC curves show the results of predicting the correct disease gene using the functional similarity scores *BPscore*, *CCscore*, and *MFscore* as well as *rfunSim* and *rfunSimAll*.

GO annotation coverage was already described previously for benchmark set 2. In case of benchmark set 3, term filtering reduces the coverage to between 36 % and 59 % of the cross validations for the single ontology scores (Table 6.2). To be able to calculate either the *rfunSim* score or the *rfunSimAll* score, the functional profiles have to contain terms from BP and MF, or all three ontologies, respectively. Consequently, term filtering has a higher impact on the combined scores and reduces the coverage to about 10 % compared to around 95 % without term filtering.

6.3.6 Potential Dataset Bias

Biomedical research preferentially targets known disease genes and their products, which can lead to biases in the datasets. First, interaction screens might include a high number of known disease associated proteins, and therefore, more interaction partners might be known for these proteins. About 80 % of the disease proteins used in cross validations participate in interactions in our dataset, compared to about 20 % of the proteins contained in QTLs. However, our results show that including interactions leads to higher coverage

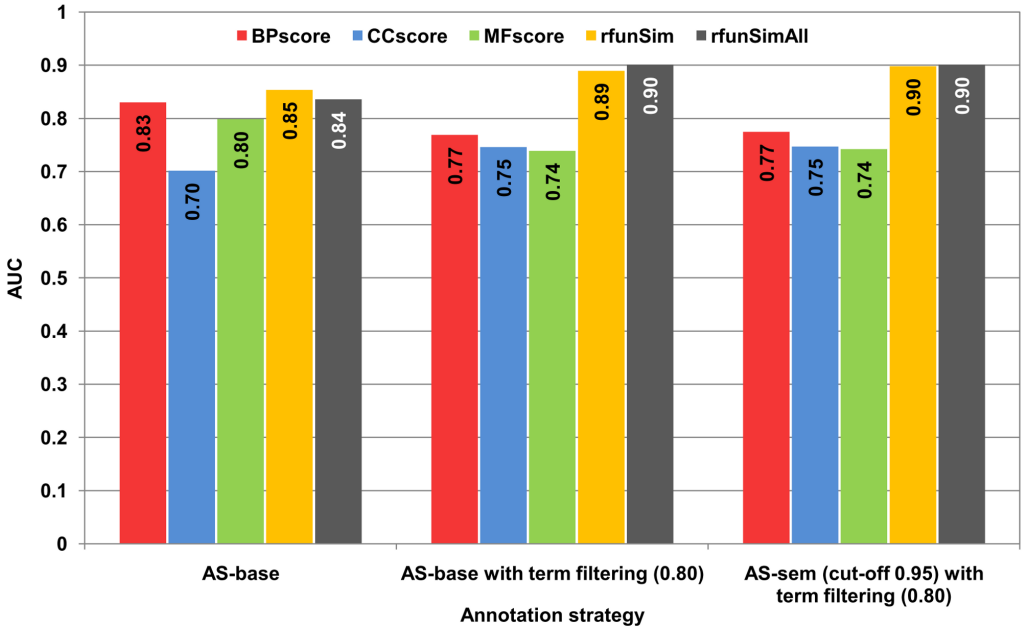

Figure 6.12: AUC values for prioritization of benchmark set 3 with different annotation strategies. The bars depict the AUC values achieved by MedSim utilizing the different functional similarity scores when the functional profiles were derived using AS-base without and with term filtering (threshold 0.80), or AS-sem (cut-off 0.95) and term filtering (threshold 0.80).

Table 6.3: Number of functional profiles in benchmark set 3 that contain terms from a given ontology (BP/MF/CC) using different annotation strategies. The columns for leave-one-out (LOO) give the number of cases in which the disease and the randomly selected disease protein are both annotated with the respective ontology. The total number of cross validations is 287. The rQTL columns show the number of annotated proteins averaged over all rQTLs that could be ranked using the respective annotation strategy.

Annotation strategy	BP		MF		CC	
	LOO	rQTL	LOO	rQTL	LOO	rQTL
AS-base	283.0	100.0	280.0	100.0	271.0	100.0
AS-base (term filtering)	127.0	86.8	101.0	88.0	161.0	97.0
AS-sem (term filtering)	127.0	86.8	101.0	88.0	161.0	97.0

but not increased performance. Second, disease genes might be more thoroughly annotated with GO terms than non-disease genes. Notably, the best performance was achieved using benchmark set 3, which was created using only proteins annotated with terms from

all three ontologies. From Table 6.3, it can be seen that in benchmark set 3 more disease proteins lack GO annotation than proteins in QTLs on average. Third, if disease genes were annotated with similar terms due to their linkage to the same disease, benchmark performance could be artificially boosted. According to the GO annotation guidelines, every annotation of a gene or gene product with a GO term is supplemented with an evidence code describing the support for this annotation (Section 2.1.3). These guidelines suggest to use the IGI (Inferred by Genetic Interaction) code if an annotation is based on the similarity of phenotypes resulting from mutations in more than one gene or gene product. However, only four proteins in our benchmark sets are annotated using this code, which makes an artificially boosting of the benchmark results unlikely.

6.3.7 Results for Exemplary Diseases

In the following, we present prioritization results for different exemplary diseases, which have been selected from different disease classes. The benchmark results showed that utilizing protein interaction data can improve candidate ranking as well as the coverage with GO annotation. One such example is photosensitive trichothiodystrophy (OMIM #601675), which causes brittle hair and nails, physical and mental retardation, and photosensitivity (Faghri *et al.*, 2008; Niedernhofer, 2008). UniProtKB contains three proteins that are known to be related to this disease: the TFIIH basal transcription factor complex helicase subunit (CXPD, P18074), the TFIIH basal transcription factor complex helicase XPB subunit (BTF2-p89, P19447), and the general transcription factor IIH subunit 5 (TFB5 ortholog, Q6ZYL4). The first two proteins are annotated with terms from all three ontologies, but the last one is completely lacking annotation. In benchmark set 2, MedSim ranks CXPD at the top of the list using all functional similarity scores. BTF2-p89 is placed within the top 5 % with all scores except the *MFscore*, which ranks the protein in the top 50 % using strategies AS-base, AS-ortho, and AS-inter. Classification of this protein using MF annotation fails because most MF terms in the functional profiles of the disease and of BTF2-p89 are only distantly related. Adding annotation from protein interaction partners and subsequent term filtering (AS-inter with term filtering), however, allows for ranking BTF2-p89 in the top 3 % of the list using the *MFscore*. The third protein, a TFB5 ortholog, lacks GO annotation and can only be analyzed using PPI data. Using AS-inter without term filtering, MedSim ranks this protein in the top 6 % using all functional similarity scores except the *CCscore* (top 11 %).

Several inherited diseases are influenced by cellular processes whose functional relationship on the molecular level has not been resolved yet, for instance inflammatory bowel disease (OMIM #266600) (Schreiber *et al.*, 2005). UniProtKB currently maps five of the proteins reported by genome-wide association studies to this disease (Cho, 2008): the nucleotide-binding oligomerization domain-containing protein 2 (NOD2, Q9HC29), the solute carrier family 22 members 4 and 5 (SLC22A4, Q9H015; SLC22A5, O76082), interleukin 10 (IL10, P22301), and the interleukin 23-receptor (IL23R, Q5VWK5). All proteins except NOD2 are ranked in the top 22 % by MedSim applying strategy AS-inter

and the *rfunSimAll* score. Notably, SLC22A5 and SLC22A4 are ranked in the top 6 % and top 11 %, respectively. NOD2 is ranked in the top 11 % using the *rfunSim* score and strategy AS-base. We compared these results to prioritizations obtained from ToppGene using GO and interaction information, applying the default options. With this configuration, ToppGene is able to rank the two solute carrier family 22 members at the top of the list. IL10 and NOD2 are ranked only in the top 66 %, and IL23R cannot be ranked at all because it is not contained in the database. This example demonstrates that, despite its simple approach, MedSim can improve on the results and coverage of other methods.

OMIM also contains information on mutations that do not cause genetic diseases, but have an effect on susceptibility to infections with pathogens. Consequently, MedSim could be used for discovering proteins that modulate the risk of viral or bacterial infections. The OMIM entry #609423 subsumes genes that influence susceptibility and resistance to human immunodeficiency virus type 1 (HIV-1, Brass *et al.*, 2008). In UniProtKB, three human proteins are cross-referenced with this phenotype: C-C chemokine receptor type 2 (CCR2, P41597), CX3C chemokine receptor 1 (CX3CR1, P49238), and C-C chemokine receptor type 5 (CCR5, P51681). In benchmark set 2, CX3CR1 is ranked within the top 10 % of its 10 Mbp aQTL by all functional similarity scores except the *BPscore*, which ranks the protein in the top 15 %. The other two proteins are ranked in the top 40 % of their respective aQTLs. This difference is observed in case of annotation strategies AS-base, AS-ortho, and AS-sem. Concerning AS-inter, *BPscore* and *CCscore* rank CX3CR1 only in the top 30 % and top 43 %, respectively. If an rQTL in benchmark set 3 is used instead, MedSim ranks the three proteins in the top 10 % using AS-base and the *rfunSim* score or the *rfunSimAll* score. This example illustrates that MedSim is, in principle, able to accurately identify human proteins involved with infectious diseases. However, it is to note that not all human proteins relevant for HIV-1 infection are currently mapped to the used OMIM entry #609423. One such exception is the C-X-C chemokine receptor type 4 (CXCR-4, P61073), which is of interest as potential target for future drugs (Hunt and Romanelli, 2009).

Our benchmark sets contain 13 proteins associated with familial hypertrophic cardiomyopathy (OMIM #192600). From these, 10 proteins are annotated with BP and MF terms and are ranked in the top 10 % utilizing this annotation with AS-base and the *rfunSim* score. Importantly, eight of these 10 proteins are ranked within the top 3 % of the list. One more protein is ranked on top of the list using either its annotated MF or CC terms. Seven proteins associated with the disease are annotated using all three ontologies and ranked within the top 3 % of the list using this information.

6.3.8 Comparison with Other Prioritization Methods

Several aspects hamper a fully objective comparison between different disease gene prioritization methods. Many methods are not readily available making it impossible to apply them on exactly the same benchmark set. Furthermore, the biological content of the

datasets used by the different methods influences the prediction results, and thereby, limits any detailed comparison. Nevertheless, it is possible to perform a general performance comparison by utilizing large-scale benchmark sets that are created in a methodologically similar way. To facilitate such a general performance comparison, the approach for creating benchmark set 3 is methodologically similar to previous publications.

Endeavour is a state-of-the-art method based on the integration of multiple data sources (Aerts *et al.*, 2006). It allows for prioritizing genes based on single data sources or combinations of selected sources. Endeavour was validated using a set of rQTLs constructed in a similar way as benchmark set 3. With GO annotation as the only data source, Endeavour achieved an AUC of slightly above 0.75. MedSim, on the other hand, reached an AUC value of up to 0.90 at a sensitivity of 73 % when relying only on GO annotation. For prioritization using all data sources, Aerts *et al*. reported an AUC value of 0.87 and a sensitivity of 74 % (at 90 % specificity), which is comparable to the performance of MedSim. Therefore, it appears that MedSim is able to significantly outperform Endeavour when prioritization is based solely on GO annotation. MedSim achieves a slightly better AUC and similar sensitivity compared to Endeavour using all available data sources.

Recently, Chen *et al*. devised the ToppGene method (Chen *et al.*, 2007). Their approach integrates annotation with terms from the Mammalian Phenotype (MP) ontology (Smith *et al.*, 2005) with other data sources, such as biomedical literature and protein interactions. The employed benchmark is comparable to benchmark set 3. The authors reported AUC values of 0.91 and 0.89 with and without using MP annotation, respectively, and a sensitivity of 74 % with MP annotation. This means that MedSim performs comparatively while using a much simpler prediction model with GO annotation alone. Relying only on GO annotation is one of the major benefits of the MedSim method because GO annotation is available for many genes and proteins. This allows to apply the method in a broad range of settings and still be able to achieve the best possible performance. Notably, in contrast to GO annotation, the availability of MP annotation seems to be quite limited for new candidate disease genes. The dataset used by Chen *et al*. contained only 4,280 mouse genes annotated with phenotype terms. Since human genes are not directly annotated with MP terms, they have to be transferred from mouse orthologs. It is likely that this negatively affects the performance of ToppGene.

In contrast to the previously discussed methods, PROSPECTR leverages sequence features to distinguish between disease and non-disease genes (Adie *et al.*, 2005). Adie *et al*. selected a set of about 1,000 known disease genes and about 18,000 non-disease genes for training and validating their method using ten-fold cross validation on this set. The resulting AUC was reported to be 0.70. Since MedSim achieves significantly better results than this sequence-based approach, it becomes evident that functional features are better predictors for gene-disease relationships.

Ortutay and Vihinen developed a method leveraging GO annotation and protein interactions in a way that is fundamentally different from the approach taken by MedSim (Ortutay and Vihinen, 2009). First, proteins are selected based on three different network

topology parameters, that is, degree, vulnerability, and closeness centrality, and then, enrichment of GO terms in the selected sets of proteins is calculated. Finally, genes are predicted to be associated with an immune disease if they receive significant values for the network parameters and are annotated with enriched GO terms. Benchmarking was conducted by cross validation with 144 genes related to primary immunodeficiency (PID). From those 144 known genes, 84 could be ranked within the top 50 in the respective cross validation run, which corresponds to ranking the known gene in the top 20 % in about 59 % of the cases. MedSim performs significantly better and ranks 85 % of the proteins in the top 20 % using the *rfunSim* score on benchmark set 3 annotated with AS-sem and term filtering (threshold 0.80).

Chen *et al*. recently studied the applicability of methods developed for the analysis of social and web networks to disease gene prioritization (Chen *et al.*, 2009). They applied the PageRank and HITS algorithms as well as the K-step Markov method to prioritization of candidate disease genes based on a protein interaction network. They conclude, however, that network-based methods are inferior to functional annotation-based approaches. Our benchmark results support this conclusion as the network methods by Chen and colleagues achieve an AUC of up to 0.80, which is below the best AUC 0.90 of MedSim.

6.4 Conclusions

In this chapter, we described the new method MedSim for disease gene prioritization. We introduced several novel strategies for automatically annotating diseases with GO terms from known disease genes or proteins and their mouse orthologs or interacting human proteins. In addition, we explored the possibility of increasing prediction performance by enriching the functional profiles by adding semantically similar terms and filtering of dissimilar terms. The results obtained with several extensive benchmark experiments show that MedSim is able to specifically associate diseases with known proteins. MedSim achieves high AUC (up to 0.90) and sensitivity (up to 73 %) values and performs at least as well as more complex state-of-the-art methods like Endeavour (Aerts *et al.*, 2006) and ToppGene (Chen *et al.*, 2007). MedSim further significantly outperforms other recent methods using GO annotation and interaction data (Chen *et al.*, 2009; Ortutay and Vihinen, 2009) as well as sequence-feature based methods like PROSPECTR (Adie *et al.*, 2005). Additionally, our results suggest that prediction performance is dependent on the methodological details of the construction of the benchmark sets. Moreover, we find that functional similarity can be used to distinguish diseases with a common functional basis from unrelated diseases, which enables clustering diseases based on functional criteria.

Using the different benchmark sets and annotation strategies, the functional similarity scores *BPscore*, *rfunSim*, and *rfunSimAll* overall performed best. Transferring GO annotation of mouse orthologs to functional profiles proved useful for increasing the coverage with GO annotation without lowering performance. Adding annotation from protein interaction partners greatly increased coverage (up to 41 %), but can have a negative impact

on the overall performance. Nevertheless, our results provide evidence for the fact that the use of GO annotations from orthologous mouse proteins or protein interaction partners aids in ranking candidate genes and proteins accurately if the latter have not yet been mapped to GO terms. In particular, term filtering increases the performance and allows for finding a tradeoff between high coverage and high performance. This is especially important if the functional profiles are created with the help of protein interaction data.

Generally, our comparison of prediction results obtained with different benchmarks demonstrated that the performance of a method depends on the actual construction of the benchmark set. The AUC and sensitivity values for benchmark set 3 are generally higher than for benchmark set 2 using the same annotation strategy for both sets. This effect was also observed in our exemplary study of susceptibility to HIV-1. Likely reasons for this observation are that the rQTLs in benchmark set 3 contain fewer proteins on average and that the unrelated proteins are randomly drawn from the whole proteome. Therefore, it is important to take into account how a benchmark set was constructed when comparing the performance of different prioritization approaches.

All benchmarks used for validating the MedSim approach were constructed such that every candidate list contains exactly one true positive. In reality, however, it is possible that none of the candidates is related to the disease of interest. In such situations, the whole list might be rejected if no candidate scores significantly better than the rest of the candidates. If the functional similarity scores obtained for different disease are compared, it is important to normalize the absolute values because they are not directly comparable.

One major benefit of MedSim is that its prioritization procedure relies only on the presence of GO annotations. In contrast to previously published approaches that utilize GO annotations, MedSim introduces two new features. First, it automatically annotates diseases with functional terms, making them suitable for large-scale analysis. Moreover, if the molecular basis of the disease of interest is yet unknown, GO terms can be manually added to the functional profile, which allows for applying MedSim without prior knowledge of disease genes or proteins. Alternatively, text-mining techniques may be used to extract GO terms from full-text descriptions or scientific articles about the disease of interest. This is in contrast to most existing prioritization methods, which necessarily require a set of known proteins for training. Second, MedSim makes use of functional similarity measures instead of exact matching and overrepresentation analysis. These similarity measures allow for quantifying the similarity between seemingly unrelated GO terms. Therefore, MedSim is able to identify and quantify more distant similarities between candidate disease genes and diseases.

In addition, we presented strategies for automatically extending the existing GO annotation of human genes and proteins using orthologs from model organisms or interaction partners. Notably, our approach is not restricted to GO as functional annotation source. Since the semantic and functional similarity measures used are applicable to any vocabulary that is organized as a tree or directed acyclic graph, MedSim could also leverage annotations with other vocabularies like FunCat (Ruepp *et al.*, 2004) or the Human Phe-

notype Ontology (Robinson *et al.*, 2008). Importantly, the availability of functional annotations is expected to improve considerably in the near future because of comprehensive annotation efforts like the Reference Genome Annotation Project (Reference Genome Group of the Gene Ontology Consortium, 2009).

Finally, the most promising MedSim annotation strategy, AS-base with term filtering (threshold 0.80), is available via our FunSimMat online service (Chapter 4.4). In particular, FunSimMat contains functional profiles for all OMIM entries and human proteins derived by annotation strategy AS-base without and with term filtering (threshold 0.80). The precomputation of functional similarity scores affords the fast ranking of genes in QTLs or even the whole genome with respect to the disease of interest.

Chapter 7

Conclusions

In this closing chapter, we summarize the work presented in this thesis and give a concise overview of our newly developed methods for utilizing semantic and functional similarity in different applications. Finally, possible future directions for research on these similarity measures and their applications will be discussed.

7.1 Summarizing Remarks

The wide-spread adoption of ontologies in the biomedical domain and the resulting availability of knowledge in a standardized, computer-readable format holds great promises for improving existing methods and for developing new approaches. There are, however, several challenges that have to be addressed. First, it is essential to develop advanced semantic and functional similarity measures that allow for performing detailed comparisons of annotations with ontology terms. It is likely that these similarity measures perform differently in varying applications. Therefore, the similarity measures have to be compared amongst each other and to established methods for identifying possible improvements and applications. Second, software is required that enables biological and medical researchers to easily access and apply these similarity measures. Finally, new methods have to be developed that take advantage of the available ontological annotation. With the work presented in this thesis, we contribute to these three challenges.

In Chapter 3, we presented an extensive analysis of our semantic and functional similarity measures. Using several sets of protein pairs with varying levels of sequence similarity, we compared our functional similarity to homology detection methods and to a previously published functional similarity measure. Then, we applied our new approaches in different medically relevant settings that can help to identify new targets for anti-infective drugs. We assessed the differences in the biological processes and molecular functions of various taxa, and determined functionally similar proteins between human and the yeast *Saccharomyces cerevisiae*. Furthermore, we provided maps of the functional space of yeast proteins and of Pfam families derived from an analysis of MF annotations. Method-

ological extensions of our functional scores presented in Section 4.2 allow for comprehensively quantifying the functional similarity of entities that are annotated with GO terms.

Many different semantic and functional similarity measures have been published, but available tools have limitations and a comprehensive web service was lacking. Moreover, the increasing number of available annotations made it necessary to develop new tools that allow for efficiently utilizing this information. To address the issues of existing programs, we developed the Functional Similarity Search Tool (FSST, Section 4.3) and the Functional Similarity Matrix (FunSimMat, Section 4.4). FSST is applicable to a wide range of tasks due to several unique features. These include the ability to utilize annotations that are not publicly available, its multi-threaded design, and the embedded database. Since its publication, FSST has been downloaded more than 140 times and was cited in 5 scientific publications. FunSimMat is the first comprehensive database of precomputed semantic and functional similarity measures. It provides several software interfaces for manually and automatically accessing the similarity scores and supports a large number of similarity measures. This enables users to test several measures in a specific application. Since it became publicly available, FunSimMat received almost 2-million queries from about 300 users and was cited in 15 published papers, which illustrates its relevance for the community.

Two important applications of functional similarity measures are the analysis of interaction networks and the prioritization of disease genes and proteins. In Chapter 5, we analyzed predicted and experimentally derived domain-domain interaction datasets. We could show that domain pairs with experimentally verified interactions have a very high functional similarity. Based on this observation, we were able to infer confidence score thresholds for dividing predicted interactions into subsets of low- and high-confidence. Further, we analyzed a number of predicted and experimental human protein-protein interaction datasets. Our results indicate that interacting proteins in manually curated interaction sets have a higher average functional similarity than interactions in predicted datasets. This could either point to the fact that the latter sets contain more false positive interactions or that many proteins are involved in novel processes.

The new MedSim method described in Chapter 6 prioritizes candidate disease genes or proteins with respect to their functional similarity to the disease of interest. In order to be able to apply functional similarity measures, we introduced annotation strategies that automatically create a functional profile of a disease. This functional profile contains GO terms that are annotated to known disease genes or proteins. Using three extensive benchmark sets, we could show that MedSim performs at least as well as more complex state-of-the-art methods and significantly outperforms other recently developed approaches. The restriction to functional annotation facilitates the interpretation of the prioritization results. Furthermore, MedSim can be applied as stand-alone method and also be integrated into other prioritization methods.

In conclusion, semantic and functional similarity measures proved to be valuable new approaches for complementing established bioinformatics methods. The development

of new tools was a necessary prerequisite for successfully employing these similarity measures in new application scenarios.

7.2 Perspectives

Currently, the most commonly used ontologies in the biomedical domain are the vocabularies provided by the Gene Ontology Consortium. Although coverage of gene product annotations with these ontologies is still far from complete, large-scale annotation projects such as the Reference Genome Annotation Project (Reference Genome Group of the Gene Ontology Consortium, 2009) will increase the amount of available data in the near future. This creates the opportunity for using semantic and functional similarity measures in new applications.

Generally, ontologies aim at representing the current state of knowledge in a specific research area. In order to achieve this goal, they are steadily developed further; new terms are added, and relationships between existing terms are constantly redefined. Identifying inconsistencies and underdeveloped parts of the ontology is an important step towards this goal. Very recently, Alterovitz *et al*. used the information content to identify suboptimally organized areas in the Gene Ontology and to automatically optimize the ontology structure (Alterovitz *et al.*, 2010). In Section 2.2.2, we described two empirical approaches for determining the probability of an ontology term to occur, annotation-based (p_{anno}, Equation 2.2) and graph-based (p_{graph}, Equation 2.3). In both cases, a term is thought to be more generic, the higher its probability is. The first measure determines the probability based on the number of times a term occurs in a large annotation database. Therefore, the measure is influenced by annotation biases in this database. The second measure uses the ontology hierarchy to obtain the probability of a term, which makes it prone to biases in the ongoing development of the ontology. A direct comparison of both approaches may help to identify terms that are either very frequently used or have only few descendants although being general. These terms might be good starting points for further development of the ontology.

A variety of semantic similarity approaches have already been proposed for comparing two terms from an ontology. However, as outlined in Chapter 3 and the recent literature (Pesquita *et al.*, 2009; Xu *et al.*, 2008), a thorough comparison of methods for calculating semantic and functional similarity measures is hampered by the lack of positive and negative gold standard sets. Semantic similarity measures have been evaluated by comparing the resulting scores to similarities assigned by human experts (Resnik, 1995; Lin, 1998; Jiang and Conrath, 1997). Manually curated sets of terms from GO or other biomedical ontologies could facilitate a comparison of different measures and an assessment of their applicability to certain problems. Similarly, sets of functionally similar and dissimilar pairs of proteins and protein families that were not derived using ontological annotation could serve as standards for assessing the performance of approaches for quantifying functional similarity. Furthermore, a gold standard would allow for identi-

fying the most appropriate method for different applications, for instance, prediction and evaluation of interacting proteins or families, for finding functionally similar proteins in different organisms, or for prioritizing candidate disease genes.

One of the first applications of GO-based functional similarity measures was the prediction and validation of protein and domain interactions. In Chapter 5, we summarized our analysis of domain and protein interactions, which was based on available annotations with BP and MF terms but excluded information on the cellular component. In order to test whether detected interactions are biologically meaningful, however, it is also important to take into account the cellular compartments that two domains or proteins occur in. Therefore, the *funSimAll* and *rfunSimAll* scores may be used for improving the prediction and validation of interactions and for deriving better confidence thresholds for predicted interactions.

The increasing availability of protein interaction networks for several species has prompted the development of network alignment methods, which, for instance, allow for detecting conserved functional modules and protein complexes across interaction networks from different species (Flannick *et al.*, 2006; Singh *et al.*, 2008). Most of the currently available approaches apply sequence similarity to determine nodes from two interaction networks that should be aligned. Ontology-based similarity may be used in such methods to align nodes based on their shared functional features. Recently, Ali and Deane devised a network alignment method that takes into account functional similarity based on the BP ontology (Ali and Deane, 2009). They could show that functional mapping of proteins across species networks improves functional coherence of the resulting protein modules and their overlap with experimentally verified complexes. Semantic similarity based on annotation with phenotype and disease ontologies could also be integrated into network alignment methods for creating a new type of prioritization methods for candidate disease genes. Additionally, alignments of interaction networks from human and model organisms using similarity based on phenotype ontologies could be applied to identify genes and proteins in other organisms that can be used to model for human diseases.

In Chapter 6, we described our new MedSim method, which applies GO-based functional similarity for prioritizing candidate disease genes or proteins. However, this annotation is not yet available for all human genes and proteins. To alleviate this problem, we utilized GO annotations from orthologous or interacting proteins. Since most biological knowledge is available in the form of scientific publications written in free text rather than ontological annotations, text-mining methods could be utilized for annotating genes and proteins with terms from ontologies (Yuan and Zhou, 2008; Yu *et al.*, 2010).

There are several other possibilities for improving the MedSim approach in the future. In our benchmarks, MedSim performed better if the true disease protein was embedded in a set of random human proteins (rQTLs in benchmark set 3) than if it was contained in a set of proteins encoded by genes from a single chromosomal region (aQTLs in benchmark set 2). This property could be exploited in a two-step prioritization approach that may reduce the number of false positives. In the first step, sets consisting of one candi-

date protein and random proteins are scored against the disease and ranked accordingly. In the second step, only candidates that received a high rank in the first step are taken into account for producing the final ranking. This stepwise approach could help to remove potential false positives and improve the prediction performance of MedSim and other prioritization methods. A second possibility for improvement is to consider annotations with terms from other ontologies to more accurately assess the similarities between a disease and the candidate genes and proteins. Examples for such ontologies are the Human Phenotype Ontology (HPO, Robinson *et al.*, 2008) and the Mammalian Phenotype Ontology (MPO, Smith *et al.*, 2005). In a recent paper, Chen *et al.* employed annotations with terms from the MPO among other information for prioritizing candidates in their ToppGene method (Chen *et al.*, 2007). If annotations with these two ontologies become readily available, the MedSim scoring scheme could be extended to take them into account.

All above mentioned applications of semantic and functional similarity depend on the availability of high-quality, ontological annotation. This is mostly obtained by manual curation, but this process is costly and time-consuming. Since the number of scientific publications and the availability of high-throughput datasets is steadily increasing, there is a growing need for automatic methods aiding in the curation process (Altman *et al.*, 2008). To this end, text-mining approaches are commonly utilized (Krallinger *et al.*, 2008). In 2004, Müller *et al.* developed the Textpresso system, which allows for extracting gene-gene interactions from text (Müller *et al.*, 2004). Kuhn *et al.* applied text-mining to create a database connecting drugs with side effects extracted from the package inserts (Kuhn *et al.*, 2010). Recently, Wiegers *et al.* described their efforts for utilizing text-mining for prioritizing publications for manual curation to extract relationships between genes, chemicals, and diseases (Wiegers *et al.*, 2009). A related task for text-mining applications is the automatic extraction of ontology terms from publications (Blaschke *et al.*, 2005; Camon *et al.*, 2005). In both settings, semantic similarity may help to increase the effectiveness of current methods. One possible approach is to augment the list of identified ontology terms with semantically similar terms, thereby helping human curators to identify the correct terms. Additionally, by improving the identification of important sentences, semantic similarity can be employed to support the prioritization of publications for manual curation.

Another research area where semantic similarity could be successfully applied is the development of the semantic web in general, and in particular, the semantic web for biomedical applications. By enhancing data repositories with computer-interpretable semantics, these efforts aim at improving the interoperability of different data sources and at enhancing search functionalities. Three examples for such systems are the Ontological Gene Orthology (OGO) system (Miñarro-Gimenez *et al.*, 2009), GoPubMed (Doms and Schroeder, 2005), and GoWeb (Dietze and Schroeder, 2009). The OGO system consists of three parts, an ontology about orthology, an orthology knowledge base, and an user-interface for querying this semantic repository. GoPubMed and GoWeb perform a search using standard technology and identify ontology terms in the search results, which can then be used for refining the search. Semantic similarity can be applied in several ways

for improving the results returned by such tools. First, terms that are semantically similar to the ones identified in the search results can help the user to refine the query. This approach was taken in the Phenomizer developed by Köhler *et al.* (2009), which is a system that was designed for improving the diagnosis of inherited diseases. Second, an automatic expansion of the query with semantically similar terms might improve the results without further user action (Voorhees, 1994; Liu *et al.*, 2008).

In conclusion, the importance of ontologies in biomedical research is constantly increasing. Semantic and functional similarity measures facilitate the utilization of ontological annotation in diverse applications. As more and more biological knowledge is captured in a computer-accessible format, new methods can be developed to test existing and generate new hypotheses by automatic inference.

Bibliography

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J., and Pickard, B.S. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 2006. **22**(6):773–774.

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J., and Pickard, B.S. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 2005. **6**:55.

Aerts, S., Lambrechts, D., Maity, S., Loo, P.V., Coessens, B., Smet, F.D., Tranchevent, L.C., Moor, B.D., Marynen, P., Hassan, B. *et al.* Gene prioritization through genomic data fusion. *Nat Biotechnol*, 2006. **24**(5):537–544.

Al-Halimi, R., Berwick, R.C., Burg, J.F.M., Chodorow, M., Fellbaum, C., Grabowski, J., Harabagiu, S., Hearst, M.A., Hirst, G., Jones, D.A. *et al.* *WordNet - An Electronic Lexical Database*. MIT Press, 1998.

Alexa, A., Rahnenführer, J., and Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 2006. **22**(13):1600–1607.

Ali, W. and Deane, C.M. Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics*, 2009. **25**(23):3166–3173.

Alterovitz, G., Xiang, M., Hill, D.P., Lomax, J., Liu, J., Cherkassky, M., Dreyfuss, J., Mungall, C., Harris, M.A., Dolan, M.E. *et al.* Ontology engineering. *Nat Biotechnol*, 2010. **28**(2):128–130.

Altman, R.B., Bergman, C.M., Blake, J., Blaschke, C., Cohen, A., Gannon, F., Grivell, L., Hahn, U., Hersh, W., Hirschman, L. *et al.* Text mining for biology–the way forward: opinions from leading scientists. *Genome Biol*, 2008. **9 Suppl 2**:S7.

Altshuler, D., Daly, M.J., and Lander, E.S. Genetic mapping in human disease. *Science*, 2008. **322**(5903):881–888.

Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res*, 2009. **37**(Database issue):D793–D796.

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J.P., Chothia, C., and Murzin, A.G. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res*, 2004. **32**(Database issue):D226–D229.

Antezana, E., Kuiper, M., and Mironov, V. Biological knowledge management: the emerging role of the Semantic Web technologies. *Brief Bioinform*, 2009. **10**(4):392–407.

Antezana, E., Tsiporkova, E., Mironov, V., and Kuiper, M. A Cell-Cycle Knowledge Integration Framework. In *Data Integration in the Life Sciences*. Springer (Berlin / Heidelberg, Germany), Hinxton, UK, volume 4075/2006 of *Lecture Notes in Computer Science*, 2006 pages 19–34.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 2000. **25**(1):25–29.

Azuaje, F., Wand, H., and Bodenreiter, O. Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*. Detroit, Michigan, USA, 2005 pages 9–10.

Bader, G.D. and Hogue, C.W.V. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 2002. **20**(10):991–997.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. The GOA database in 2009-an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*, 2009. **37**(Database issue):D396–D403.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., and Edgar, R. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 2007. **35**(Database issue):D760–D765.

Berglund, A.C., Sjölund, E., Ostlund, G., and Sonnhammer, E.L.L. InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 2007. **36**(Database issue):D263–D266.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res*, 2000. **28**(1):235–242.

Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J., and Elofsson, A. Domain rearrangements in protein evolution. *J Mol Biol*, 2005. **353**(4):911–923.

Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Group, M.G.D. The Mouse Genome Database genotypes::phenotypes. *Nucleic Acids Res*, 2009. **37**(Database issue):D712–D719.

Blaschke, C., Leon, E.A., Krallinger, M., and Valencia, A. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics*, 2005. **6 Suppl 1**:S16.

Bodenreider, O., Aubry, M., and Burgun, A. Non-lexical approaches to identifying associative relations in the gene ontology. *Pac Symp Biocomput*, 2005. **2005**:91–102.

Bodenreider, O. and Stevens, R. Bio-ontologies: current trends and future directions. *Brief Bioinform*, 2006. **7**(3):256–274.

Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I., and Marcotte, E.M. Protein interaction networks from yeast to human. *Curr Opin Struct Biol*, 2004. **14**(3):292–299.

Brameier, M. and Wiuf, C. Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps. *J Biomed Inform*, 2007. **40**(2):160–173.

Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., and Elledge, S.J. Identification of Host Proteins Required for HIV Infection Through a Functional Genomic Screen. *Science*, 2008. **319**(5865):921–926.

Brown, K.R. and Jurisica, I. Online predicted human interaction database. *Bioinformatics*, 2005. **21**(9):2076–2082.

Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. *et al.* The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*, 2003. **13**(4):662–672.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*, 2004. **32**(Database issue):D262–D266.

Camon, E.B., Barrell, D.G., Dimmer, E.C., Lee, V., Magrane, M., Maslen, J., Binns, D., and Apweiler, R. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*, 2005. **6 Suppl 1**:S17.

Cao, S.L., Qin, L., He, W.Z., Zhong, Y., Zhu, Y.Y., and Li, Y.X. Semantic search among heterogeneous biological databases based on gene ontology. *Acta Biochim Biophys Sin (Shanghai)*, 2004. **36**(5):365–370.

Caruana, R. and Niculescu-Mizil, A. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *KDD '04: Proceedings of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press (New York, NY, USA), Seattle, WA, USA, 2004 pages 69–78.

Caspi, R., Foerster, H., Fulcher, C.A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S.Y., Shearer, A.G., Tissier, C. *et al.* The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res*, 2008. **36**(Database issue):D623–D631.

Chabalier, J., Garcelon, N., Aubry, M., and Burgun, A. A transversal approach to compute semantic similarity between genes. In *Proceedings of the Workshop on Biomedical Ontologies and Text Processing - European Conference on Computational Biology*. Madrid, Spain, 2005 .

Chagoyen, M., Carmona-Saez, P., Gil, C., Carazo, J., and Pascual-Montano, A. A literature-based similarity metric for biological processes. *BMC Bioinformatics*, 2006. **7**(1):363.

Chatr-Aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. MINT: the Molecular INTeraction database. *Nucleic Acids Res*, 2007. **35**(Database issue):D572–D574.

Chen, J., Aronow, B.J., and Jegga, A.G. Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, 2009. **10**:73.

Chen, J., Xu, H., Aronow, B.J., and Jegga, A.G. Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, 2007. **8**:392.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D., and Siani-Rose, M.A. A knowledge-based clustering algorithm driven by Gene Ontology. *J Biopharm Stat*, 2004. **14**(3):687–700.

Chiang, J.H., Ho, S.H., and Wang, W.H. Similar genes discovery system (SGDS): Application for predicting possible pathways by using GO semantic similarity measure. *Expert Syst Appl*, 2008. **35**(3):1115 – 1121.

Cho, J.H. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat Rev Immunol*, 2008. **8**(6):458–466.

Cho, Y.R., Hwang, W., Ramanathan, M., and Zhang, A. Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 2007. **8**:265.

Cho, Y.R., Zhang, A., and Xu, X. Semantic similarity based feature extraction from microarray expression data. *Int J Data Min Bioinform*, 2009. **3**(3):333–345.

Choi, I.G., Kwon, J., and Kim, S.H. Local feature frequency profile: a method to measure structural similarity in proteins. *Proc Natl Acad Sci USA*, 2004. **101**(11):3797–3802.

Corazzon, R. Theory and History of Ontology. A Resource Guide for Philosophers. Web site, 2009.

Cordell, H.J. and Clayton, D.G. Genetic association studies. *Lancet*, 2005. **366**(9491):1121–1131.

Couto, F.M., Silva, M.J., and Coutinho, P.M. Measuring semantic similarity between Gene Ontology terms. *Data Knowl Eng*, 2007. **61**(1):137–152.

Deng, M., Mehta, S., Sun, F., and Chen, T. Inferring domain-domain interactions from protein-protein interactions. *Genome Res*, 2002. **12**(10):1540–1548.

Devos, D. and Valencia, A. Practical limits of function prediction. *Proteins*, 2000. **41**(1):98–107.

Devos, D. and Valencia, A. Intrinsic errors in genome annotation. *Trends Genet*, 2001. **17**(8):429–431.

Dietze, H. and Schroeder, M. GoWeb: a semantic search engine for the life science web. *BMC Bioinformatics*, 2009. **10 Suppl 10**:S7.

Domingues, F.S. and Lengauer, T. Inferring Protein Function from Protein Structure. In T. Lengauer, editor, *Bioinformatics - From Genomes to Therapies*, Wiley-VCH, Weinheim, volume 3, pages 1211–1252. 33 edition, 2007.

Doms, A. and Schroeder, M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*, 2005. **33**(Web Server issue):W783–W786.

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S.A., and Tainsky, M.A. Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, 2003. **31**(13):3775–3781.

Echt, S., Bauer, S., Steinbacher, S., Huber, R., Bacher, A., and Fischer, M. Potential anti-infective targets in pathogenic yeasts: structure and properties of 3,4-dihydroxy-2-butanone 4-phosphate synthase of Candida albicans. *J Mol Biol*, 2004. **341**(4):1085–1096.

Faghri, S., Tamura, D., Kraemer, K.H., and Digiovanna, J.J. Trichothiodystrophy: a systematic review of 112 published cases characterises a wide spectrum of clinical manifestations. *J Med Genet*, 2008. **45**(10):609–621.

Feldman, I., Rzhetsky, A., and Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proc Natl Acad Sci USA*, 2008. **105**(11):4323–4328.

Finn, R.D., Marshall, M., and Bateman, A. iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, 2005. **21**(3):410–412.

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. *et al.* Pfam: clans, web tools and services. *Nucleic Acids Res*, 2006. **34**(Database issue):D247–D251.

Fischer, M. and Bacher, A. Biosynthesis of flavocoenzymes. *Nat Prod Rep*, 2005. **22**(3):324–350.

Flannick, J., Novak, A., Srinivasan, B.S., McAdams, H.H., and Batzoglou, S. Graemlin: general and robust alignment of multiple large interaction networks. *Genome Res*, 2006. **16**(9):1169–1181.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 1995. **269**(5223):496–512.

Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M., and Wijmenga, C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 2006. **78**(6):1011–1025.

Freudenberg, J. and Propping, P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, 2002. **18 Suppl 2**:S110–S115.

Friedberg, I. Automated protein function prediction–the genomic challenge. *Brief Bioinform*, 2006. **7**(3):225–242.

Friedberg, I. and Godzik, A. Connecting the protein structure universe by using sparse recurring fragments. *Structure (Camb)*, 2005. **13**(8):1213–1224.

Futschik, M.E., Chaurasia, G., and Herzel, H. Comparison of Human Protein-Protein Interaction Maps. *Bioinformatics*, 2007. **23**(5):605–611.

Gabaldon, T. and Huynen, M.A. Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci*, 2004. **61**(7-8):930–944.

Ganem, C., Devaux, F., Torchet, C., Jacq, C., Quevillon-Cheruel, S., Labesse, G., Facca, C., and Faye, G. Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast. *EMBO J*, 2003. **22**(7):1588–1598.

Gentleman, R. Visualizing and Distances Using GO. Technical report, Bioconductor, 2007.

Gibson, G. Decanalization and the origin of complex disease. *Nat Rev Genet*, 2009. **10**(2):134–140.

Gómez-Pérez, A., Fernandez-Lopez, M., and Corcho, O. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer, 1st edition, 2004.

Goble, C. and Stevens, R. State of the nation in data integration for bioinformatics. *J Biomed Inform*, 2008. **41**(5):687–693.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K.S., Knoblich, M., Haenig, C. *et al.* A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*, 2004. **15**(6):853–65.

Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabási, A.L. The human disease network. *Proc Natl Acad Sci USA*, 2007. **104**(21):8685–8690.

Grenon, P., Smith, B., and Goldberg, L. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform*, 2004. **102**:20–38.

Guo, X., Liu, R., Shriver, C.D., Hu, H., and Liebman, M.N. Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics*, 2006. **22**(8):967–973.

Heller, B. and Herre, H. Ontological Categories in GOL. *Axiomathes*, 2004. **14**(1):57–76.

Hill, D.P., Berardini, T.Z., Howe, D.G., and Auken, K.M.V. Representing ontogeny through ontology: A developmental biologist's guide to the gene ontology. *Mol Reprod Dev*, 2009. **77**(4):314–329.

Hooper, S.D. and Bork, P. Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, 2005. **21**(24):4432–4433.

Hou, J., Sims, G.E., Zhang, C., and Kim, S.H. A global representation of the protein fold space. *Proc Natl Acad Sci USA*, 2003. **100**(5):2386–2390.

Huang, D.W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., and Lempicki, R.A. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 2007. **8**(9):R183.

Huang, T.W., Tien, A.C., Huang, W.S., Lee, Y.C., Peng, C.L., Tseng, H.H., Kao, C.Y., and Huang, C.Y. POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome. *Bioinformatics*, 2004. **20**(17):3273–3276. Le.

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F. *et al.* Ensembl 2005. *Nucleic Acids Res*, 2005. **33**(Database issue):D447–D453.

Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L. *et al.* Ensembl 2009. *Nucleic Acids Res*, 2009. **37**(Database issue):D690–D697.

Hunt, J.S. and Romanelli, F. Maraviroc, a CCR5 coreceptor antagonist that blocks entry of human immunodeficiency virus type 1. *Pharmacotherapy*, 2009. **29**(3):295–304.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res*, 2009. **37**(Database issue):D211–D215.

Ideker, T. and Sharan, R. Protein networks in disease. *Genome Res*, 2008. **18**(4):644–652.

Jensen, L.J., Gupta, R., Staerfeldt, H.H., and Brunak, S. Prediction of human protein function according to Gene Ontology categories. *Bioinformatics*, 2003. **19**(5):635–642.

Jiang, J.J. and Conrath, D.W. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of the 10th International Conference on Research on Computational Linguistics*. Tapei, Taiwan, 1997 pages 19–33.

Jimenez-Sanchez, G., Childs, B., and Valle, D. Human disease genes. *Nature*, 2001. **409**(6822):853–855.

Kann, M.G. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform*, 2007. **8**:333–346.

Kaplan, N., Sasson, O., Inbar, U., Friedlich, M., Fromer, M., Fleischer, H., Portugaly, E., Linial, N., and Linial, M. ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res*, 2005. **33**(Database issue):D216–D218.

Kemerling, G. Philosophy Pages. Web site, 2010.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R. *et al.* IntAct-open source resource for molecular interaction data. *Nucleic Acids Res*, 2007. **35**(Database issue):D561–D565.

Khatri, P. and Draghici, S. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 2005. **21**(18):3587–3595.

Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 2009. **85**(4):457–464.

Krallinger, M., Valencia, A., and Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*, 2008. **9 Suppl 2**:S8.

Kuhn, M., Campillos, M., Letunic, I., Jensen, L.J., and Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*, 2010. **6**:343.

Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N. *et al.* A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 2007. **25**(3):309–316.

Lee, D.S., Park, J., Kay, K.A., Christakis, N.A., Oltvai, Z.N., and Barabási, A.L. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci USA*, 2008. **105**(29):9880–9885.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 2004. **14**(6):1085–1094.

Lee, P.H. and Lee, D. Modularized learning of genetic interaction networks from biological annotations and mRNA expression data. *Bioinformatics*, 2005. **21**(11):2739–2747.

Lehner, B. and Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol*, 2004. **5**(9):R63.

Leibniz, G.W. *Opuscules et fragments inédits de Leibniz*. Georg Olms Verlag, Hildesheim, extraits des manuscrits de la bibliothèque royale de hanovre. paris 1903. 2. reprint edition, 1988.

Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 2006. **34**(Database issue):D257–D260.

Li, Y., Bandar, Z.A., and McLean, D. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE T Knowl Data En*, 2003. **15**(4):871–882.

Liao, D.I., Viitanen, P.V., and Jordan, D.B. Cloning, expression, purification and crystallization of dihydroxybutanone phosphate synthase from Magnaporthe grisea. *Acta Crystallogr D Biol Crystallogr*, 2000. **56**(Pt 11):1495–1497.

Lim, J., Hao, T., Shaw, C., Patel, A.J., Szabó, G., Rual, J.F., Fisk, C.J., Li, N., Smolyar, A., Hill, D.E. *et al.* A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 2006. **125**(4):801–814.

Lin, D. An information-theoretic definition of similarity. In J.W. Shavlik, editor, *Proc 15th Int'l Conf. on Machine Learning (ICML-98)*. Morgan Kaufmann Publishers (San Francisco, CA, USA), Madison, Wisconsin, USA, 1998 pages 296–304.

Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 2004. **5**:154.

Liu, H., Hu, Z.Z., and Wu, C.H. DynGO: a tool for visualizing and mining of Gene Ontology and its associations. *BMC Bioinformatics*, 2005a. **6**:201.

Liu, Y., Liu, N., and Zhao, H. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 2005b. **21**(15):3279–3285.

Liu, Y., Li, C., Zhang, P., and Xiong, Z. A Query Expansion Algorithm Based on Phrases Semantic Similarity. In *ISIP '08: Proceedings of the 2008 International Symposiums on Information Processing*. IEEE Computer Society (Washington, DC, USA), Moscow, Russia, 2008 pages 31–35.

Lord, P.W., Stevens, R.D., Brass, A., and Goble, C.A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 2003. **19**(10):1275–1283.

Lowe, H.J. and Barnett, G.O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, 1994. **271**(14):1103–1108.

Lu, L.J., Xia, Y., Paccanaro, A., Yu, H., and Gerstein, M. Assessing the limits of genomic data integration for predicting protein networks. *Genome Res*, 2005. **15**(7):945–953.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 2005. **33**(Database issue):D54–D58.

Marchini, J., Donnelly, P., and Cardon, L.R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet*, 2005. **37**(4):413–417.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, 2004. **5**(12):R101.

Mathur, S. and Dinakarpandian, D. A New Metric to Measure Gene Product Similarity. In *Proc. IEEE International Conference on Bioinformatics and Biomedicine BIBM 2007*. Sillicon Valley, CA, USA, 2007 pages 333–338.

McCray, A.T. Conceptualizing the world: lessons from history. *J Biomed Inform*, 2006. **39**(3):267–273.

McDermott, J., Bumgarner, R., and Samudrala, R. Functional annotation from predicted protein interaction networks. *Bioinformatics*, 2005. **21**(15):3217–3226.

McGinnis, S. and Madden, T.L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 2004. **32**(Web Server issue):W20–W25.

Merkl, R. and Wiezer, A. GO4genome: a prokaryotic phylogeny based on genome organization. *J Mol Evol*, 2009. **68**(5):550–562.

Miñarro-Gimenez, J.A., Madrid, M., and Fernandez-Breis, J.T. OGO: an ontological approach for integrating knowledge about orthology. *BMC Bioinformatics*, 2009. **10 Suppl 10**:S13.

Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M. *et al.* Human protein reference database - 2006 update. *Nucleic Acids Res*, 2006. **34**(Database issue):D411–D414.

Mistry, M. and Pavlidis, P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 2008. **9**:327.

Müller, H.M., Kenny, E.E., and Sternberg, P.W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, 2004. **2**(11):e309.

Morgunova, E., Meining, W., Illarionov, B., Haase, I., Jin, G., Bacher, A., Cushman, M., Fischer, M., and Ladenstein, R. Crystal structure of lumazine synthase from Mycobacterium tuberculosis as a target for rational drug design: binding mode of a new class of purinetrione inhibitors. *Biochemistry*, 2005. **44**(8):2746–2758.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L. *et al.* InterPro, progress and status in 2005. *Nucleic Acids Res*, 2005. **33**(Database issue):D201–D205.

Musen, M.A., Lewis, S., and Smith, B. Wrestling with SUMO and bio-ontologies. *Nat Biotechnol*, 2006. **24**(1):21; author reply 23.

Ng, S.K., Zhang, Z., Tan, S.H., and Lin, K. InterDom: a database of putative interacting protein domains for validating predicted protein interactions and complexes. *Nucleic Acids Res*, 2003. **31**(1):251–254.

Niedernhofer, L.J. Nucleotide excision repair deficient mouse models and neurological disease. *DNA Repair (Amst)*, 2008. **7**(7):1180–1189.

Novère, N.L. Model storage, exchange and integration. *BMC Neurosci*, 2006. **7 Suppl 1**:S11.

Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G. *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res*, 2009. **37**(Web Server issue):W170–W173.

O'Connor, T.P. and Crystal, R.G. Genetic medicines: treatment strategies for hereditary disorders. *Nat Rev Genet*, 2006. **7**(4):261–276.

Ortutay, C. and Vihinen, M. Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res*, 2009. **37**(2):622–628.

Othman, R.M., Deris, S., and Illias, R.M. A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. *J Biomed Inform*, 2008. **41**(1):65–81.

Oti, M. and Brunner, H. The modular nature of genetic diseases. *Clin Genet*, 2007. **71**(1):1–11.

Ozgür, A., Vu, T., Erkan, G., and Radev, D.R. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 2008. **24**(13):i277–i285.

Park, M., Li, Q., Shcheynikov, N., Zeng, W., and Muallem, S. NaBC1 is a ubiquitous electrogenic Na+ -coupled borate transporter essential for cellular boron homeostasis and cell growth and proliferation. *Mol Cell*, 2004. **16**(3):331–341.

Pekar, V. and Staab, S. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th international conference on Computational Linguistics*. Morgan Kaufmann Publishers (San Francisco, CA, USA), Taipei, Taiwan, 2002 pages 1–7.

Perez-Iratxeta, C., Bork, P., and Andrade, M.A. Association of genes to genetically inherited diseases using data mining. *Nat Genet*, 2002. **31**(3):316–319.

Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., Florio, A., and Cesareni, G. HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. *BMC Bioinformatics*, 2005. **6 Suppl 4**:S21.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E.N., Falcão, A.O., and Couto, F.M. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 2008. **9 Suppl 5**:S4.

Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. Semantic similarity in biomedical ontologies. *PLoS Comput Biol*, 2009. **5**(7):e1000443.

Pidcock, W. What are the differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model? Web site, 2010.

Plomin, R., Haworth, C.M.A., and Davis, O.S.P. Common disorders are quantitative traits. *Nat Rev Genet*, 2009. **10**(12):872–878.

Popescu, M., Keller, J.M., and Mitchell, J.A. Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006. **3**(3):263–274.

Pozo, A.D., Pazos, F., and Valencia, A. Defining functional distances over Gene Ontology. *BMC Bioinformatics*, 2008. **9**(1):50.

Prasad, T.S.K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. *et al.* Human Protein Reference Database-2009 update. *Nucleic Acids Res*, 2009. **37**(Database issue):D767–D772.

Pu, S., Vlasblom, J., Emili, A., Greenblatt, J., and Wodak, S.J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics*, 2007. **7**(6):944–960.

Qu, Y. and Xu, S. Supervised cluster analysis for microarray data based on multivariate Gaussian mixture. *Bioinformatics*, 2004. **20**(12):1905–1913.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. Development and application of a metric on semantic nets. *IEEE T Syst Man Cyb*, 1989. **19**(1):17–30.

Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., and Albrecht, M. Computational analysis of human protein interaction networks. *Proteomics*, 2007. **7**(15):2541–2552.

Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. *PLoS Comput Biol*, 2009. **5**(7):e1000431.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.J., Hon, G.C., Myers, C.L., Parsons, A., Friesen, H., Oughtred, R., Tong, A. *et al.* Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, 2006. **5**(4):11.

Remm, M., Storm, C.E., and Sonnhammer, E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*, 2001. **314**(5):1041–1052.

Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers (San Francisco, CA, USA), Madison, Wisconsin USA, 1995 pages 448–453.

Rhee, S.Y., Wood, V., Dolinski, K., and Draghici, S. Use and misuse of the gene ontology annotations. *Nat Rev Genet*, 2008. **9**(7):509–515.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A.M. Probabilistic model of the human protein-protein interaction network. *Nat Biotechnol*, 2005. **23**(8):951–959.

Richardson, R. and Smeaton, A. Using WordNet in a knowledge-based approach to information retrieval. In *Proceedings of the BCS-IRSG Colloquium, Crewe*. Robert Gordon University, Aberdeen, 1995 .

Riensche, R.M., Baddeley, B.L., Sanfilippo, A.P., Posse, C., and Gopalan, B. XOA: Web-Enabled Cross-Ontological Analytics. In *Proc. IEEE Congress on Services*. Salt Lake City, Utah, USA, 2007 pages 99–105.

Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. Inferring protein domain interactions from databases of interacting proteins. *Genome Biol*, 2005. **6**(10):R89.

Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 2008. **83**(5):610–615.

Rossi, S., Masotti, D., Nardini, C., Bonora, E., Romeo, G., Macii, E., Benini, L., and Volinia, S. TOM: a web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Res*, 2006. **34**(Web Server issue):W285–W292.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 2005. **437**(7062):1173–1178.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O.N., Stümpflen, V. *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 2008. **36**(Database issue):D646–D650.

Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Güldener, U., Mannhaupt, G., Münsterkötter, M. *et al.* The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, 2004. **32**(18):5539–5545.

Ruiz-Herrera, J. and San-Blas, G. Chitin Synthesis as a Target for Antifungal Drugs. *Current Drug Targets*, 2003. **3**:77–91.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*, 2004. **32**(Database issue):D449–D451.

Sanfilippo, A., Posse, C., Gopalan, B., Riensche, R., Beagley, N., Baddeley, B., Tratz, S., and Gregory, M. Combining hierarchical and associative gene ontology relations with textual evidence in estimating gene and gene product similarity. *IEEE Trans Nanobioscience*, 2007. **6**(1):51–59.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2009. **37**(Database issue):D5–D15.

Schlicker, A. *A Global Approach to Comparative Genomics: Comparison of Functional Annotation over the Taxonomic Tree*. Master's thesis, Center for Bioinformatics, Saarland University, 2005.

Schlicker, A. and Albrecht, M. FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, 2008. **36**(Database Issue):D434–D439.

Schlicker, A. and Albrecht, M. FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res*, 2010. **38**(Database Issue):D244–D248.

Schlicker, A., Domingues, F., Rahnenführer, J., and Lengauer, T. A new Measure for functional Similarity of Gene Products based on Gene Ontology. *BMC Bioinformatics*, 2006a. **7**(1):302.

Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. Functional evaluation of domain-domain interactions and human protein interaction networks. In *German Conference on Bioinformatics*. Bonner Köllen Verlag, Tübingen, Germany, Lecture Notes in Informatics, 2006b pages 115–126.

Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 2007a. **23**(7):859–865.

Schlicker, A., Rahnenführer, J., Albrecht, M., Lengauer, T., and Domingues, F.S. GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol*, 2007b. **8**(3):R33.

Schreiber, S., Rosenstiel, P., Albrecht, M., Hampe, J., and Krawczak, M. Genetics of Crohn disease, an archetypal inflammatory barrier disease. *Nat Rev Genet*, 2005. **6**(5):376–388.

Schulze-Kremer, S. Ontologies for molecular biology and bioinformatics. *In Silico Biol*, 2002. **2**(3):179–193.

Sen, T., Kloczkowski, A., and Jernigan, R. Functional clustering of yeast proteins from the protein-protein interaction network. *BMC Bioinformatics*, 2006. **7**(1):355.

Sevilla, J.L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J.M., Martinez-Cruz, L.A., Corrales, F.J., and Rubio, A. Correlation between Gene Expression and GO Semantic Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2005. **2**(4):330–338.

Shalgi, R., Lapidot, M., Shamir, R., and Pilpel, Y. A catalog of stability-associated sequence elements in 3' UTRs of yeast mRNAs. *Genome Biol*, 2005. **6**(10):R86.

Sharan, R. and Ideker, T. Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 2006. **24**(4):427–433.

Sheehan, B., Quigley, A., Gaudin, B., and Dobson, S. A relation based measure of semantic similarity for Gene Ontology annotations. *BMC Bioinformatics*, 2008. **9**:468.

Shriner, D., Baye, T.M., Padilla, M.A., Zhang, S., Vaughan, L.K., and Loraine, A.E. Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies. *Nucleic Acids Res*, 2008. **36**(4):e26.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics*, 2005. **21**:3940–3941.

Singh, R., Xu, J., and Berger, B. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc Natl Acad Sci USA*, 2008. **105**(35):12763–12768.

Smith, B. *Blackwell Guide to the Philosophy of Computing and Information*, Blackwell Publ, chapter Ontology, pages 155–166, 2003.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 2007. **25**(11):1251–1255.

Smith, B., Williams, J., and Schulze-Kremer, S. The ontology of the gene ontology. *AMIA Annu Symp Proc*, 2003. **2003**:609–613.

Smith, C.L., Goldsmith, C.A.W., and Eppig, J.T. The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol*, 2005. **6**(1):R7.

Soldatova, L.N. and King, R.D. Are the current ontologies in biology good ontologies? *Nat Biotechnol*, 2005. **23**(9):1095–1098.

Spaltmann, F., Blunck, M., and Ziegelbauer, K. Computer-aided target selection-prioritizing targets for antifungal drug discovery. *Drug Discov Today*, 1999. **4**(1):17–26.

Speer, N., Spieth, C., and Zell, A. A Memetic Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology. In *Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004)*. IEEE Press (San Diego, CA, USA), La Jolla, California, USA, 2004 .

Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D.G., Knight, J. *et al.* The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res*, 2008. **36**(Database issue):D768–D772.

Stein, A., Russell, R.B., and Aloy, P. 3did: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res*, 2005. **33**(Database issue):D413–D417.

Stein, L. Genome annotation: from sequence to biology. *Nat Rev Genet*, 2001. **2**(7):493–503.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 2005. **122**(6):957–968.

Stevens, R., Goble, C.A., and Bechhofer, S. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform*, 2000. **1**(4):398–414.

Sussna, M. Word sense disambiguation for free-text indexing using a massive semantic network. In *CIKM '93: Proceedings of the second international conference on Information and knowledge management*. ACM (New York, NY, USA), Washington, D.C., USA, 1993 pages 67–74.

Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., and Ideker, T. A direct comparison of protein interaction confidence assignment schemes. *BMC Bioinformatics*, 2006. **7**(1):360.

Tao, Y., Sam, L., Li, J., Friedman, C., and Lussier, Y.A. Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics*, 2007. **23**(13):i529–i538.

Tatusova, T.A. and Madden, T.L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*, 1999. **174**(2):247–250.

Teare, M.D. and Barrett, J.H. Genetic linkage studies. *Lancet*, 2005. **366**(9490):1036–1044.

The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*, 2001. **11**(8):1425–1433.

The UniProt Consortium. The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*, 2009. **37**(Database issue):D169–D174.

Theus, M. Interactive Data Visualization using Mondrian. *Journal of Statistical Software*, 2002. **7**(11):1–9.

Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, P., Moras, D., Westhof, E., and Poch, O. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. *Nucleic Acids Res*, 2005. **33**(13):4164–4171.

Tiffin, N., Kelso, J.F., Powell, A.R., Pan, H., Bajic, V.B., and Hide, W.A. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res*, 2005. **33**(5):1544–1552.

van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., and Leunissen, J.A.M. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 2006. **14**(5):535–542.

van Driel, M.A. and Brunner, H.G. Bioinformatics methods for identifying candidate disease genes. *Hum Genomics*, 2006. **2**(6):429–432.

Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R., and Henrick, K. E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*, 2005. **33**(Database issue):D262–D265.

von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., and Bork, P. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*, 2005. **33**(Database issue):D433–D437.

Voorhees, E.M. Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. Springer-Verlag (New York, NY, USA), Dublin, Ireland, 1994 pages 61–69.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S., and Chen, C.F. A new method to measure the semantic similarity of GO terms. *Bioinformatics*, 2007. **23**(10):1274–1281.

Watson, J.D., Laskowski, R.A., and Thornton, J.M. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol*, 2005. **15**(3):275–284.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A., and Rapp, B.A. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 2000. **28**(1):10–14.

White, T.A. and Kell, D.B. Comparative genomic assessment of novel broad-spectrum targets for antibacterial drugs. *Comp Funct Genom*, 2004. **5**(4):304–327.

Wiegers, T., Davis, A., Cohen, K., Hirschman, L., and Mattingly, C. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD). *BMC Bioinformatics*, 2009. **10**(1):326.

Wilkinson, M.D. and Links, M. BioMOBY: an open source biological web services proposal. *Brief Bioinform*, 2002. **3**(4):331–341.

Wojcik, J. and Schachter, V. Protein-protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, 2001. **17 Suppl 1**:S296–S305.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 2006a. **34**(Database issue):D187–D191.

Wu, H., Su, Z., Mao, F., Olman, V., and Xu, Y. Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res*, 2005. **33**(9):2822–2837.

Wu, X., Zhu, L., Guo, J., Zhang, D.Y., and Lin, K. Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic Acids Res*, 2006b. **34**(7):2137–2150.

Wu, X., Jiang, R., Zhang, M.Q., and Li, S. Network-based global inference of human disease genes. *Mol Syst Biol*, 2008. **4**:189.

Wu, Z. and Palmer, M. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics (Morristown, NJ, USA), Las Cruces, NM, USA, 1994 pages 133–138.

Xu, T., Du, L., and Zhou, Y. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 2008. **9**:472.

Yang, D., Li, Y., Xiao, H., Liu, Q., Zhang, M., Zhu, J., Ma, W., Yao, C., Wang, J., Wang, D. *et al.* Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, 2008. **24**(2):265–271.

Ye, P., Peyser, B.D., Pan, X., Boeke, J.D., Spencer, F.A., and Bader, J.S. Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol*, 2005. **1**:2005.0026.

Yilmaz, S., Jonveaux, P., Bicep, C., Pierron, L., Smaïl-Tabbone, M., and Devignes, M.D. Gene-disease relationship discovery based on model-driven data integration and database view definition. *Bioinformatics*, 2009. **25**(2):230–236.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, 2004. **14**(6):1107–1118.

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

Yu, H., Gao, L., Tu, K., and Guo, Z. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, 2005. **352**:75–81.

Yu, S., Tranchevent, L.C., Moor, B.D., and Moreau, Y. Gene prioritization and clustering by multi-view text mining. *BMC Bioinformatics*, 2010. **11**(1):28.

Yu, S., Vooren, S.V., Tranchevent, L.C., Moor, B.D., and Moreau, Y. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. *Bioinformatics*, 2008. **24**(16):i119–i125.

Yuan, F. and Zhou, Y. CDGMiner: A New Tool for the Identification of Disease Genes by Text Mining and Functional Similarity Analysis. In *ICIC '08: Proceedings of the 4th international conference on Intelligent Computing*. Springer-Verlag (Berlin / Heidelberg, Germany), Shanghai, China, 2008 pages 982–989.

Zhang, P., Zhang, J., Sheng, H., Russo, J.J., Osborne, B., and Buetow, K. Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, 2006. **7**:135.

Appendix A

List of OMIM Phenotypes

List of OMIM phenotypes and known disease proteins used in the different benchmark sets for validating MedSim. The first column contains the OMIM accession number and the second column disease name given by OMIM. The third column contains the UniProtKB accessions of the known disease proteins.

OMIM accession	Disease name	UniProtKB accession
#104300	ALZHEIMER DISEASE	P78380, P49810, P05067
#105200	AMYLOIDOSIS	P02647, P02671, P61626
#105830	ANGELMAN SYNDROME	P51608, Q05086, O60312
#109100	AUTOIMMUNE DISEASE	P16410, Q15116, O60602, O43918
#109800	BLADDER CANCER	P22607, P06400, P01112
#114480	BREAST CANCER	P38398, P51587, Q9BX63, Q96BI1, P42336, Q06609, O60934
#114550	HEPATOCELLULAR CARCINOMA	Q16667, P08581, Q9ULD2, O15169
#115150	CFC SYNDROME	P15056, Q02750, P01116, P36507
#115200	DILATED CARDIOMYOPATHY 1A	O75112, P02545, P12883, P68032
#125853	NONINSULIN-DEPENDENT DIABETES MELLITUS	Q9HC96, P14672, P35680, P06213, P35568, Q9UQF2, Q13562, P52945, Q16821, O15357, Q8IWU4, P22413
#130600	ELLIPTOCYTOSIS 2	P02730, P11277, P02549
#133239	ESCC	Q9Y238, P37173, Q9NZC7, P04637

#142623	HIRSCHSPRUNG DISEASE	P42892, P14138, P24530, P32004, Q99748, P07949, P39905
#144700	RENAL CELL CARCINOMA 1	Q96SL1, P49789, Q96EW2, O15527, P40337, Q8WU17
#145900	DEJERINE-SOTTAS SYNDROME	P08034, P11161, Q01453, Q9BXM0, P25189
#146110	HYPOGONADOTROPIC HYPOGONADISM	P11362, P30968, Q6X4W1, Q969F8
#149730	LACRIMOAURICULODENTODIGITAL SYNDROME	P21802, P22607, O15520
#155255	MEDULLOBLASTOMA	P25054, Q9Y6C5, Q9UMX1, P35222
#158000	MONILETHRIX	Q14533, P78385, O43790
#158810	BETHLEM MYOPATHY	P12109, P12111, P12110
#168600	PARKINSON DISEASE	P04062, O43464, Q5S007, Q99497, Q9BXM7, O60260, Q9Y6H5, P37840
#170400	HYPOKALEMIC PERIODIC PARALYSIS	Q9Y6H6, P35499, Q13698
#171300	PHEOCHROMOCYTOMA	P21912, O14521, P07949, P40337
#176807	PROSTATE CANCER	O96017, P29323, P60484, Q05823, Q9BQ52, P50539
#180300	RHEUMATOID ARTHRITIS	Q9UBC1, Q9UM07, Q8TDQ0, P11021
#187500	TETRALOGY OF FALLOT	Q8WW38, P78504, P52952
#188050	THROMBOPHILIA VENOUS THROMBOEMBOLISM	P01008, P05546, P05121, P00747, P04070
#188550	PAPILLARY CARCINOMA OF THYROID	Q16204, Q8TBA6, Q13772, P04629, Q15154, Q8IUD2, P07949, Q92734, O15164, P06753, P12270, Q9UPN9, P14373
#192600	FAMILIAL HYPERTROPHIC CARDIOMYOPATHY	P56539, P10916, P13533, P12883, P08590, Q9H1R3, Q9UM54, Q14896, O15273, P19429, P45379, P09493, P68032

#194050	WILLIAMS-BEUREN SYNDROME	Q9UIG0, Q9BQE9, Q9UDT6, P15502, O75344, Q9UHL9, P78347, Q15056, P53667, P35250, Q9Y4P3, Q9NP71, Q96I51, O43709, Q9H6D5, Q8N6F8, Q9GZY6
#202370	NEONATAL ADRENOLEUKODYSTROPHY	Q92968, O43933, Q7Z412, P50542, O60683
#203300	HERMANSKY-PUDLAK SYNDROME	O00203, Q6QNY0, Q96EV8, Q969F9, Q9NQG7, Q9UPZ3, Q86YV9, Q92902
#203450	ALEXANDER DISEASE	P14136, P49821, P02511
#209880	CONGENITAL ONDINE CURSE	P23560, P14138, Q99453, P07949, P39905
#209900	BARDET-BIEDL SYNDROME	Q9H0F7, Q8TAM1, Q6ZW61, Q8NFJ9, Q9BXC9, Q8N3I7, Q8IWZ6, Q9NPJ1, Q3SYG4, Q8TAM2, Q96RK4
#209950	FAMILIAL ATYPICAL MYCOBACTERIOSIS	P42701, P29460, P15260, P38484, P42224
#211980	ADENOCARCINOMA OF LUNG	P15056, Q9Y238, P00533, Q96BI1, P04637
#214100	ZELLWEGER SYNDROME	O60683, O75381, Q9Y5Y5, P40855, Q7Z412, P28328, P56589, P50542, Q13608, O00623
#219080	ACTH-INDEPENDENT MACRONODULAR ADRENAL HYPERPLASIA	P84996, Q5JWF2, O95467, P63092
#219100	AUTOSOMAL RECESSIVE CUTIS LAXA	O95967, P28300, Q9UBX5
#220110	MITOCHONDRIAL COMPLEX IV DEFICIENCY	Q12887, Q7KZN9, P00395, P00414, O75880, O43819, Q15526, P00403
#226650	PROGRESSIVE EPIDERMOLYSIS BULLOSA JUNCTIONALIS	Q9UMD9, Q13751, P16144
#226700	EPIDERMOLYSIS BULLOSA LETALIS	Q13751, Q13753, Q16787

#227650	FANCONI ANEMIA	P51587, Q9BXW9, O15360, Q8NB91, Q9HB96, Q9NPI8, O15287, Q9NVI1, Q9BX63, Q9NW38, Q8IYD8, Q86YC2, Q00597
#231200	GIANT PLATELET SYNDROME	P07359, P14770, P13224
#235400	ATYPICAL HEMOLYTIC UREMIC SYNDROME	P08603, P05156, P15529, Q76LX8
#236670	WALKER-WARBURG SYNDROME	O75072, Q9H9S5, Q8WZA1, Q9UKY4, Q9Y6A1
#236750	IDIOPATHIC HYDROPS FETALIS	P08236, P04062, Q04446, P10746, Q9NRA2, P69905
#242100	CONGENITAL NONBULLOUS ICHTHYOSIFORM ERYTHRODERMA	Q9BYJ1, O75342, P22735
#248600	MAPLE SYRUP URINE DISEASE	P09622, P11182, P12694, P21953
#252010	MITOCHONDRIAL COMPLEX I DEFICIENCY	Q8N183, O15239, P28331, O43181, O75251, P49821, P03897, O75306
#252150	MOLYBDENUM COFACTOR DEFICIENCY	Q9NQX3, Q9NZB8, O96007, O96033, O14940
#254200	MYASTHENIA GRAVIS	P02708, P11230, Q04844, P28329, Q07001
#254500	PRIMARY MYELOMA, MULTIPLE AMYLOIDOSIS	P24385, P01857, Q15306, P22607
#256000	LEIGH SYNDROME	P00846, Q12887, Q7KZN9, P31040, O00217, P49821, P03897, Q15526, O75251
#259700	OSTEOPETROSIS, AUTOSOMAL RECESSIVE 1	P51798, Q86WC4, Q13488
#262600	PITUITARY DWARFISM III	Q9UBX0, O75360, Q9UBR4
#265120	SURFACTANT METABOLISM DYSFUNCTION, PULMONARY, 1	P07988, P11686, P32927
#266600	INFLAMMATORY BOWEL DISEASE 1	P22301, Q5VWK5, Q9H015, O76082, Q9HC29
#267430	RENAL TUBULAR DYSGENESIS	P12821, P00797, P30556

#268000	RETINITIS PIGMENTOSA	P29973, P82279, Q5IJ48, O43186, Q12866, P08100, P16499, P35913, P23942, P47804, P12271, Q03395, Q8TA86, O75445
#268220	RHABDOMYOSARCOMA 2	P23760, P23759, Q12778
#272200	MULTIPLE SULFATASE DEFICIENCY	P15289, Q8NBK3, P15848
#275355	SQUAMOUS CELL CARCINOMA	Q9UK53, Q9NXR8, P04637, O14763, P60484
#276300	TURCOT SYNDROME	P25054, P54278, P40692
#276900	USHER SYNDROME, TYPE I	Q9H251, Q13402, Q96QU1, Q495M9, Q9Y6N9
#277580	WAARDENBURG-SHAH SYNDROME	P24530, P56693, P14138
#535000	LEBER OPTIC ATROPHY	P00846, P00395, P00414, P00156, P03886, P03901, P03905, P03915, P03923, P03891
#540000	MELAS SYNDROME	P03905, P03923, P03886
#580000	STREPTOMYCIN OTOTOXICITY	Q969Y2, Q8WVM0, O75648
#601367	ISCHEMIC STROKE	P12821, P20292, P12259, P05112, P24723, P42898, P00734, P16109
#601462	MYASTHENIC SYNDROME, CONGENITAL, SLOW-CHANNEL	P11230, Q07001, Q04844, P02708
#601495	AGAMMAGLOBULINEMIA, NON-BRUTON TYPE, AUTOSOMAL RECESSIVE	P15814, Q8IWT6, P11912
#601539	PEROXISOME BIOGENESIS DISORDERS	O60683, O00623, Q9Y5Y5, P40855, O43933, Q7Z412, P28328, P56589, O00628, Q13608
#601634	NEURAL TUBE DEFECTS, FOLATE-SENSITIVE	P11586, Q99707, P42898, Q9UBK8
#601665	OBESITY LEANNESS	O00253, P41159, P32245, P37231, Q15466, P55916, P01189
#601675	TRICHOTHIODYSTROPHY, PHOTOSENSITIVE	P19447, Q6ZYL4, P18074

#603174	HOMOCYSTEINEMIA	Q99707, Q9Y4U1, P42898, P35520
#603233	PSEUDOHYPOPARATHYROIDISM, TYPE IB	P84996, Q5JWF2, P63092, O95467, O14662
#603554	OMENN SYNDROME	Q96SD1, P15918, P55895
#603896	OVARIOLEUKODYSTROPHY	Q14232, P49770, Q9UI10, Q13144, Q9NR50
#604219	CATARACT, AUTOSOMAL DOMINANT	Q13515, P43320, P07315, P07320, P02489
#604229	PETERS ANOMALY	Q16678, Q12948, P26367, Q99697, P61812, Q9UKV0
#604233	GENERALIZED EPILEPSY WITH FEBRILE SEIZURES PLUS	P35498, Q07699, Q99250, P18507
#604271	SHORT STATURE, IDIOPATHIC, AUTOSOMAL	P10912, P01241, Q92847
#604967	PROTOCADHERIN-BETA GENE CLUSTER	Q9Y5F3, Q9Y5E7, Q9Y5E6, Q9Y5E5, Q9Y5E4, Q9Y5E3, Q9Y5E2, Q9UN66, Q9Y5E1, Q9UN67, Q9Y5F1, Q9Y5F0, Q9Y5E9, Q9Y5E8, Q9NRJ7, Q9Y5F2
#605074	RENAL CELL CARCINOMA, PAPILLARY	Q9BZE9, Q92733, P19532, P08581
#605899	GLYCINE ENCEPHALOPATHY	P23378, P48728, P23434
#606391	MATURITY-ONSET DIABETES OF THE YOUNG	P19835, P20823, P35680, P41235, P35557, Q13562, O14901
#606904	JUVENILE MYOCLONIC EPILEPSY	O00305, Q5JVL4, P14867
#607748	FAMILIAL HYPERCHOLANEMIA	P07099, Q9UDY2, Q14032
#607785	JUVENILE MYELOMONOCYTIC LEUKEMIA	Q06124, P01116, P01111, Q9UNA1, P21359
#608089	ENDOMETRIAL CANCER	P12830, P40692, P43246, P20585, P52701, P60484
#608930	MYASTHENIC SYNDROME, CONGENITAL, FAST-CHANNEL	P02708, Q04844, Q07001
#608931	MYASTHENIC SYNDROME, CONGENITAL, ASSOCIATED WITH ACETYLCHOLINE RECEPTOR DEFICIENCY	P11230, Q04844, Q13702, O15146

#608971	SEVERE COMBINED IMMUNODEFICIENCY	P04234, P16871, P08575
#609423	SUSCEPTIBILITY TO HUMAN IMMUNODEFICIENCY VIRUS TYPE 1	P51681, P49238, P41597
#609830	ABDOMINAL BODY FAT DISTRIBUTION	P37231, P18031, P01189
#610424	SUSCEPTIBILITY TO HEPATITIS B VIRUS	P01903, P16410, Q08334, P48551, P11226, P01375

Appendix B

List of Publications

Andreas Schlicker, Thomas Lengauer, and Mario Albrecht. Improving Disease Gene Prioritization using the Semantic Similarity of Gene Ontology Terms. *submitted*

Andreas Schlicker and Mario Albrecht (2010) FunSimMat update: new features for exploring functional similarity. *Nucleic Acids Res*, 38(Database Issue): D244-D248

Christoph Welsch, Francisco S Domingues, Simone Susser, Iris Antes, Christoph Hartmann, Gabriele Mayr, **Andreas Schlicker**, Christoph Sarrazin, Mario Albrecht, Stefan Zeuzem, and Thomas Lengauer (2008) Molecular basis of telaprevir resistance due to V36 and T54 mutations in the NS3-4A protease of HCV. *Genome Biol*, 9(1):R16

Andreas Schlicker and Mario Albrecht (2008) FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res*, 36(Database Issue): D434-D439

Fidel Ramírez, **Andreas Schlicker**, Yassen Assenov, Thomas Lengauer, and Mario Albrecht (2007) Computational analysis of human protein interaction networks. *Proteomics*, 7(15): 2541-2552

Michael L Tress, Pier Luigi Martelli, Adam Frankish, Gabrielle A Reeves, Jan Jaap Wesselink, Corin Yeats, Páll Ísólfur Ólason, Mario Albrecht, Hedi Hegyi, Alejandro Giorgetti, Domenico Raimondo, Julien Lagarde, Roman A Laskowski, Gonzalo López, Michael I Sadowski, James D Watson, Piero Fariselli, Ivan Rossi, Alinda Nagy, Wang Kai, Zenia Størling, Massimiliano Orsini, Yassen Assenov, Hagen Blankenburg, Carola Huthmacher, Fidel Ramírez, **Andreas Schlicker**, France Denoeud, Phil Jones, Samuel Kerrien, Sandra Orchard, Stylianos E Antonarakis, Alexandre Reymond, Ewan Birney, Søren Brunak, Rita Casadio, Roderic Guigo, Jennifer Harrow, Henning Hermjakob, David T Jones, Thomas Lengauer, Christine A Orengo, László Patthy, Janet M Thornton, Anna Tramontano, and Alfonso Valencia (2007) The implications of alternative splicing in the

ENCODE protein complement. *Proc Natl Acad Sci USA*, 104(13): 5495-5500.

Andreas Schlicker, Jörg Rahnenführer, Mario Albrecht, Thomas Lengauer, and Francisco S Domingues (2007). GOTax: investigating biological processes and biochemical activities along the taxonomic tree. *Genome Biol*, 8(3): R33.

Andreas Schlicker, Carola Huthmacher, Fidel Ramírez, Thomas Lengauer, and Mario Albrecht (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7): 859-865.

Andreas Schlicker, Carola Huthmacher, Fidel Ramírez, Thomas Lengauer, and Mario Albrecht (2006). Functional evaluation of domain-domain interactions and human protein interaction networks. *In German Conference on Bioinformatics (GCB 2006)*, Tuebingen, Germany, 115-126.

Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer (2006). A new Measure for functional Similarity of Gene Products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302.