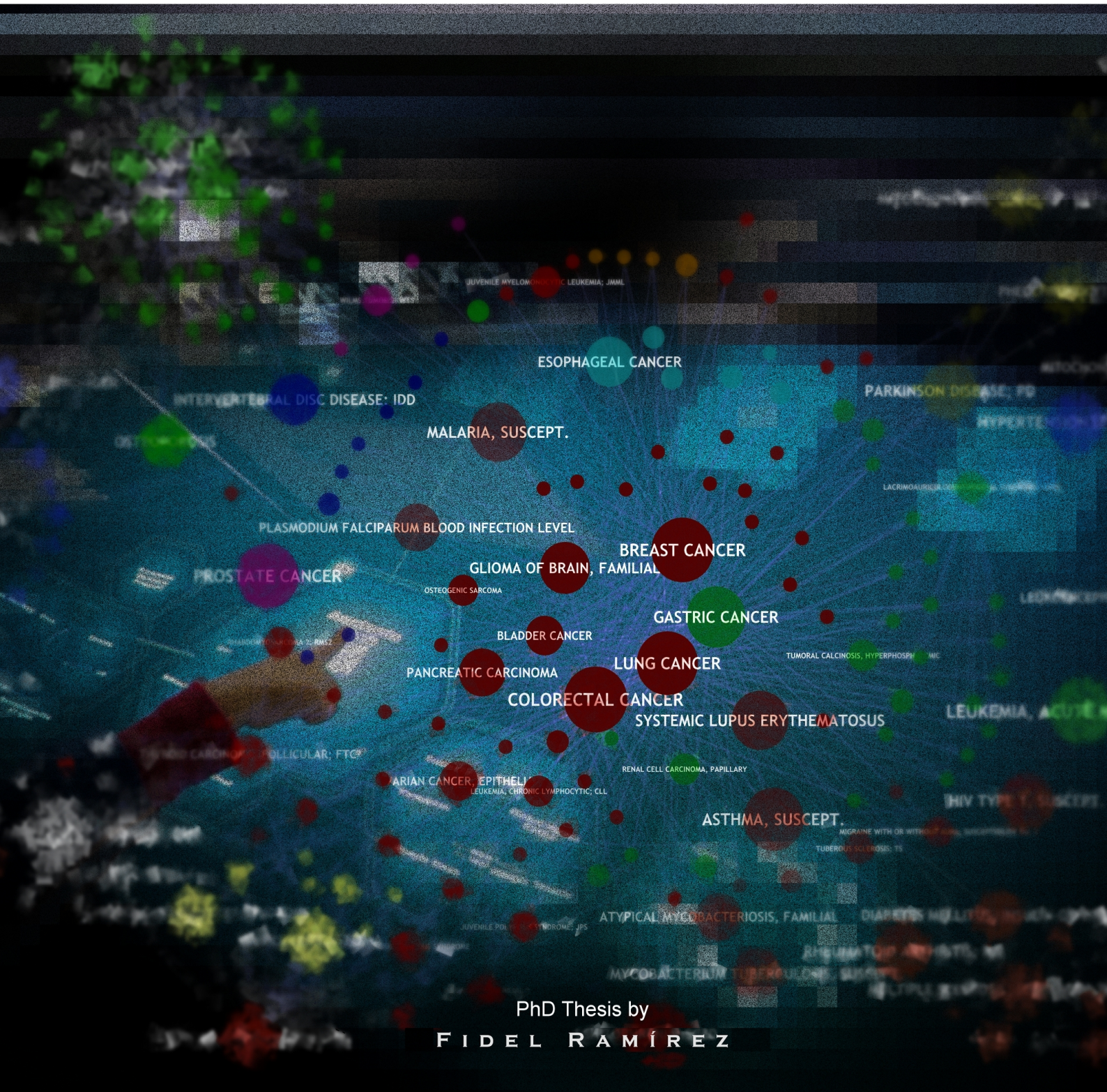
max planck institut informatik

Novel Approaches to the Integration and Analysis of Systems Biology Data



PhD Thesis by

FIDEL RAMÍREZ

Novel Approaches to the Integration and Analysis of Systems Biology Data

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften der
Naturwissenschaftlich-Technischen Fakultäten III
Chemie, Pharmazie, Bio- und Werkstoffwissenschaften
der Universität des Saarlandes

eingereicht von
Fidel Ramírez

Saarbrücken 2011

Fidel Ramírez: *Novel Approaches to the Integration and Analysis of Systems Biology Data.* Ph.D. thesis for obtaining the academic degree of a doctor of the natural sciences in bioinformatics, submitted to the Faculty III of Natural Sciences and Technology (chemistry, pharmacy, biological and materials science) of the Saarland University, Saarbrücken, Germany, 2011.

Dekan (Dean of the Faculty):	Univ.-Prof. Dr. Wilhelm F. Maier
Einreichung (Submission):	9. December 2011
Tag des Kolloquiums (Colloquium):	19. April 2012
Vorsitzerin (Chairwoman):	Prof. Dr. Alexandra Kiemer
Berichterstatter (Reviewers):	Prof. Dr. Thomas Lengauer
	Prof. Dr. Mario Albrecht
	Prof. Dr. Volkhard Helms
Akad. Mitarbeiter (Research Assistant):	Dr. Konstantin Lepikhov

I hereby swear in lieu of an oath that I have independently prepared this thesis and without using other aids than those stated. The data and concepts taken over from other sources or taken over indirectly are indicated citing the source. The thesis was not submitted so far either in Germany or in another country in the same or a similar form in a procedure for obtaining an academic title.

Saarbrücken, 9 December 2011

(Fidel Ramírez)

Contents

List of Figures

List of Tables

Abstract

The opportunity to investigate whole cellular systems using experimental and computational high-throughput methods leads to the generation of unprecedented amounts of data. Processing of these data often results in large lists of genes or proteins that need to be analyzed and interpreted in the context of all other biological information that is already available. To support such analyses, repositories aggregating and merging the biological information contained in different databases are required.

To address this need, we created an integrative data warehouse containing millions of up-to-date annotations related to human genes and proteins from over thirty major molecular biology databases. In particular, this data warehouse was instrumental in assessing the data quality of human protein interactions and in predicting an important, but largely unidentified, group of proteins that function as molecular scaffolds in the formation of signaling cascades. Additionally, the data warehouse enabled us to devise the novel computational method BioSim for the discovery of biological relationships based on the functional similarity of gene and protein annotations. Furthermore, we showed how this method allows identifying disease-associated genes.

To facilitate the analysis and interpretation of large lists of genes or proteins derived from high-throughput methods, we built the new web portal BioMyn. It provides a powerful search engine with public access to the data warehouse and the BioSim method. BioMyn also offers a number of useful tools for own functional enrichment analysis and the visualization of the results.

Kurzfassung

Die Möglichkeit, ganze zelluläre Systeme mit experimentellen und computerbasierten Hochdurchsatz-Methoden zu erforschen, führt zur Generierung beispielloser Datenmengen. Die Verarbeitung dieser Daten ergibt oft große Listen von Genen oder Proteinen, die im Kontext all der anderen bereits vorhandenen, biologischen Informationen analysiert und interpretiert werden müssen. Um solche Analysen zu unterstützen, werden Datensammlungen benötigt, die die in verschiedenen Datenbanken enthaltenen biologischen Informationen zusammenführen und verknüpfen.

Um diesem Bedarf Rechnung zu tragen, wurde ein integratives Data-Warehouse angelegt, das Millionen aktueller Annotationen bezüglich humaner Gene und Proteine aus über dreißig wichtigen Datenbanken der Molekularbiologie beinhaltet. Insbesondere war dieses Data-Warehouse nützlich bei der Bewertung der Datenqualität humaner Proteininteraktionen und bei der Vorhersage einer bedeutenden, jedoch größtenteils unidentifizierten Gruppe von Proteinen, die als molekulares Gerüst der Bildung von Signalkaskaden dienen. Zudem ermöglichte es das Data-Warehouse, die neuartige Computermethode BioSim zur Aufdeckung biologischer Ähnlichkeiten, basierend auf funktionellen Ähnlichkeiten von Gen- und Proteinannotationen, zu entwickeln. Des Weiteren wurde gezeigt, wie diese Methode die Identifizierung krankheitsassoziierter Gene erlaubt.

Um die Analyse und Interpretation großer Listen von Genen oder Proteinen, die aus Hochdurchsatz-Methoden stammen, zu erleichtern, wurde das neue Webportal BioMyn geschaffen. Es bietet eine starke Suchmaschine, die das Data-Warehouse und die BioSim-Methode öffentlich zugänglich macht. Auch stellt BioMyn eine Reihe praktischer Tools für eigene Funktionsanalysen und die Visualisierung der Ergebnisse zur Verfügung.

Acknowledgements

Of the many people who have been enormously helpful in the preparation of this thesis, I am especially thankful to Mario Albrecht who offered me the opportunity to join the Max Planck Institute for Informatics and whose further guidance, support and never-ending corrections challenged me through my Ph.D. studies. I would also like to express my gratitude to Thomas Lengauer who gathered an incredible team of researchers under a wonderful academic atmosphere that I was lucky to join. I am also grateful with him for his constant support and critical revision of this dissertation. Also, I thank Jörn Walter for his trust and dedicated efforts that finally made possible my registration as Ph.D. student in the Saarland University.

I specially want to thank Hagen Blankenburg for his insightful suggestions and comments on the manuscript and for his constant technical requests about the BioMyn web portal that led to large improvements. I would also like to acknowledge my office mates, Dorothea Emig and Andreas Schlicker for their continuous help and support. Also, I thank Andreas for introducing me into the topic of functional similarity.

Furthermore, I thank Ingolf Sommer, Francisco Domingues, Jörg Rahnenfürer, Jochen Maydt, Christoph Bock, Oliver Sander, Tobias Sing and Elena Zotenko for kindly sharing their knowledge and expertise. I greatly acknowledge Joachim Büch and Georg Friedrich for their technical support and Ruth Schneppen-Christmann for her support with all administrative issues.

I would like to thank Yassen Assenov, who developed the software NetworkAnalyzer that I used to investigate the topological properties of predicted protein-protein interaction networks; Mike Wininger who developed the plug-in to interconnect the software platform Cytoscape with our data warehouse; Adrian Alexa who provided me optimized functions for R to dramatically speed up the computation of functional similarities; and Glenn Lawyer who advised and guided my research on functional similarity methods. Also, I particularly thank Adrian

for his constant motivation to bike and Glenn for his friendship, wonderful book suggestions and lively discussions.

I am thankful to many friends and colleagues from the Max Planck Institute for Informatics that directly or indirectly helped me towards the completion of this dissertation: Laura Tolosi, Jasmina Bogojeska, Gabriele Mayr, Sven-Eric Schelhorn, Sarah Diehl, Alexander Thielen, André Altmann, Hongbo Zhu and Friederike Mühlpfordt. Specially, I would like to extend my heartfelt gratitude to Andreas Steffen, his wife Kerstin and their little son Frederik for their selfless and trustworthy friendship.

I am also indebted to Miguel Granados, Lina Ruiz, Jose David Gomez and Sergio Roa for their kind friendship and profound discussions that always challenged me to see new perspectives.

Naturally, none of this work would have been possible without the help from my family here in Germany and my parents in Colombia. I am specially grateful with my wife Diana and my son Sergio that always have stood by me and who give meaning to my life and work.

This PhD thesis was supported by basic funding of the Max Planck Society; the German National Research Network (NGFN); and German Research Foundation (DFG) contract number KFO 129/1-2. The research was conducted in the context of the Cluster of Excellence for Multimodal Computing and Interaction.

1

Introduction

1.1 Motivation

Thanks to numerous technological breakthroughs, large scientific endeavors, and computational power, the turn of the century has seen an exponential growth of available biological data. Not only an increasing number of genomes is being sequenced, but also large amounts of data that characterize genes, proteins and their interrelations is nowadays available in specialized biological and medical databases. This deluge of information has enabled a transition of the biological sciences from a reductionistic approach where the focus lies on individual cellular parts to a more holistic, modular approach where biological complexity is studied through the structure, interaction and dynamics of the different biological systems (Kitano, 2002; Hartwell et al., 1999).

This new approach, called *systems biology*, aims to understand emergent cellular properties such as division, growth and other complex processes resulting from the myriad of interaction between cellular constituents (Sauer et al., 2007). Those emergent properties are not properties of the individual constitutive elements, but instead arise as a result of the functioning of the system as a whole. For this reason, *systems biology* studies are often based on high-throughput methods to comprehensively interrogate different biological states. Current high-throughput methods include expression arrays, RNA interference screens, genome-wide association studies, yeast two-hybrid screens, proteomic methods based on mass spectrometry and, recently, next-generation sequencing. These experimental methods usually explore different aspects of the cell function and produce complementary results that need to be procesed and combined through bioinformatic methods (Ge et al., 2003).

The sheer amount of data produced by systemic approaches has brought forward new informatic challenges concerning the storage, processing, statistical analysis and biological interpretation of the data. In particular, the analysis and biological interpretation of the experimental data requires the use of information

found in the literature or stored in specialized databases in order to make useful inferences. This important analysis in the context of all other information available is often cited as a critical step to achieve a systems level understanding of biological processes (Reiss et al., 2011; Sauer et al., 2007). However, such information is not readily available but scattered in multiple databases under different formats. Moreover, the biological information found in databases comprise diverse aspects of gene and protein function, regulation, location, phenotype, interaction, etc. that are difficult to merge and prepare for analysis specially because each database tends to use a different reference system to identify genes and proteins. For these reasons, the integration of this information in order to be used for the analysis and interpretation of high-throughput data, has been acknowledged as one of the greatest challenges in bioinformatics (Goble & Stevens, 2008).

Although many bioinformatic tools had been developed to match the analytical demands of large-scale studies, there are only few comprehensive and centralized repositories integrating existing biological information with experimental and computationally generated data. Yet, the demand for such tools is growing as more and more laboratories become involved in large-scale studies due to increased availability and low costs of high-throughput methods. The unprecedented speeds in which this changes are occurring has met many institutions and laboratories unprepared to face the challenges associated with the analysis of large quantities of data. Also, most biologist and biochemists usually do not have the time and knowledge to integrate available biological information to properly analyze and interpret their results. Instead, they rely on the existence of ready-to-use tools to carry out their analysis. Unfortunately, major established biological databases are just starting to adapt to the rapid changes and are still unprepared for analyzing large datasets. For instance, the popular web portals of EBI (McWilliam et al., 2009), NCBI Entrez (Sayers et al., 2011) and UniProtKB (The UniProt Consortium, 2011), and the genome browsers from Ensembl (Flicek et al., 2011) and UCSC (Fujita et al., 2011) do not allow the submission of large list of genes or proteins for analysis. Instead the information contained in such databases has to be downloaded and processed for local analysis.

Considering the great importance to offer data mining and analysis tools that are easily accessible, this work aims to: create the appropriate infrastructure to unify and integrate biological information for genes and proteins from diverse specialized databases; develop methods for the analysis of large sets of genes or proteins; and make this information and methods freely available through simple and easy-to-use web interfaces to promote its use.

1.2 Overview

In this work, a data warehouse model was chosen to create a central repository for storing the accumulated information of genes and proteins. This information is usually referred as gene or protein *annotations* and describes different aspects of gene and protein function, location, structure, relations and expression.

The data warehouse supports efficient searches over the integrated annotations in which the annotations can be easily combined and filtered. The use of this data warehouse enabled system-wide bioinformatic investigations to study the quality of protein-protein interactions (Ramírez et al., 2007) and the prediction human scaffold proteins (Ramírez & Albrecht, 2010).

Apart from facilitating large-scale bioinformatic analyses, the integrated data warehouse allowed the development of a comprehensive search method that identifies closely annotated genes and proteins that are likely to be functionally related (**?**). Access to the centralized information also facilitated the application of enrichment analysis based on the combined annotations. Similar available tools for enrichment analysis of high-throughput data are currently based on only one or two of the specialized databases. Instead, the tools developed in this work allow the quick analysis of data using all major biological databases and offer methods for searching, combining, filtering and mining the data. Moreover, the development of a web interface to easily interact with the integrated annotations simplifies the analysis of large sets of genes or proteins.

1.3 Outline

The remainder of this dissertation is composed of seven chapters followed by summarizing conclusions and an appendix. Chapter 2 describes the guidelines for the construction of the data warehouse and the different types of biological knowledge integrated.

The following two chapters describe bioinformatic studies based on integrated data. In Chapter 3, predicted and experimental protein-protein interactions were assessed by using functional annotations and domain-domain interactions. Our results confirmed the low quality of large-scale yeast-two hybrid screens and highlighted the good quality of some prediction methods.

Chapter 4 describes the prediction of scaffold proteins, important members of signaling cascades. This research used the interaction-based definition of scaffold proteins from Zeke et al. (2009) to obtain a set of candidates that was subsequently validated and characterized. This set of candidates was obtained by cross

linking protein functions, biological processes and protein-protein interactions from the different biological databases stored in the data warehouse.

Chapters 5 and 6 of the thesis present a number of tools that were developed to facilitate the mining of data within the data warehouse. Chapter 5 introduces a method for quantifying the pairwise similarity of genes and proteins based on their integrated annotations. The data warehouse can be quickly searched by using this novel method to identify other genes or proteins matching the annotations of a query gene or protein. The advantages of this method over similar methods are thoroughly demonstrated using known functional relations. Moreover, new genes associated with a disease were correctly inferred by searching the data warehouse for genes, similarly annotated, to those already associated with the disease.

Chapter 6 presents the new web portal, *BioMyn*, that allows public access to the data warehouse and to the functional similarity method. The aim of the web portal is to open the use of the data warehouse to study large gene and protein sets. For this, *BioMyn* offers enrichment analysis over all integrated data and novel visualization methods to explore the results. The development of this portal was based on the experiences presented in previous chapters and established workflows from other studies.

Chapter 7 summarizes and evaluates the main achievements of this work and draws conclusions on the completed research. It also discusses methodological improvements and future directions.

The Appendix lists all validated scaffold proteins described in Chapter 4 and the disease-associated genes correctly discovered by the similarity method introduced in Chapter 5. The appendix also contains a list own of publications related to this thesis.

2

Integration of Molecular Biology Databases

This chapter introduces the development of a data warehouse to integrate, into a single database, the information about human genes and proteins that is usually found scattered in multiple molecular biology databases. First, this chapter discusses the challenges behind biological data integration and the different approaches that had been proposed. Then, the integrative approach adopted for this dissertation, namely the data warehousing model, is presented along with the design guidelines that have led its development. Finally, the description of the relational database used for the data warehouse and the methods for acquiring and storing the data are presented.

2.1 Introduction

The number and variety of biological databases have increased steadily during the past years (Fig. 2.1) reaching over 1,000 molecular biology databases in the year 2011 according to the compilation prepared by the *Nucleic Acids Research* journal (Galperin & Cochrane, 2011). Researchers can make inferences about the functioning of complex biological systems by contrasting experimental results with the complementary information deposited in these databases. (Goble & Stevens, 2008).

Despite the large amounts of available data, the spread of the information across multiple locations hampers cross-database searches useful for data mining and analysis. Working with available biological information is time-consuming as often databases have to be visited and queried one by one using a web browser to follow the links. Manual inspection of the data is not only impractical but hardly scalable. Instead, methods to do batch searches over the multiple databases to combine their information are more useful. This is particularly needed for the interpretation of the results from high-throughput experiments that require the analysis of hundreds to thousands of genes or proteins in the context of all available knowledge.

Numerous approaches are currently available to facilitate searches over numerous biological databases. Roughly they fall into four categories (Stein, 2003; Goble & Stevens, 2008): approaches that facilitate the navigation of the information by cross-linking the information using hypertext links, approaches that query multiple databases on-the-fly and present the collected results to the user, approaches based on workflows where databases are queried one after the other and in which the results from one step feed the next, and finally, approaches in which large amounts of information is downloaded from the biological databases in advance and stored locally for efficient searches. Of these approaches, the most useful for complex and large searches is the last approach called data warehousing. The first approach does not allow batch searches while the other two require a long time to fetch the results from all sources.

This chapter presents a solution, using the data warehouse approach, to integrate information for human genes and proteins found in numerous molecular biology databases. Locally, the information from the source databases is stored using a relational database that allows the efficient querying of large amounts of information. The development of such a data warehouse was motivated by the need to collect, maintain, and analyze the available information on human genes and proteins to support our research projects.

Data warehouse systems for integrating a large number of biological databases include Atlas (Shah et al., 2005), BioWarehouse (Lee et al., 2006), and BN++ (Küntzer et al., 2006). These data warehouses contain a large number of tables and relations to store the heterogeneous information available. For instance, BioWarehouse has over fifty tables and BN++ over 200 tables. A large fraction of these tables is dedicated to store the details of different types of biological information such as the sequence of genes and proteins; signaling and metabolic pathways, including reactions, reactants, and chemical compounds; protein-protein interactions, and many other features of genes and gene products. Our approach is different from these databases; instead of having tables for each type of information we use a much simpler database schema in which the heterogeneous information available is converted to a common model requiring only over a dozen tables. This simplified schema facilitates the maintenance and development of the data warehouse.

Furthermore, in current data warehouses the methods to merge the multiple identifier systems for referencing genes and proteins from the source databases are not clearly exposed. This is a critical step required for properly integrating biological information; in our approach, it is clearly documented.

There are two other data warehouses that are specific for signaling and metabolic pathways: ConsensusPathDB (Kamburov et al., 2009) and PathwayCommons (Cerami et al., 2011). These two databases, however, do not allow the

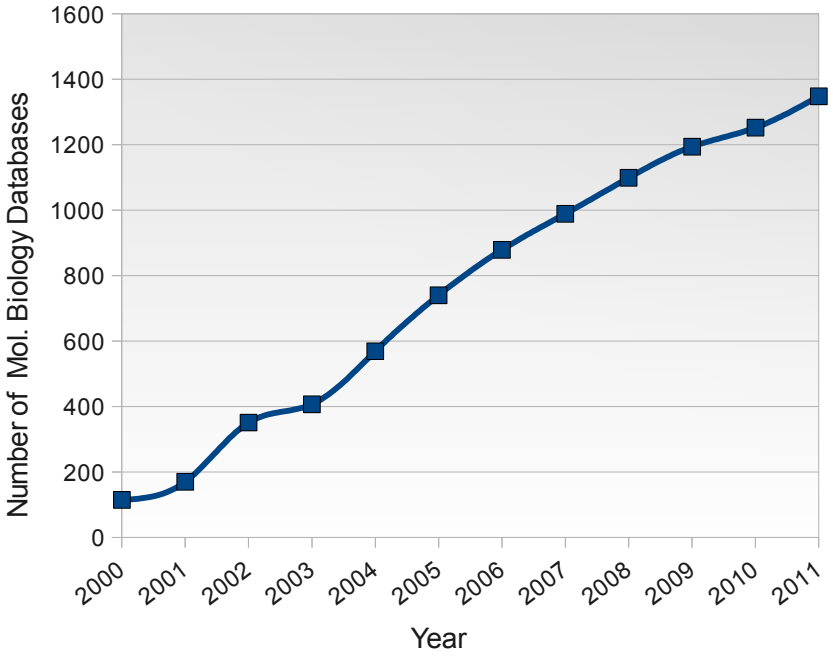


Fig. 2.1: Growth in the number of molecular biology databases. This figure describes the cumulative growth in the number of molecular biological databases that are part of the *Nucleic Acids Research* online Database Collection. The figure was generated using data compiled from the yearly summary accompanying the Database Issue of the *Nucleic Acids Research* journal from 2000–2011.

integration of other important biological databases like the Gene Ontology (GO) (Barrell et al., 2009) and OMIM (Amberger et al., 2009).

2.2 Challenges of data integration

There are numerous difficulties when integrating molecular biology information besides the spread of the information in different locations. Some of these problems are technical and deal with the diverse computational methods needed to obtain the information from the databases and the different file formats in which the information is stored. However, most difficulties stem from the complexity of the information that reflects the underlying complexities of the biological objects described. The following is a list of the most prominent problems encountered.

Heterogeneous information

One of the challenges hindering data integration is the diversity of information that needs to be gathered, ranging from annotations containing the description

of the molecular function of a protein to the complex biochemical steps involved in a metabolic pathway.

Most molecular biology databases collect and catalogue published experimental results from the scientific literature; however the scope and coverage of these databases vary extensively. Some databases contain several types of information for one particular organism, for instance, the Saccharomyces Genome Database (SGD) (Engel et al., 2010) and the Mouse Genome Database (MGD) (Blake et al., 2010); while other databases specialize in cataloguing specific gene and protein information across all organisms, as is the case of the ENZYME database (Bairoch, 2000) which contains descriptions of all characterized enzymes and the Protein Data Bank (PDB) (Rose et al., 2011) that stores all known 3D structures of proteins. Other databases maintain smaller collections of genes or proteins pertaining to specific cellular processes such as the Autophagy Database (Homma et al., 2011). Some databases serve as storage for data coming from high-throughput methods, for instance GEO (Barrett et al., 2010) and ArrayExpress (Parkinson et al., 2010) contain the raw data from microarray experiments. Several databases contain computationally derived knowledge such as protein domains, linear motifs, predicted protein-protein interactions and microRNA target predictions.

Several of the existing biological databases, mostly supported by large organizations and containing a wide range of information, have become an important part of biological research. Among these popular databases are the genome browsers Ensembl (Flicek et al., 2011) and UCSC (Fujita et al., 2011); the National Center for Biotechnology Information (NCBI) (Sayers et al., 2011), the European Bioinformatics Institute (EBI) (McWilliam et al., 2009) and the Universal Protein Resource (UniProtKB) (The UniProt Consortium, 2011).

This diversity of databases containing different types of data complicates the design of storage frameworks (e.g. relational database schemata), to store and manipulate all data types. Furthermore, if a novel type of data becomes available, either from new experimental or computational results, such a framework has to be upgraded to accommodate the data type.

Multiple gene and protein identifiers

One particular problem of data integration is caused by the multiple means for identifying or referring to genes and proteins. Biological data tends to have numerous synonyms that are difficult to recognize without expert knowledge and that are often ambiguous. Some research communities use a terminology in which the same words are used in different contexts with diverging meanings. An example is the word ‘pseudogene’ which can be a (i) gene-like structure containing in-frame stop codons or evidence of reverse transcription, (ii) a sequence with a

full reading frame that is not transcribed or (iii) a transposable cassette that is rearranged in the course of antigenic variation (Goble & Stevens, 2008). Moreover, large databases cataloguing genes and protein sequences had progressed more or less independently from one another, resulting in a proliferation of reference systems to identify genes, transcripts and proteins.

The following list of commonly used identifier systems illustrates the different manner in which gene and protein identifiers are being used.

GeneInfo Identifier (GI)

The GeneInfo Identifier (Benson et al., 2006) was introduced for the release of GeneBank 81.0 (Febrary, 1994) to identify every sequence (DNA, RNA and protein translation) in the database. A GI number was assigned to each nucleotide and protein sequence accessible through the NCBI search systems. When a sequence changes (even a single nucleotide or amino acid), a new GI number is assigned to the sequence. Changes not related to the sequence itself, such as a change in some annotation of the sequence (e.g. PubMed id), do not alter the GI.

Since the GI identifier is a number that is incrementally assigned to each new sequence submitted to the GeneBank, it is not possible to differentiate DNA, RNA or amino acid sequences just by analyzing the GI number.

DDBJ/EMBL/GenBank accession

This reference system arose when the GenBank, EMBL and the DDBJ databases started to collaborate and share information (Cochrane et al., 2011). The accession number assignment process is managed by prefix agreements from the collaborating databases. Nucleotide sequence version identifiers contain two letters followed by six digits. For older nucleotide records the format is one letter followed by five digits. Protein sequence version identifiers contain three letters followed by five digits. In contrast to GI, updates to sequences in the DDBJ/EMBL/GenBank accession are differentiated by adding a dot and a version number after the accession number.

RefSeq

RefSeq is a curated non-redundant database for sequences that includes features and bibliographic annotation (Pruitt et al., 2005). The RefSeq database is built and distributed by the NCBI and the identifiers used in this database are not related to a particular submission to GeneBank. RefSeq identifiers are stable over time and have a version system similar to DDBJ/EMBL/GenBank accession numbers; when a sequence is changed, only the version portion of the identifier is modified.

The RefSeq database aims to be a non-redundant collection of sequences, summarizing and representing the "current" view of the sequence information,

names and other annotations (Pruitt et al., 2005). The RefSeq identifier system distinguish between DNA, RNA and protein sequences by a characteristic prefix in the accession number followed by underscore. NM stands for RNA, NC for DNA and NP for protein. There are some other prefixes for untranslated RNA and for predicted assemblies. Identical gene sequences at different location within the genome have the same RefSeq identifier.

Entrez Gene ID

Entrez Gene id is part of the Entrez retrieval system of NCBI (Maglott et al., 2007). The primary goals of Entrez Gene are to provide unique identifiers for genes from a subset of model organisms and to report information associated with those identifiers. The main difference with other identifiers is that Entrez identifiers refer to any identified gene even if the sequence is not known. When the gene sequence is known, the references to sequence identifiers are reported in the gene description page part of the NCBI website. Duplicate sequences in a genome have different Entrez ids.

UniProtKB

The Universal Protein Resource Knowledge Base (UniProtKB) maintains two types of protein sequences, the manually curated sequences in UniProtKB/SwissProt and the unreviewed and automatically annotated sequences of UniProtKB/TrEMBL (The UniProt Consortium, 2010). Each Swiss-Prot and TrEMBL entry has two identifiers, the *entry name* which is a mnemonic string containing the protein and species name or, in the case of TrEMBL, an alphanumeric string instead of the protein name. The second identifier is the *accession number* which is a stable identifier of six alphanumeric characters. Swiss-Prot and TrEMBL can not be distinguished by inspecting the *accession numbers*. Alternative products generated from the same gene caused by alternative splicing, alternative promoters or alternative initiation are differentiated by adding a dash and a number to the *accession number*. Identical protein sequences produced by duplicated genes have the same identifiers. In contrast to Entrez gene, UniProtKB does not limit protein sequences to any set of organisms.

UniParc

UniParc, which stands for UniProt Archive (Leinonen et al., 2004), is a non-redundant archive of protein sequences extracted from different publicly accessible databases. Each unique sequence identifier has the prefix `UPI` followed by ten hexadecimal numbers. In UniParc, identical sequences from different organisms receive the same identifier, but sequence variants of the same gene are given different identifiers. UniParc does not maintain any annotations, only contains protein sequences and cross-references to the source databases.

Ensembl

Ensembl identifiers (Flicek et al., 2011) are assigned automatically by the Ensembl pipeline to genes, transcripts and gene products. Ensembl contains genome information for a number of organisms that have been fully sequenced. For the human genome the Ensembl identifiers have the prefixes ENSG, ENST and ENSP for gene, transcript and protein, respectively, followed by ten digits. In Ensembl, different identifiers are assigned to identical protein sequences if they are produced by distinct genes.

HGNC

The Human Genome Organization (HUGO), through its Gene Nomenclature Committee (HGNC) is in charge of assigning unique gene symbols and names to human genes (Seal et al., 2011). HGNC symbols only contain upper-case letters and Arabic numerals and must be at most six characters long. Duplicated genes have different identifiers.

This information is summarized in Table 2.1 which also includes example identifiers for the BRCA1 gene. Currently, a human gene has numerous name synonyms and at least five identifiers: the approved symbol given the the HUGO organization, the Gene Identifier assigned to its GenBank sequence, a RefSeq identifier, an Ensembl identifier and an Entrez identifier. The gene product have a similar proliferation of identifiers, including the UniProtKB accession numbers and Ensembl identifiers for transcripts and peptides.

In order to effectively integrate molecular biology databases, the unification of these diverse gene and protein identifiers is critical. Otherwise, it would be difficult and time consuming to recognize and combine the information associated with different instances of the same gene or protein. Yet mapping of all different identifiers that refer to the same gene or gene product is a difficult task, normally requiring the alignment of a growing amount of sequences or the use of mapping tables. Even when using mapping tables, the mapping of identifiers is a complicated process as demonstrated by Razick et al. (2008) who listed a number of problems found when mapping identifiers from protein-protein interaction databases. Some of the common problems they found are: the identifier was retired from the source database, the identifier corresponds to a gene or sequence that was updated and assigned to a new identifier, the identifier maps to several other identifiers, the given taxonomy or source database for the identifier are ambiguous, and the identifier contains typographical errors.

For these reasons, Alibés et al. (2007) and Côté et al. (2007) consider that one of the most challenging task for data integration is precisely the unification of gene and protein identifiers.

identifier	cellular entity	manually reviewed	redundant	examples of identifiers		
				gene	transcript	protein
GeneInfo Identifier (GI)	DNA, RNA and protein sequences	no	yes	30039658	1498736	30039659
DDBJ/EMBL/GenBank accession	DNA, RNA and protein sequences	no	yes	AY273801.1 DQ190453.1	U64805.1 U14680.1	AAC00049.1 AAN61423.1
RefSeq	DNA, RNA and protein sequences	yes	no	NG_005905.2	NM_007294.3 NM_007297.3	NP_009225.1 NP_009228.2
Entrez Gene	genes (even when sequence is not known)	yes	yes	3039	—	—
UniProtKB/Swiss-Prot	protein sequences	yes	no	—	—	P38398 BRCA1_HUMAN
UniProtKB/Swiss-Prot (alternative variants)	protein sequences	yes	no	—	—	P38398-1 P38398-2
UniProtKB/TrEMBL	protein sequences	no	yes	—	—	Q3LRJ6 Q3LRJ6_HUMAN
UniParc	protein sequences	no	no	—	—	UPI0000126AC8 UPI000013ECD3
Ensembl	DNA, RNA and protein sequences	no	yes	ENSG00000012048	ENST00000357654 ENST00000393691	ENSP00000350283 ENSP00000377294
HGNC	human genes	yes	yes	BRCA1	—	—

Table 2.1: Gene, transcript and protein identifier systems commonly found in molecular biology databases. Biological databases have developed different systems to catalog and identify nucleotide and amino acid sequences. In this table, the most commonly used identifier systems are listed. The 'manually reviewed' column indicates whether the database providing the identifiers manually revises and annotates each entry. The 'redundant' column informs if the database providing the identifier contains redundant information for sequences. Non-redundant databases aggregate sequences that correspond to the same gene, transcript or protein. In the case of Entrez Gene identifier and HGNC, the databases assign different identifiers to identical genes that are found at different location in the genome. All the example identifiers refer to the breast cancer 1, early onset gene (BRCA1) for which four alternative products are known, although in the table only two such alternative products are shown.

Information available for genes, proteins and proteins variants

While biological information is often annotated to genes, some databases contain specific knowledge at the level of proteins, protein variants or even only for proteins that are expressed in a particular cellular location or have specific post-transcriptional modifications. If the integrated information is supposed to be useful for analysis, the unification around a single type of cellular entity (gene, protein, protein variant) is often required. However, the unification of the information between different cellular entities requires transferring information from one type of entity to the other, potentially creating wrong associations or erasing valuable information. As an example, consider a disease-associated gene that has

two splice variants, each of which is expressed in a different tissue. By annotating the disease association to both gene products, new disease associations are made, but there is no certainty whether both or only one of the protein variants is responsible for the disease. When annotating the different tissue expression from the two gene products to the gene, the information on the specificity of the annotation to each protein variant is lost.

Moreover, when mapping from gene to protein identifiers a one-to-many relation is generally observed as each gene is translated into one or more products. However, it is not rare to find a many-to-one mapping for duplicated genes that produce identical proteins. Also a number of genes are not translated into proteins, like microRNA genes, and their associated information can not be transferred to proteins.

Different file formats

A technical issue aggravating the integration of data are the different file formats in which the data is found. Most often data is found in plain text files containing two-dimensional arrays based on a tab-delimited or comma-delimited structure, in which the contents (the internal table columns) are specific for each database.

Some of the biological information is recently becoming available using standards developed by consortia willing to facilitate the sharing of data. Some of this standards are based on the Extensible Markup Language (XML), as for instance, the BioPAX language to represent biological pathways (Demir et al., 2010) and the Proteomics Standards Initiative for Molecular Interactions (PSI-MI) language for molecular interaction data (Isserlin et al., 2011). Other standards are available in different file formats, for example the Minimum Information About a Microarray Experiment (MIAME) (Brazma et al., 2001) and the Open Biomedical Ontologies (OBO) can be stored both in XML or plain text formats.

Although these standards are meant to facilitate the data sharing, the complexity of some formats has limited their acceptance in laboratories lacking dedicated bioinformatics support. However, the biggest problem that entail the use of standards for data integration are the quirks found in the implementation of the standards by the biological databases offering such standards to share their data. Frequently, the files available differ in the interpretation of the standard or contain errors that need to be identified on a case-by-case basis.

Diverse coverage and quality of the data

The quality of the information contained in the databases is expected to vary due to the different rates of false positives and false negatives of experimental and

computational methods. The coverage of information is also not uniform. While system-wide studies generate comprehensive results, most manually curated data is usually focused in a limited number of medically relevant genes or proteins. For most databases, however, the reliability of the data is unknown. Generally, it is accepted that manually curated databases are more reliable than collections containing high-throughput data. Nevertheless, a recent publication that analyzed the reliability of manually curated databases has challenged this assumption, claiming that some manually curated databases may contain up to 40% erroneous annotations (Schnoes et al., 2009).

Few standard methods to access the data

Although most of the biological databases are public, the programmatic access to the collected information is hampered by each site providing their own methods for searching and retrieving the information. Moreover, there is no efficient way to detect which information has been updated or newly added. For most cases, there is a download folder where the required information is contained in which the time stamp of the files give an indication of the last update. Other sites have to be manually accessed in order to obtain the information because forms need to be filled to download the data. Finally, in some databases the information is provided as single large files whereas others deliver multiple small files.

2.3 Data integration approaches

Numerous methods to support the data integration of biological databases have been developed and explored during the past years. Roughly, the available methods can been categorized as *link integration*, *view integration*, *workflow integration*, and *data warehousing* (Stein, 2003; Goble & Stevens, 2008).

Link integration

This is the usual type of integration found in many biological databases. Also, it is the most natural as it relies on the hypertext, one of the basic features of the World Wide Web. *Link integration* directly cross references entry data from one site with another entry in a different site. Known tools that use this method are SRS (Etzold et al., 1996), LinkDB (Fujibuchi et al., 1998), GeneCards (Safran et al., 2002), and the Bioinformatic Harverster (Liebel et al., 2004).

Among the usually cited drawbacks of link integration is the difficulty to maintain unbroken links caused by name ambiguities and updates. Moreover, when

following cross-references, the integration is actually being undertaken manually by the researcher.

This integration method also lacks scalability. If more than a handful of entities are to be inspected, the process of gathering information through cross-references from relevant databases becomes difficult and time consuming.

View integration

This type of integration collects information on-the-fly from different databases and aggregates it for the user. This integration system uses a strategy where the data are acquired using remote access to web services provided by the original databases. The *view integration* clients responsible for querying the database servers use defined web service protocols, for instance RPC, REST, WSDL, and DAS (Dowell et al., 2001). *View integration* has the advantage of being always up-to-date. An example of this type of integration is DASMI (Blankenburg et al., 2009a), a web-site that collects information from different protein-protein interaction databases.

A drawback of this approach is the time required to gather the data from all sources. For this reason, analyses involving thousands of entities are impractical because of the amount of data needed to be transferred from each of the sources. Moreover, view integration relies on the existence of services, when such services are missing the data cannot be integrated.

Workflow integration

Workflows describe, execute and monitor a number of processes that can be chained into a pipeline, where the results from one step are required to start the next. In this approach the users can design their own workflows or use pre-existing ones. Examples of workflow software are Galaxy (Goecks et al., 2010) and Taverna (Hull et al., 2006).

The software implementing the workflow is responsible for formatting the data passed or obtained between the pipeline steps. In this approach the integration model is exposed, allowing the users to create their own specific integration. A disadvantage of workflows is that they are difficult to design and require good knowledge of the services available and their formats. Moreover, the software for workflow integration is not optimal to handle large quantities of data compared to specialized software such as relational databases. Also, *workflows* rely on the existence of web services, when such services are missing the information can not be integrated.

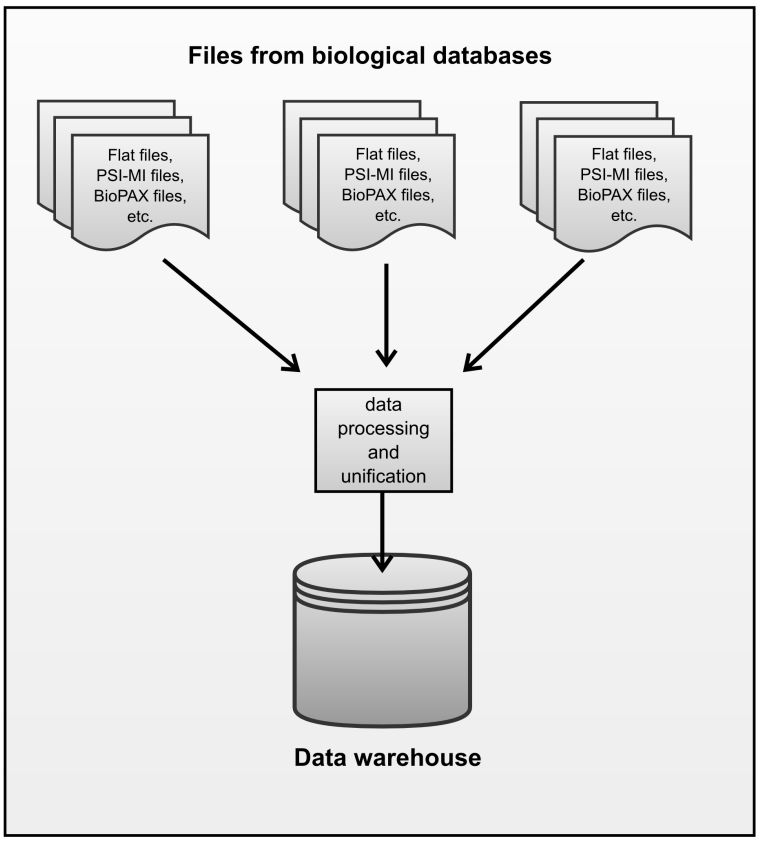


Fig. 2.2: Data warehousing model for integrating molecular biology databases. Data warehouses periodically download the information found in molecular biology databases, process this information and stores it into a central database. Different software is required to acquire the data, unify it and store it.

Workflow integration is partially similar to *view integration* because in this approach information is collected on-the-fly as well. However, the workflow software does not query a predefined list of services but instead those services are chosen by the user.

Data warehousing

In data warehousing the information from biological databases is periodically downloaded, cleaned, and re-formated for local storage (Fig. 2.2). In contrast to other approaches, data warehouses truly integrate many databases into one central repository (Stein, 2003). For this reason, they are traditionally considered as the most useful for data mining purposes (Han et al., 2005).

Data warehouses employ an update-driven approach in which information from multiple, heterogeneous sources is integrated and stored in a relational database for direct querying and analysis. Thus, data warehouses not only integrate, generalize and consolidate data to facilitate the mining of knowledge, but also speed-up queries and data retrieval due to the central and local access to the integrated data. Existing biological data warehouses include Atlas (Shah et al., 2005), BioWarehouse (Lee et al., 2006), Biozon (Birkland & Yona, 2006), BN++ (Küntzer et al., 2006), ConsensusPathDB (Kamburov et al., 2009) and PathwayCommons (Cerami et al., 2011).

A major shortcoming of data warehouses is the dependence on stable data formats from the biological databases. If such formats are changed, the appropriate parsers that feed the data warehouse will have to be updated as well. Another drawback of warehouses is that they may not contain the most up-to-date information and therefore may lag behind compared to the source databases (Stein, 2003).

2.4 Implementation of a local data warehouse to integrate multiple molecular biology databases

Despite the shortcomings of data warehousing this is the most efficient method for large-scale data analysis and was adopted for integrating molecular biology databases that characterize human genes and proteins. Aggregating approaches, such as *workflows* or *view integration* to integrate data are sub-optimal and time inefficient to answer queries that span the domain of more than one data source. Also, analysis involving thousands of entities are impractical because of the amounts of data that would need to be transferred from each of the sources. For example, if a user wants to know all proteins expressed in brain that are also kinases, the aggregating software (workflow or view integration) collecting data will have to query the two sources that have the information, download all proteins expressed in brain and all proteins that are kinases, and finally find the intersection from these two results. Moreover, data mining tools like and enrichments analysis (see Chapters 5 and 6) that require access to all available data are impractical under the *workflow* or *view integration methods*. In contrast, in a data warehouse the execution of queries is more efficient and reliable because the data is stored locally on a powerful database system and not distributed over the Internet.

The proposed approach is composed of three layers for processing, storing and unifying the acquired information that would be discussed in detail in the next sections.

In the following, a number of constrains that guided the design of the data warehouse are first presented:

Simplicity
The data warehouse must be a simple system easy to maintain and extend while being able to include as many different types of biological information as possible.

Incremental updates
It is required to keep track of the changes in the database content through time without creating copies of the database. The aim is to recover the status of the data warehouse at any time point and to easily identify recently added information as well as information that is no longer available.

Independent unification of identifiers
The mapping between the identifiers from different references systems changes continuously and the mapping process in charge of the unification of gene and protein identifiers should be able to remap identifiers once updates become available. For this reason, unified identifiers should not be considered definitive and should be stored separately from the original identifiers obtained from the data sources to allow future re-mappings.

Standard identifiers
The creation of new identifiers for genes and proteins should be avoided by relying as much as possible on existing identifiers.

Cellular entities
Integrated data should be easily available at the gene and protein level, but also be flexible enough to incorporate other cellular entities, for example, splice variants or protein complexes.

2.4.1 Design of the relational database

Integrated data types

A novel relational database schema was designed to store the following frequent types of information associated to genes and proteins:

Annotations based on categorical classifications
Often genes and proteins are annotated with categorical classifications of diseases, protein complexes, protein domains, sequence clusters etc. These types of associations are definitions assigned to one or many genes or gene products.

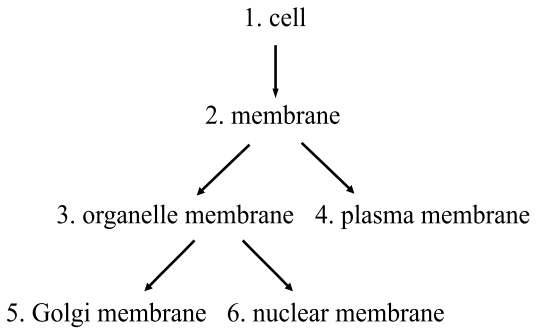


Fig. 2.3: Gene Ontology (GO) Simplified example of a cellular component ancestor-child relation of ontological terms. A gene or protein associated with any of the terms is considered to be also associated with all the ancestor terms. In this figure, if a gene is associated to the term *nuclear membrane* then, the gene is also associated to *organelle membrane*, *membrane*, and *cell*.

Annotations based on ontological terms

Ontologies are controlled vocabularies composed of hierarchically organized terms to describe the concepts and relationships of a specific knowledge domain. In the hierarchical structure of ontologies there are few high-level terms having broad definitions that become more and more precise down the hierarchy (Fig. 2.3). Ontological terms are annotated to genes and proteins at any level of the hierarchy and, because of the so-called *true path rule*, all ancestor terms in the hierarchy are considered to be annotated to the gene or protein as well. Each annotation of a gene with a term also includes an evidence code to indicate how the respective annotation was inferred.

Binary relations

Protein-protein interactions and gene co-expression are the most frequent types of binary relations found in biological databases. In the case of protein-protein interactions, some experimental methods like those based on co-immunoprecipitation usually identify groups of interacting proteins for which the physical pairwise interactions are unknown (Gavin et al., 2002). These groups tend to be small but in a few cases they comprise hundreds of proteins. Protein-protein interaction databases also inform about the experimental method used to detect the interaction and, in some cases, also contain information about the experimental conditions, such as the host organism, or whether protein participated as bait or prey.

Participation in signalling and metabolic pathways

Biological pathways summarize a series of biochemical reactions occurring within a cell to effect a cellular process such as DNA repair, apoptosis and protein folding. They describe the sequential steps in which enzymatic re-

actions occur, proteins complexes are formed, cargo is moved, genes are expressed, signals are transduced, etc. Biological pathways are rich in all sorts of information from the different processes occurring within the cell.

Numerical features

Some gene and protein features are given as numerical values, for example, sequence length, molecular weight or the fold change in the expression of a gene in microarray experiments.

Integration around *molecular concepts*

In order to integrate the different types of information available in the biological databases, the idea of *molecular concepts* (Tomlins et al., 2007) was used. This term refers to sets of related cellular entities (genes or proteins) with shared properties. The use of *molecular concepts* allows to develop a common framework to store most of the previously defined types of information annotated to genes and proteins as follows: For categorical classifications, each set of genes annotated to a particular classification constitutes a *molecular concept*. For ontologies, each set of genes annotated to a term is also treated as a *molecular concept*. In the case of ontologies the true path rule is followed and the association of genes is extended to all ancestor terms while the hierarchical structure is stored separately. Binary relations are considered as *molecular concepts* involving two genes or proteins or, in the case of protein-protein interactions, involving as many interaction partners as reported in the original database providing the information (although usually there are only two). The experimental method and other information is stored separately. For signalling and metabolic pathways, all genes or proteins members are considered as belonging to a single *molecular concept*. The information about the formation of protein complexes is stored as in the case of protein-protein interactions. No other information from biological pathways is included, instead links to the source database are kept such that the details can be accessed in the respective source. Finally, numerical values of gene expression from microarray and RNA-seq sources are treated as *molecular concepts* by using the threshold values, as defined in the respective publications, that indicate expression in a certain tissue or cell line. The numeric expression value is stored separately. Because numerical values are binned, in order to store them as *molecular concept*, they are considered in the data warehouse as categorical classifications.

In the following chapters, however, we will refer to the *molecular concepts* as *annotations*. Even though we think that the term *molecular concepts* might be more appropiate, the use of *annotation* for referring to any gene or protein feature such as function, location, structure, relations, etc., has gained wide adoption and we prefer to use it instead.

type of data	source databases
co-expressed genes	COXPRESdb
disease associations	PharmGKB, OMIM, UniProtKB keywords
drug associations	PharmGKB
metabolic and signaling pathways	BioCarta, KEGG, HumanCyc, Reactome, PID
molecular functions	GO, UniProtKB keywords, ENZYME
orthologous species	OrthoMCL
phenotype associations	MGI mammalian phenotype
protein complexes	CORUM, PID, BioCarta, PDB, Reactome
protein domain classifications	Pfam (family, clan and architecture), UniProtKB keywords, Interpro
protein-protein interactions	BioGRID, DIP, HPRD, IntAct, MINT
predicted protein-protein interactions	HiMAP, HomoMINT, OPHID, PIPs, STRING
sequence clusters	Ensembl Family, UniRef90
sub-cellular locations	GO, UniProtKB keywords
tissue expression	GNF expression, BurgeLab tissue expression

Table 2.2: Classification of data sources included in the data warehouse.

2.4.2 Molecular biology databases integrated

Currently, 31 publicly available databases containing characterizations for human genes and gene products have been integrated into the data warehouse.

These sources are: Gene Ontology annotations from the European Bioinformatics Institute (Barrell et al., 2009); clusters of similar sequences from Ensembl protein families (Flicek et al., 2008); protein domain architectures from Pfam (Finn et al., 2008) and InterPro (Hunter et al., 2008); metabolic and signaling pathways from BioCarta (Nishimura, 2001), HumanCyc (Romero et al., 2005), KEGG (Kanehisa et al., 2008), PID (Schaefer et al., 2009), and Reactome (Matthews et al., 2009); protein interactions and protein complexes from BioGRID (Stark et al., 2011), CORUM (Ruepp et al., 2008), DIP (Salwinski et al., 2004), HiMAP (Rhodes et al., 2005), HomoMINT (Persico et al., 2005), HPRD (Prasad et al., 2009), IntAct (Kerrien et al., 2007), MINT (Ceol et al., 2010), PDB (Velankar et al., 2005; Berman et al., 2003), OPHID (Brown & Jurisica, 2005), PIPs (McDowall et al., 2009), and STRING (Jensen et al., 2009); disease associations from OMIM (Amberger et al., 2009); enzyme classifications

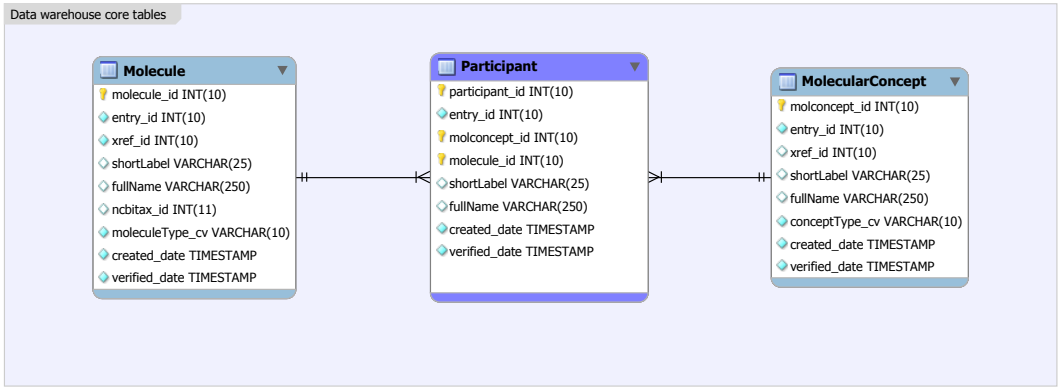


Fig. 2.4: Core data warehouse tables. The main tables of the database schema are `Molecule` where the references to genes and proteins from the integrated biological databases are stored, the `MolecularConcept` table that contains the *molecular concept* definitions and source database, and the `Participant` table which allows a many-to-many relation between the previous two tables.

from the Enzyme nomenclature database (Bairoch, 2000); gene tissue expression from the Novartis Gene Atlas (Su et al., 2002) and BurgeLab (Ramsköld et al., 2009); mammalian phenotype ontology classification of human genes by the Mouse Genome Database (MGD) (Blake et al., 2010); drug associations from PharmGKB (Klein et al., 2001); pairs of co-expressed proteins from COXPRESdb (Obayashi & Kinoshita, 2011); ortholog grouping of protein sequences from multiple genomes from OrthoMCL (Chen et al., 2006); UniRef90 similar sequence clusters (Suzek et al., 2007) and UniProt Keywords (The UniProt Consortium, 2010).

These sources are classified into the following fourteen categories: co-expressed genes, disease associations, drug associations, metabolic and signaling pathways, molecular function, orthologous species groups, phenotype associations, protein complexes, protein domain classifications, binary relations interactions, predicted protein-protein interactions, sequence clusters, sub-cellular locations and gene expression in tissues. (Table 2.2).

Database schema

The core of the database schema is composed of three tables: `MolecularConcept`, `Molecule` and `Participant` (Fig. 2.4). The first table holds the *molecular concepts*, the second contains the gene or protein identifiers and the third table relates one with the other. All tables have time stamps to identify when an entity was first added (`created_date`) and when was the last time it was verified to be up to date (`verified_date`). Data in the `Molecule` table contains the original identifiers reported by each integrated biological database.

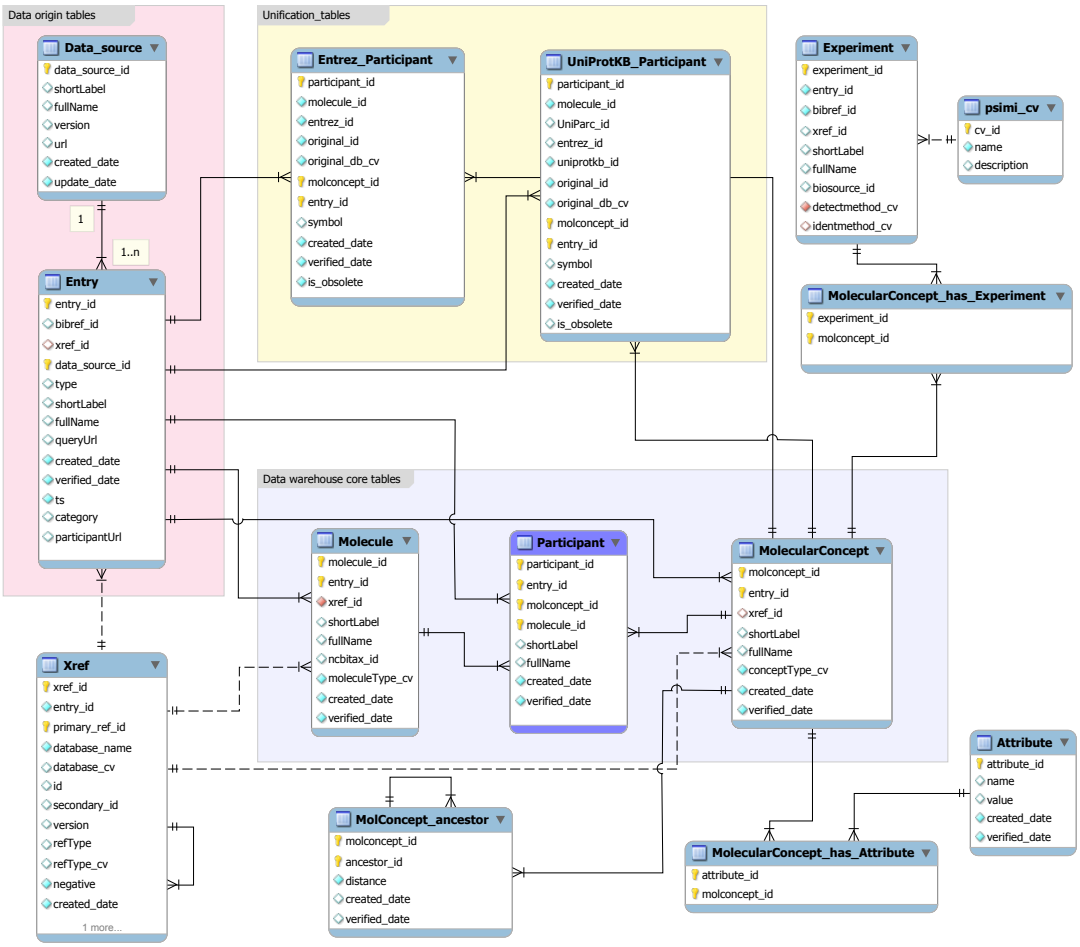


Fig. 2.5: Relational database schema of the data warehouse. This schema shows the relations of the core tables with different other tables that contain other information: The Data_source and Entry tables store the provenance of the integrated data, the Xref table stores the external identifiers of genes, proteins molecular concepts and literature references. The MolecularConcept_has_Experiment and the Experiment tables contain information about the experimental conditions that allowed the association of a gene or protein with the *molecular concept*. This information is usually associated to protein-protein interactions. The MolecularConcept_has_Attribute and the Attribute tables store miscellaneous information about molecular concepts, for example synonyms and descriptions. The dotted arrows shown represent 1:n non-identifying relations and the solid arrows represent 1:n identifying relations.

The complete database schema is depicted in Fig. 2.5. The tables `Data_source` and `Entry` contain the descriptions of all databases integrated in the data warehouse. This information is stored in two tables because a single biological database often provides different types of gene and protein features. In the database schema, the `Data_source` table contains entries for each biological database integrated and the `Entry` table contains the different types of knowledge found in the original database. For example, the Gene Ontology Association database (GOA) contains three distinct categories: *Biological Process*, *Cellular Component*, and

Molecular Function. In this case, the table `Entry` has an entry for each subdivision and relates them to the parent `Data_source` table containing the description of the GOA database. The `Entry` table classifies the integrated information following the data types described earlier: categorical classifications, ontologies, protein-protein interactions and pathways. Also, the `Entry` table subdivides the integrated data into the fourteen categories shown in Table 2.2.

All data loaded into the core tables `Molecule`, `MolecularConcept` and `Participant` is associated with the `Entry` table to discriminate the provenance of the information. The `Xref` table stores all identifiers, either for gene and proteins, *molecular concepts* identifiers, and literature references. The `MolecularConcept` table is associated with a number of other tables: the `MolConcept_ancestor` stores the ontological relations within *molecular concepts*, the `Experiment` table stores information about the experimental conditions and the `Attribute` table contains miscellaneous information like expression values.

The `Entrez_Participant` and `UniProtKB_Participant` tables contain unified information for gene and protein identifiers, respectively, and are fed by an identifier mapping process described in the Section 2.4.4.

The proposed database schema allows fast queries requiring any combination of *molecular concepts* and/or biological sources by joining the unification tables. For example, to find all diseases associated to kinase proteins the UniProtKB_Participant table is joined with itself to yield the required information:

```
SELECT distinct m.fullName
  FROM UniProtKB_Participant p1
   JOIN UniProtKB_Participant p2 using(UniProtKB_id)
   JOIN MolecularConcept m on m.molconcept_id = p2.molconcept_id
  WHERE p1.molconcept_id = <molecular concept id for the kinase function>
   AND  p2.entry_id  = <entry id corresponding to disease associations>
```

Because all imported information is stored as *molecular concepts* there is no need to use additional tables to query the data warehouse to search for other annotations. This is in contrast to other existing databases where, depending on the type of information being queried, different tables needs to be joined. Naturally, this demands a deeper understanding of the underlying schema.

Currently, this schema is implemented using the relational database MySQL (`http://www.mysql.com`); however, any other relational database using the SQL query language can be used.

2.4.3 Processing of source database data

Software loaders are in charge of translating the original data files from the source databases to the data warehouse schema. The loaders try to keep as much infor-

integrated database	download method
BioCarta	manual download
BioGRID	manual download
BurgeLab tissue expression	not updated
CORUM	automatic
COXPRESdb	manual download
DIP	manual download
Ensembl Family	automatic
ENZYME	automatic
GNF expression	not updated
GO	automatic
HiMAP	not updated
HomoMINT	automatic
HPRD	manual download
HumanCyc	manual download
IntAct	automatic
Interpro	automatic
KEGG	automatic
MGI mammalian phenotype	automatic
MINT	automatic
OMIM	automatic
OPHID	automatic
OrthoMCL	manual download
PDB	automatic
Pfam	manual download
PharmGKB	automatic
PID	manual download
PIPs	automatic
Reactome	automatic
STRING	manual
UniProtKB keywords	automatic
UniRef90	automatic

Table 2.3: Download method of integrated databases. The data in most biological databases is automatically downloaded and integrated into the data warehouse. About a third of the data sources need to be manually downloaded for diverse reasons: in some cases files are only available for download upon request while, for other cases, there are no standard places to find downloadable files, which have to be found manually. The GNF expression and BurgeLab datasets that indicate the expression of genes in the different tissues, as well as the predicted protein-protein interactions from HiMAP have not been modified after publication.

mation as possible from the source database while converting the information to molecular concepts.

Using the Python programming language (`http://python.org`), we developed loaders to read the standard file formats PSI-MI (for molecular interaction data Isserlin et al. (2011)) and BioPAX (for biological pathways Demir et al. (2010)). However, the implementation of the standard varies among the databases offering the data and it was necessary to create specific programs to read the files from each database. Particular programs to extract information from flat files contain-

ing tables where columns are separated by tabs, were also developed. Ontological information was acquired by reading and processing OBO files[1].

For each biological source database, a loader was created to read, extract and write the information into the data warehouse. In most cases, the loader automatically downloads the data from the source database; however, for several databases the data has to be downloaded manually, either because permission to download the file has to be granted or because the files are not stored in standard places or with standard names and must be searched for manually. However, only the download of the information demands personal attention; once the data is obtained the respective program is triggered to load the data into the data warehouse.

In three cases the available data is part of a publication and does not requires updating (Table 2.3).

2.4.4 Mapping and unification of gene and protein identifiers

The mapping of identifiers attempts to unify corresponding genes or proteins labeled under different reference systems. This is done by selecting a target identifier system (e.g. Entrez Gene) to which all other identifiers are mapped. During the unification process all information associated to the original identifier, as reported in the source database, is translated to the target identifier system. In the data warehouse we chose a gene and a protein identifier systems for unification, namely: Entrez Gene (Maglott et al., 2007) and UniProtKB (The UniProt Consortium, 2010). A gene and a protein identifier system were selected to reduce, for each system, the trade-offs involved when mapping from gene to protein identifiers and vice versa (Section 2.2).

The unification of identifiers uses mapping tables downloaded from third parties. This method was preferred over web-only applications, such as the Protein Identifier Cross-Reference (PICR) service (Côté et al., 2007) and IDconverter (Alibés et al., 2007), in which a mapping has to be requested from their sites on a one-by-one basis.

Several mapping tables are available. The International Protein Index (IPI) (Kersey et al., 2004) was considered as a good candidate to obtain mapping tables. IPI, although not a mapping service, is suitable for mapping identifiers because it maintains an updated and downloadable database of cross references between gene and protein sequence sources. Although IPI focuses on protein sequences, identifiers for gene sequences are also contained, thus permitting protein to gene mappings. The IPI algorithm works by creating pairwise alignments from all the sequences from the data sources and the reciprocally best matching pairs are com-

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

bined to form clusters of mapped proteins. However, a drawback of IPI is that protein variants are often part of the same cluster and can not be distinguished. UniProtKB also offers comprehensive mapping tables from many different identifier systems to their own identifiers. Yet the methods behind the mapping tables are not documented and protein variants are not considered either. The NCBI provides mappings only between their identifiers systems, Entrez and RefSeq, and Ensembl identifiers. None of these mapping tables contain information about obsolete or updated identifiers.

In order to obtain the best results, the mapping method implemented for the data warehouse merges different mapping sources. This method is centered around UniParc records to map human protein sequences (including protein variants) to several other protein identifiers. NCBI and UniProtKB mapping tables support mappings from protein identifiers to gene identifiers and from other identifiers not covered in UniParc. Deleted and updated identifiers from the Entrez and UniProtKB systems are obtained from NCBI and UniProtKB respectively (Fig 2.6).

The mapping algorithm is similar to the one presented by (Razick et al., 2008) to map protein-protein interactions. First, the primary identifier of a gene or protein referenced in a source database is analyzed. Although most source databases only use a single identifier, in some databases, specially in protein-protein interaction databases, each gene or protein has a primary identifier and several secondary identifiers. The primary identifier is revised to determine if it is considered obsolete and if replacement identifiers are found. This identifier is then mapped to gene and protein identifiers if mappings are available. When the identifier maps to several genes or proteins, secondary identifiers are used to resolve ambiguities. Also, if the primary identifier can not be mapped, secondary identifiers are used. The mapping algorithm is shown in Fig. 2.7.

The mapping algorithm creates two tables: `Entrez_Participant` and `UniProtKB_Participant` that contain the unified relations to the *molecular concepts*. These derived tables can be re-generated at any time from the core tables to reflect updates in the mapping tables or changes in the mapping algorithm. Furthermore, if unification is required using another identifier system, the mapping algorithm can be easily adapted because the process runs independently of the data collection and storage.

2.5 Summary

In this chapter, we presented a data warehouse to integrate the biological knowledge annotated to human genes and proteins found in several specialized databases. Currently, the data warehouse integrates almost 3 million such annota-

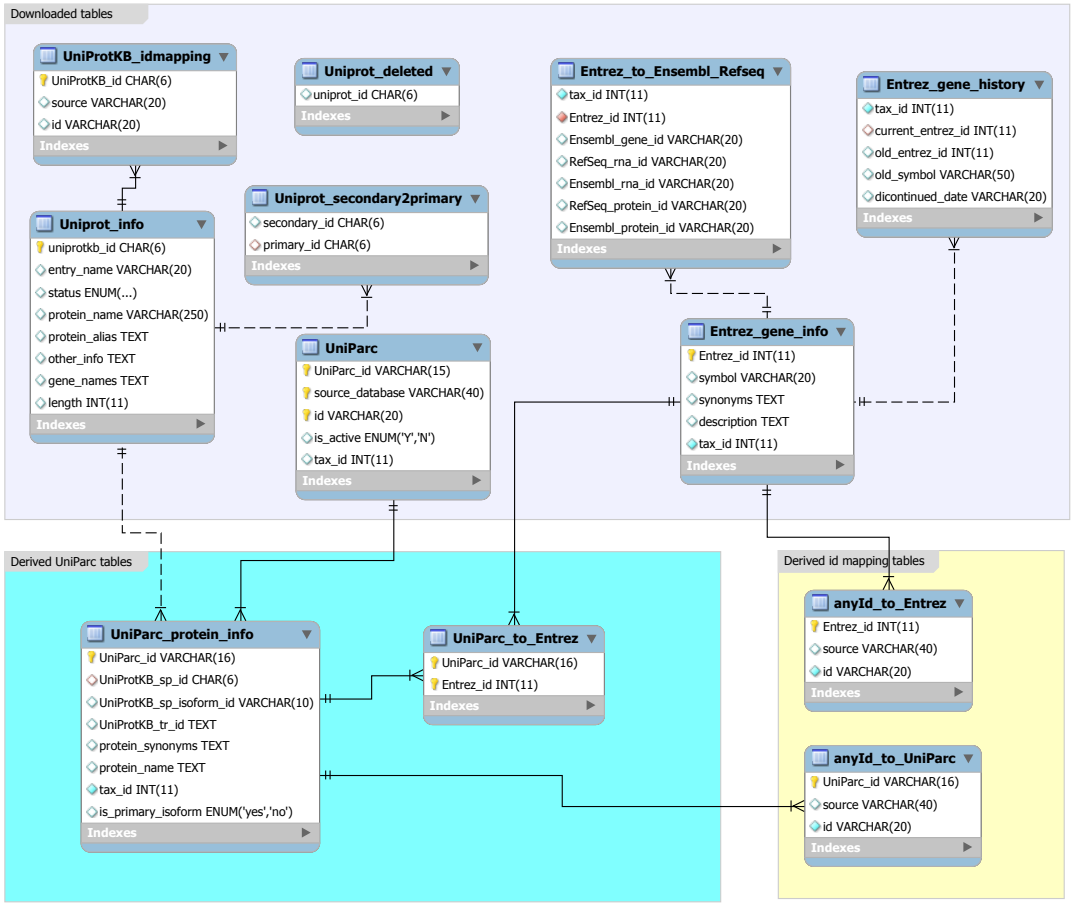


Fig. 2.6: Identifier mapping tables. Identifier mapping and identifiers information tables are obtained from third parties and stored into the database. The main source for mapping information is UniParc, which cross-references any amino acid sequence to different identifier systems. However, the UniParc table does not contain gene to protein mappings and some other mappings that are obtained from UniProtKB and NCBI and stored in the `UniProtKB_idmapping` and `Entrez_to_Ensembl_Refseq` tables. The mapping tables `anyId_to_UniParc` and `anyId_to_Entrez` merge the information for all the mapping sources mentioned.

tions from over 30 major molecular biology databases (Fig. 2.8). The integration of this large number of heterogeneous types of information is achieved through the transformation of the information acquired from the biological data sources into sets of genes and proteins sharing a common *molecular concept* or *annotation*.

The integrative solution presented is composed of a relational database that stores the information from the data sources implemented in MySQL; a series of programs written in Python that fetch, process and write the data into the database; and of a identifier mapping method developed also in Python that unifies the different identifier systems to reference genes and proteins. The data warehouse is composed over 10 thousand lines of programming code that have been extensively tested and debugged.

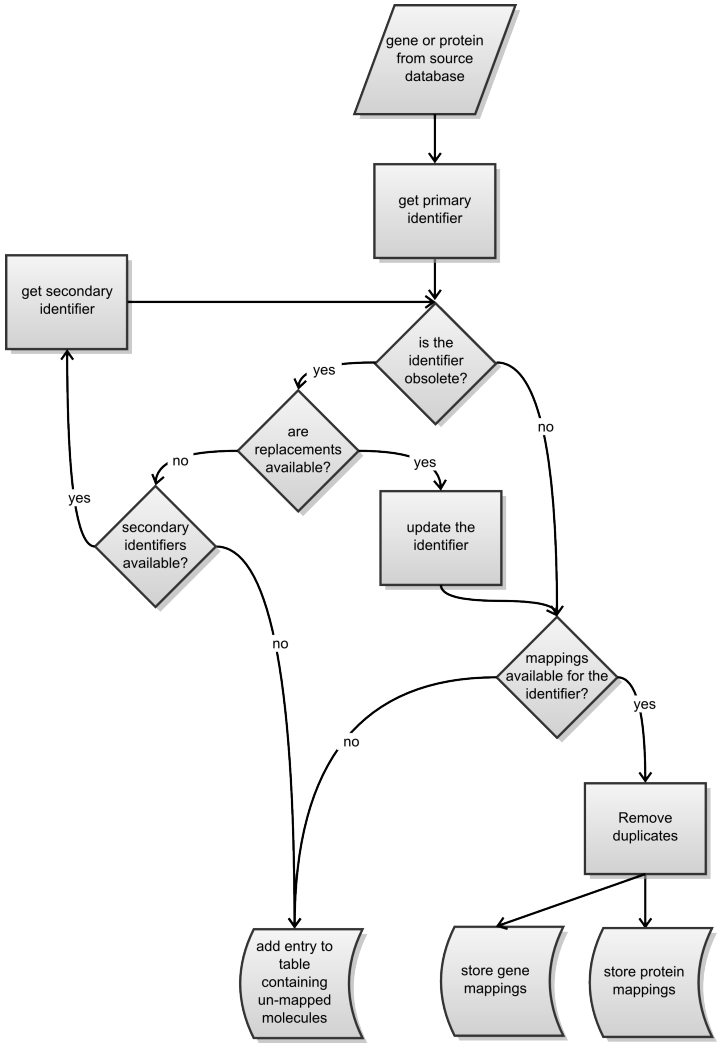


Fig. 2.7: Algorithm for mapping identifiers. An input gene or protein from a database source is always associated to a primary identifier and optionally to one or several secondary identifiers. The mapping algorithm searches for obsolete identifiers and tries to find replacements. If the primary identifier can not be mapped, secondary identifiers are used. If it is not possible to map the identifier, this issue is recorded for manual inspection.

The developed data warehouse is easy to maintain and expand because of the simplicity of the underlying database schema based on the *molecular concept* idea. This allows to store heterogeneous types of information from a large number of different sources only using few tables. As a result, the methods to access and store the information only need to interact with these few tables in a well defined and similar manner. If a new source of information needs to be added to the data warehouse, no new tables are required.

Updating of the mapping algorithm, if required, is straightforward because it is separated from the processes that acquire and store the data. This separa-

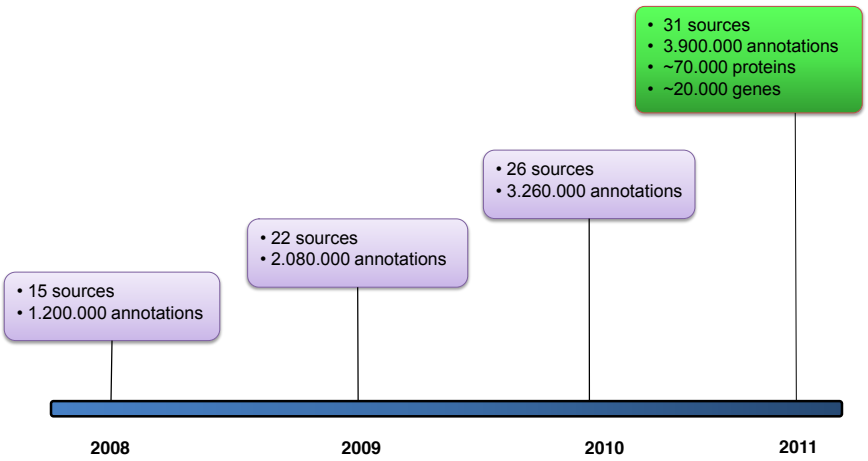


Fig. 2.8: Data warehouse growth. New biological data sources have been integrated steadily over the years into the data warehouse.

tion also makes the data warehouse flexible to consolidate unified results using different identifiers systems that may become popular in the future. Moreover, it is possible to unify information in novel ways, for instance, around protein complexes. Currently, integrated data is unified using both a gene and a protein identifier system.

3

Analysis of Human Protein Interaction Networks

This chapter reports the first use of our data warehouse for the analysis of human protein-protein interactions. The comprehensive analyses that were carried out helped to understand the quality of the publicly available data, particularly those obtained by predictions and by high-throughput techniques like yeast two-hybrid. The analysis contemplated six human predictions data-sets based on interspecies homology mapping, two yeast two-hybrid screen sets, and two sets derived from the scientific literature found on curated databases (Ramírez et al., 2007).

The assessment of the reliability and the potential bias of the datasets was analyzed through the use of novel methods which included a new graph measure called shared neighbors. The analysis revealed a lower quality of the interactions from high-throughput experiments, particularly from the yeast two-hybrid screens than that of the interactions obtained from computational predictions. Moreover, computational predictions that mapped high-throughput interactions from other organisms to humans reduced their quality. Although some computational predictions were found to be very reliable, they tended to have certain bias towards highly connected parts of the interaction network. Highly reliable interactions were those reported in at least three different experiments.

3.1 Introduction

Charting human protein-protein interactions (PPIs) on a cell-wide scale can afford key insights into the molecular basis of life and disease. Such investigations can uncover functional relationships of proteins in cellular processes and between the human host and its pathogenic intruders (Cusick et al., 2005; Uetz et al., 2006). However, we are still far from a complete human interactome map, which is estimated to contain between 200,000 to 400,000 interactions (Bork et al., 2004; Ramani et al., 2005). Most human PPIs known until recently have been obtained in small-scale experiments and collected from the literature in several databases (Zanzoni et al., 2002; Salwinski et al., 2004; Alfarano et al., 2005; Prasad et al.,

2009; Kerrien et al., 2007). Altogether, there are no more than 30,000-40,000 non-redundant, experimentally measured, human PPIs currently available in the literature and publicly accessible (Ramani et al., 2005; Gandhi et al., 2006). Since experimental data on PPIs is still scarce, bioinformatics methods have been used to predict human PPIs from experimentally derived interactions in other organisms such as yeast, worm, and fruit fly (Huang et al., 2004; Lehner & Fraser, 2004; Brown & Jurisica, 2005; McDermott et al., 2005; Rhodes et al., 2005; Persico et al., 2005). Lately, two comprehensive yeast two-hybrid (Y2H) screens have been performed with about 2,500 human proteins, resulting in about 5,000 PPIs (Rual et al., 2005; Stelzl et al., 2005).

Unfortunately, even though several publications on PPI predictions and Y2H screens offer some evaluation of the provided interaction data, no comprehensive comparison and quality assessment of the different human datasets have been performed. Therefore, the reliability, coverage, and inherent bias of predicted and new high-throughput data is unclear. This situation makes difficult the selection or the combination of one or another PPI dataset for further studies of the human interactome. Previous analyses of PPIs have mainly focused on high-throughput data for yeast, concluding that the reliability is limited and that about 50% of the interactions may be spurious (Mrowka et al., 2001; Bader & Hogue, 2002; Deane et al., 2002; Edwards et al., 2002; von Mering et al., 2002; Sprinzak et al., 2003; Bader et al., 2004; Reguly et al., 2006). Two studies on human PPIs have primarily compared literature-derived human datasets with each other or with interactions observed in other species (Ramani et al., 2005; Gandhi et al., 2006), but do not consider predicted datasets and recent Y2H screens.

In this work, the aim is to introduce a more comprehensive evaluation of the available human PPI datasets. This analysis differs from others because it additionally includes six large predicted datasets together with three high-confidence predicted subsets, two sizable high-throughput Y2H datasets, and two of the most comprehensive, manually literature-curated, datasets. It is noteworthy that all predicted datasets were published shortly before the results of the Y2H screens became available, making possible the independent assessment of different prediction methods with novel experimental data. The integrative data warehouse was used to study the similarities between the different datasets and also applied various alternative quality measures based on the Gene Ontology (GO), structurally known domain interactions, likelihood ratios, and topological network parameters.

Dataset	Original identifier type	#Original interactions	#Original identifiers	#Mappable identifiers	#Final Entrez Gene IDs	#Final interactions
			Predicted protein-protein interactions			
Bioverse	RefSeq	3218048	36996	16388	7711	233941
Bioverse-core	RefSeq	18327	1753	1481	1263	3266
HiMAP	Entrez Gene ID	38379	5790	5790	5790	38378
HiMAP-core	Entrez Gene ID	8833	2901	2901	2901	8832
HomoMINT	UniProt	10993	4129	4101	4184	10870
OPHID	UniProt	26425	4787	4738	4559	28255
POINT	GeneInfo Identifier	101783	13047	12982	12058	98528
Sanger	Ensembl (gene)	71806	6231	5788	5923	67518
Sanger-core	Ensembl (gene)	11652	3872	3661	3728	11131
			Experimental Y2H protein-protein interactions			
CCSB-HI1	Entrez Gene ID	2754	1549	1549	1549	2754
MDC	Entrez Gene ID	2124	1124	1124	1124	2033
			Manually curated protein-protein interactions			
HPRD-LS	Entrez Gene ID	3151	1983	1983	1983	3151
HPRD-SS	Entrez Gene ID	27955	7686	7686	7686	27955
IntAct	UniProtKB	6734	3484	2977	2988	5809

Table 3.1: Datasets of human protein-protein interactions included in the analysis and their conversion to Entrez Gene IDs. The number of interactions and identifiers of a certain type contained in the original datasets were obtained after removal of duplicates. The number of mappable identifiers is the number of original identifiers for which corresponding. The high throughput Y2H screens CCSB-HI1 and MDC were removed from the literature-curated sets HPRD-LS and IntAct. The complete HPRD without the Y2H dataset contains 30,956 interactions among 8,329 proteins. The numbers of final Entrez Gene IDs and interactions refer to the number of unique identifiers and interactions after performing the identifier mapping.

3.2 Data sources containing protein-protein interactions

Several human PPI datasets from different sources were used as follows (Table 3.1 and Table 3.2): six predicted datasets from Bioverse (McDermott et al., 2005), HiMAP (Rhodes et al., 2005), HomoMINT (Persico et al., 2005), OPHID (Brown & Jurisica, 2005), POINT (Huang et al., 2004), and Sanger (Lehner & Fraser, 2004); two experimental high-throughput Y2H screens CCSB-HI1 (Rual et al., 2005) and MDC (set LacZ4) (Stelzl et al., 2005); and two literature-curated datasets HPRD (release 6 of 1 January 2007) (Prasad et al., 2009) and IntAct (downloaded on 12 January 2007) (Kerrien et al., 2007). Interactions derived from protein complexes were discarded and the two high-throughput Y2H screens were removed from the literature-curated datasets. Also, the HPRD dataset was divided into two subsets of small-scale (HPRD-SS) and large-scale (HPRD-LS) experiments using 70 as threshold for the number of interactions reported in the

Dataset	Data Sources	Species	Homology Detection Method
Bioverse	DIP, GRID, PDB	50 species	PSI-BLAST (E-value < 1.0)
HiMAP	Gavin, Giot, Ito, Li, Uetz	fruit fly, worm, yeast	InParanoid
HomoMINT	MINT	15 species	InParanoid, results subsequently filtered by matching domain architecture between human and species orthologs.
OPHID	Giot, Li, Suzuki, von Mering, MIPS	fruit fly, worm, yeast	BLASTP reciprocal best-hit (E-value $< 10^{-5}$), filtered for hits with length > 50% of human protein sequence
POINT	DIP	fruit fly, worm, yeast, mouse	BLASTP (E-value not given)
Sanger	Gavin, Giot, Ho, Ito, Li, Tong, Uetz, von Mering	fruit fly, worm, yeast	InParanoid

Table 3.2: Predicted protein-protein interacting datasets. Comparison of interolog mapping methods for each predicted dataset regarding data sources of the protein interactions, species of the data sources, and homology detection methods applied. The listed data sources refer to the following studies: Gavin: *S. cerevisiae* TAP purified complexes (Gavin et al., 2002); Giot: *D. melanogaster* Y2H screen (Giot et al., 2003); Ho: *S. cerevisiae* HMS-PCI purified complexes Ho et al. (2002); Ito: *S. cerevisiae* Y2H screen Ito et al. (2001); Li: *C. elegans* Y2H screen (Li et al., 2004); Suzuki: *M. musculus* Y2H screen (Suzuki et al., 2001); Tong: *S. cerevisiae* synthetic genetic array (Tong et al., 2001); Uetz: *S. cerevisiae* Y2H screen (Uetz et al., 2000); von Mering grouped the *S. cerevisiae* data from Gavin, Ho, Ito, Uetz, Tong, and added 7,446 predicted interactions derived from gene neighborhood, gene fusion, and co-occurrence of genes (von Mering et al., 2005); further databases are: DIP for Bioverse, DIP for POINT, GRID, MINT, MIPS, and PDB.

same publication; HPRD-LS contains interactions primarily derived from Y2H experiments (Table 3.3). CCSB-HI1 was assembled using full-length proteins as baits and preys. The CCSB-HI1 authors reported that 78% of the interactions from a random sample of 217 interaction pairs could be verified using in vivo co-affinity purification assays (Rual et al., 2005). Since the MDC technology applied protein fragments of varying size as baits and preys (Stelzl et al., 2005), also the length of the fragments was considered in this study. The MDC authors described a verification rate of 62% interactions for a random sample of 116 interactions using membrane co-inmunoprecipitation assays and of 66% interactions for a random sample of 131 interactions obtained using pull-down experiments. In the following, the union of CCSB-HI1 and MDC consists of 4,770 unique interactions between 2,472 proteins and is referred to as the *combined Y2H datasets*.

Additionally, three core datasets available as subsets of Bioverse, HiMAP, and Sanger were used. This core subsets are assumed to consist of particularly reliable predictions. The Bioverse-core subset contains interactions of human proteins with a joint sequence similarity of at least 80% to the species interologs used to

Publication	Experimental method	HPRD	IntAct
Nakayama et al. (2002)	Y2H	118	125
Bouwmeester et al. (2004)	Y2H	128	1682
Colland et al. (2004)	Y2H	706	—
Goehler et al. (2004)	Y2H	154	151
Jin et al. (2004)	Co-immunoprecipitation	297	—
Lehner et al. (2004)	Y2H	110	95
Lehner & Fraser (2004)	Y2H	264	231
Ramachandran et al. (2004)	Protein array	102	109
Barrios-Rodiles et al. (2005)	LUMIER	430	—
Guo et al. (2005)	Far-western blotting	75	—
Rual et al. (2005)	Y2H	2619	2671
Stelzl et al. (2005), 2005	Y2H	3116	3137
Lim et al. (2006)	Y2H	704	706
Tsang et al. (2006)	Y2H	75	—
Camargo et al. (2007)	Y2H	—	191

Table 3.3: List of publications reporting large numbers of protein-protein interactions included in HPRD or IntAct.

predict the PPIs (Yu et al., 2004). As suggested in the original publication on HiMAP, the HiMAP-core subset consists of interactions with a 4-to-1 odds ratio that two proteins interact (that is, with a likelihood ratio exceeding the computed threshold 1,526) (Rhodes et al., 2005). The Sanger-core subset comprises predictions derived from high-throughput experiments in yeast, worm, and fly that were detected more than once in the experimental assays (Lehner & Fraser, 2004). Five consensus sets were assembled and named ConSet$n$ (with n ranging from 2 to 6) consisting solely of predicted PPIs contained in at least n of the predicted datasets. This resulted in consensus sets of decreasing size: ConSet2 with 38,258, ConSet3 with 10,844, ConSet4 with 4,747; ConSet5 with 1,565, and ConSet6 with 484 PPIs. To generate a random interaction dataset, a subset of 5,000 proteins was randomly picked from the list of all proteins contained in the different interaction datasets. From this subset, random selections with replacements were used to determine 30,000 interactions including homodimers. A second randomized set named HPRD-random was constructed by shuffling the protein identifiers of the complete HPRD dataset (HPRD-SS and HPRD-LS), thus preserving the network topology.

The diverse original protein and gene identifiers used in the interaction datasets namely Ensembl (gene) (Birney et al., 2006), Entrez Gene ID (Maglott et al., 2005), GenInfo Identifier (Benson et al., 2006), RefSeq (Pruitt et al., 2005), and UniProtKB (Wu et al., 2006a) were converted to Entrez Gene IDs to allow comparison between datasets (Table 3.1). Outdated original identifiers were substituted with new identifiers if they were available. Otherwise, identifiers

removed from the original database were ignored. An extraordinary case was Bioverse with more than half of the original RefSeq identifiers removed without a new substitute.

3.3 Network overlap computation

For pairwise comparisons of the human interaction datasets, three different matrices named F, S and T representing the mutual overlap were computed. In the following, V_i and E_i are the sets of proteins (referenced by Entrez Gene IDs) and their interactions, respectively, for an interaction dataset $i \in$ {Bioverse, Bioverse-core, HiMAP, HiMAP-core, HomoMINT, OPHID, POINT, Sanger, Sanger-core, CCSB-HI1, MDC, HPRD-LS, HPRD-SS, IntAct, HPRD-random, Random}. In the first matrix F, the fraction f_{ij} of proteins shared between the two sets V_i and V_j in ratio to the size of the smallest set used as normalizing denominator is defined as follows:

$$f_{ij} = |V_i \cap V_j| / min(|V_i|, |V_j|)$$

Similarly, the second matrix S contains the fraction s_{ij} shared between two interaction datasets E_i and E_j normalized by the size of the smallest interaction set:

$$s_{ij} = |E_i \cap E_j| / min(|E_i|, |E_j|)$$

In contrast to the second matrix, the third matrix T is normalized by another denominator $|K_{ij}|$ in order to ignore all proteins that are not part of both V_i and V_j. Here, if $(u, v) \in E_i$ represents the interaction between the proteins u and v in some interaction set i, K_{ij} is defined as:

$$K_{ij} = \{(u, v) \in E_i | u, v \in V_i \cap V_j\}.$$

Thus, K_{ij} as a subset of E_i is different from K_{ji} and lacks all interactions of proteins not shared between V_i and V_j. The values t_{ij} in the third matrix are calculated accordingly:

$$t_{ij} = |E_i \cap E_j| / min(|K_{ij}|, |K_{ji}|)$$

Fisher's exact test was applied to obtain a p-value for the observed overlap between each pair of interaction datasets. For this purpose, the total number n_{ij}

of all possible interactions including homodimers between all proteins contained in V_i or V_j was calculated as follows:

$$n_{ij} = |V_i \cup V_j| \times (|V_i \cup V_j| + 1)/2$$

3.4 Quality assessment methods

Functional similarity using Gene Ontology

To compute functional similarities of interacting proteins the *BPscore* similarity measure was used. This measure, introduced by Schlicker et al. (2006), is derived from the information contents of the annotated BP terms and their distances within the GO graph. The Gene Ontology (GO) (Harris et al., 2004) annotation was acquired on the of 12 December 2006 from the NCBI (`ftp://ftp.ncbi.nih.gov/gene/DATA/gene2go.gz`); annotations derived from protein-protein interactions where not used[1]. The calculation of the information content of each GO term was based on the term frequencies estimated using the annotation of human proteins from UniProtKB release 8.4. *BPscore* values range from 0 to 1, where 1 indicates the highest functional similarity. The *BPscore* computation did not include interactions of proteins solely annotated with quite general GO terms (very low information content ≤ 2).

To compare the functional similarity of interacting proteins qualitatively, the proteins were grouped into BP categories of top levels of the GO hierarchy using the script map2slim.pl from the GO Slim website (`http://www.geneontology.org/GO.slims.shtml`) (von Mering et al., 2002). The 20 categories used for analysis were derived from the GO Slim file at `http://www.geneontology.org/GO_slims/goslim_goa.obo` (Biswas et al., 2002). Maps of protein interaction density were constructed for every pair of BP categories (Ge et al., 2001). Homodimers were removed from the datasets for this analysis.

Structural domain interactions

To assess protein-protein interactions with structurally known domain-domain interactions, the iPfam version 20.0, containing 3,020 interactions between 2,147

[1] This is distinguished in the GO by the evidence code *inferred from physical interaction* (IPI)

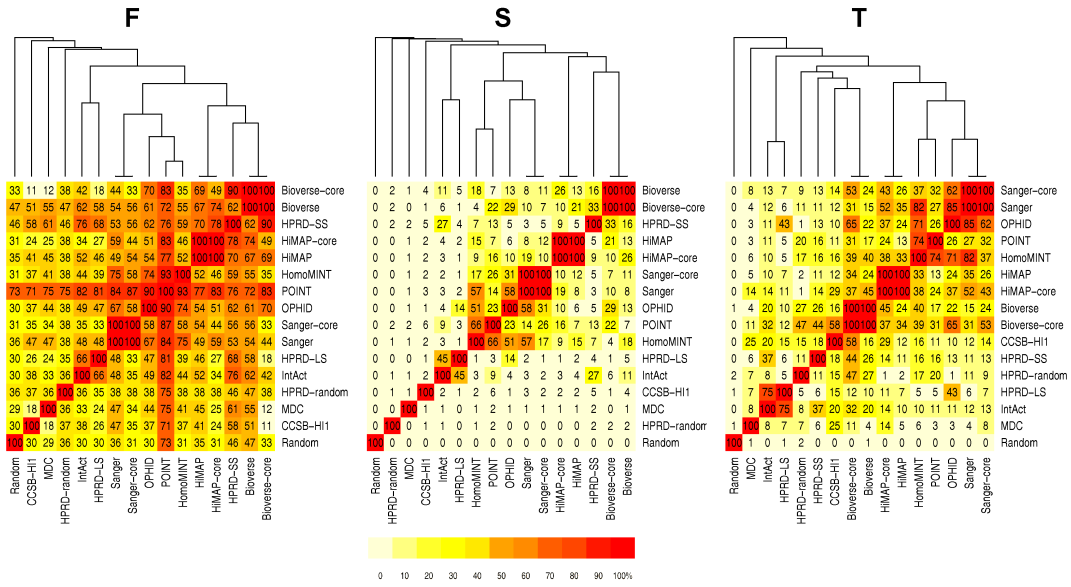


Fig. 3.1: Pairwise overlap of the human protein interaction datasets regarding proteins (matrix F), interactions (matrix S), and interactions ignoring all Entrez Gene IDs not contained in either interaction sets being compared (matrix T). The coloring schema from light yellow to dark red represents the respective overlap relative to the size of the smaller one of the two compared sets (values from 0% to 100%). Dendrograms are based on average linkage clustering using the respective matrix values.

distinct Pfam-A domains, was used (Finn et al., 2008). The domain composition of proteins was obtained from UniProtKB release 8.4.

Likelihood ratios

A set of positive and negative protein interactions was chosen to benchmark the prediction methods using likelihood ratios. The experimental dataset HPRD-SS and the combined Y2H datasets CCSB-HI1 and MDC were used as two different positive reference sets (PRS) because they have a low overlap with each other (Fig. 3.1). Unlikely protein interactions between proteins localized in the cellular membrane and proteins localized in the nucleus were used as negative reference set (NRS) (Jansen et al., 2003). 4,921 proteins localized in the membrane and 3,630 proteins localized in the nucleus were found according to the GO annotations. After removing 230 proteins annotated to both membrane and nucleus, the NRS set contains 15,949,400 protein pairs. The true positive and false positive rates and likelihood ratio (TPR, FPR and LR) were calculated as follows:

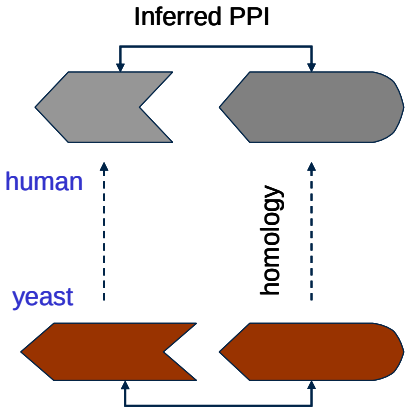


Fig. 3.2: Interolog mapping. Experimentally verified interactions in one species are transferred to other using protein homology.

$$\begin{aligned} \text{TPR} &= |E_i \cap \text{PRS}|/|\text{PRS}| \\ \text{FPR} &= |E_i \cap \text{NRS}|/|\text{NRS}| \\ \text{LR} &= \text{TPR}/\text{FPR} \end{aligned}$$

3.5 Evaluation of the interaction networks

3.5.1 Contents of the human interaction datasets

The following predicted datasets of human PPIs were used (Table 3.1): Bioverse, HiMAP, HomoMINT, OPHID, POINT and Sanger. Three core subsets with high-confidence interactions were additionally obtained from Bioverse, HiMAP, and Sanger. The prevailing bioinformatics methods used for the prediction of human PPIs is known as *interolog mapping* (Fig. 3.2). This method uses evolutionary relationships to infer interactions between human proteins from the interactions of homologous proteins of other species (Yu et al., 2004; Walhout et al., 2000). Thus, interolog mapping relies on the idea that protein interactions in one organism are likely to occur in another where homologous proteins exists (Gandhi et al., 2006). Table 3.2 summarizes the different approaches to determine human homologs from yeast, worm and the fruit fly by using either BLAST, PSI-BLAST or InParanoid (Altschul et al., 1997; O'Brien et al., 2005) in the predicted datasets studied.

Many interactions used by the PPI prediction methods originate from high-throughput experiments that are found in databases such as DIP, GRID, MINT, and MIPS (Table 3.2). Bioverse is the only predicted dataset that includes in-

teractions taken from structural complexes in the PDB (Berman et al., 2000). Additional procedures beyond interolog mapping were applied by some prediction methods to filter interactions for quality, for instance, HiMAP is assembled utilizing not only interologs, but also co-expression data, shared GO annotation of proteins, and domain-domain interactions (Rhodes et al., 2005). Interestingly, the authors of HiMAP report that interolog mapping is only moderately predictive of human interactions, which is in agreement with another recent analysis (Mika & Rost, 2006), and that confidence of an interaction is often associated with the existence of experimental data reported by different experiments (Rhodes et al., 2005).

Two comprehensive experimental Y2H datasets were used in this study: CCSB-HI1 (Rual et al., 2005) and MDC (Stelzl et al., 2005), which were published shortly after the predicted datasets became available. Also HPRD (Prasad et al., 2009) and IntAct (Kerrien et al., 2007), which consist of PPIs manually curated from the literature, were contemplated in this study. The number of curated publications by HPRD and IntAct is 18,525 and 1,098, respectively, with an overlap of 629 publications.

3.5.2 Overlap of the human interaction datasets

A distinction is made between the set of binary interactions and the set of proteins involved in the interactions. Each interaction can be formed by a pair of different proteins or by two instances of the same protein in the case of homodimers. Pairwise comparisons were performed for all datasets containing 16,318 proteins in total. The predicted dataset POINT contains the largest protein set with 12,058 different proteins. All other protein sets are mainly subsets of POINT (Fig. 3.1).

HiMAP contains more proteins in common with Bioverse and HPRD-SS (overlap $\geq 60\%$) than with the other protein sets. Also, HomoMINT, OPHID, POINT, and Sanger form a group of similar protein sets (pairwise overlap $\geq 67\%$). Regarding the three predicted datasets HiMAP, HomoMINT, and Sanger, it is interesting to note that HomoMINT and Sanger share 75% of their proteins in contrast to HiMAP that shares no more than 52% of its proteins with HomoMINT or Sanger, although all three datasets were produced using the InParanoid method to identify human homologues (O'Brien et al., 2005). This observed discrepancy may be a consequence of the different prediction methods employed by HiMAP (Rhodes et al., 2005). The proteins used by CCSB-HI1 and MDC in the Y2H screens are particularly different. Both sets share only 201 proteins overall (18%

of MDC, 13% of CCSB-H1). Regarding the literature-curated datasets, HPRD-LS and HPRD-SS share 66% and 76% proteins, respectively, with IntAct.

p-values were computed using Fisher's exact test to compare the size of the pairwise overlaps between interaction sets against randomly ocurring overlaps. The obtained p-values for the overlap between predicted data set were highly significant ($p \leq 1.44 \times 10^{-179}$) indicating a good overall agreement within these sets (Table 3.4, and Fig. 3.1**S**). Similar results were obtained when comparing the predicted datasets with HPRD-SS. However, the statistical significance of the overlap sizes with the Y2H datasets is different (Table 3.4). Larger p-values were obtained for the overlap with CCSB-HI1 ($3.64 \times 10^{-207} \leq p \leq 1.34 \times 10^{-11}$) and even larger values for MDC ($1.76 \times 10^{-41} \leq p \leq 0.018$). Presumably, since HPRD-LS is composed mainly of Y2H interactions, the p-values obtained are similar to those of the Y2H datasets CCSB-HI1 and MDC ($0 \leq p \leq 2.53 \times 10^{-4}$).

When reducing the interaction sets to those interactions that involve solely proteins contained in each of the respective two protein sets (Fig. 3.1**T**), HomoMINT, OPHID and Sanger are very similar (>70% pairwise overlaps). This presumably reflects the common methodology and data sources used by them (Table 3.2). A large fraction of 74% of the interactions in HomoMINT are also contained in POINT. The two predicted datasets Bioverse and HiMAP show only a small 24% interaction overlap. The Y2H dataset CCSB-HI1 has an overlap of 58% with Bioverse-core and fall into a group containing Bioverse, Bioverse-core, and HPRD-SS. The overlap of the two Y2H screens CCSB-HI1 and MDC amounts to 17 interactions only, and the MDC overlap with other datasets is limited to at most 41 out of 2033 interactions. Those results are in good agreement with other studies that have demonstrated a generally low overlap of high-throughput interaction data (Ramani et al., 2005; Gandhi et al., 2006; Mrowka et al., 2001; Deane et al., 2002; Edwards et al., 2002; von Mering et al., 2002; Sprinzak et al., 2003; Bader et al., 2004; Reguly et al., 2006; Goll & Uetz, 2006). However, HPRD-LS and IntAct have a large overlap due to the curation of identical high-throughput datasets (Table 3.3).

3.5.3 Assessment of protein-protein interactions

Functional analysis

Functional similarity of proteins has been used to predict (Rhodes et al., 2005; Ben-Hur & Noble, 2005; Wu et al., 2006b) and assess their interactions (Lehner & Fraser, 2004; Brown & Jurisica, 2005; Persico et al., 2005; Bader & Hogue, 2002;

	Bioverse	Bioverse core	HiMAP	HiMAP core	Homo- MINT	OPHID	POINT	Sanger	Sanger core	CCSB-HI1	MDC	HPRD-LS	HPRD-SS	IntAct	HPRD Random	Random
Bioverse	0	0	0	0	0	0	0	0	0	4.62E−55	3.14E−4	1.36E−76	0	0	1.35E−134	1.00
Bioverse core		0	0	0	1.44E−179	0	0	0	0	3.34E−21	1.79E−2	5.56E−16	0	2.05E−289	2.38E−66	1.00
HiMAP			0	0	0	0	0	0	0	1.97E−43	4.34E−5	3.47E−39	0	5.38E−230	0.82	1.00
HiMAP core				0	0	0	0	0	0	6.13E−41	2.93E−6	2.36E−32	0	3.61E−169	0.97	1.00
Homo-MINT					0	0	0	0	0	1.02E−68	2.09E−12	1.71E−26	0	3.65E−196	1.65E−75	1.00
OPHID						0	0	0	0	5.69E−20	1.44E−4	0	0	1.76E−267	0.998413	1.00
POINT							0	0	0	2.50E−192	6.47E−32	7.48E−108	0	0	0	0.87
Sanger								0	0	1.29E−43	1.24E−6	3.72E−24	0	5.84E−143	2.03E−12	1.00
Sanger core									0	2.35E−47	8.13E−7	1.72E−25	0	1.41E−132	6.40E−14	1.00
CCSB-HI1										0	1.34E−11	2.85E−18	3.64E−207	3.31E−45	1.04E−10	0.99
MDC											0	2.53E−4	1.76E−41	4.06E−11	2.06E−2	0.97
HPRD-LS												0	5.78E−160	0	5.11E−6	0.95
HPRD-SS													0	0	0	1.00
IntAct														0	1.69E−27	1.00
HPRD random															0	1.00
Random																0

Table 3.4: Evaluation of the overlap sizes of interaction sets. Fisher's exact test was used to compute the p-values.

Dataset	Functional Similarity using GO				Domain Interactions using iPfam		Overlap in percent
	Average *BPscore*	GO assignments in percent	Number of homodimers in percent	Average information content	Domain assignments in percent	Number of DDI-validated interactions (%)	
Bioverse-core	0.814	87.94	12.58	12.6	91.40	2090 (70.0)	51.92
HiMAP-core	0.793	92.07	0.00	12.6	92.33	3934 (48.2)	42.70
HiMAP	0.785	94.64	0.00	12.8	91.77	12764 (36.2)	35.02
HPRD-SS	0.634	91.20	6.78	12.8	89.36	6301 (25.2)	30.76
Bioverse	0.575	87.62	0.81	13.1	92.87	94569 (43.5)	25.70
Sanger-core	0.558	70.10	1.90	12.9	78.09	1049 (12.1)	26.62
IntAct	0.554	77.71	3.44	12.2	75.45	693 (15.8)	33.20
HomoMINT	0.525	63.85	5.81	12.3	75.90	992 (12.0)	35.24
OPHID	0.510	67.97	0.00	12.5	85.22	1528 (6.35)	16.45
HPRD-LS	0.488	78.32	1.52	12.6	81.72	169 (6.56)	20.92
CCSB-HI1	0.464	53.05	5.19	12.3	56.57	250 (16.0)	37.22
POINT	0.457	65.33	4.18	11.9	70.82	6818 (9.77)	30.32
Sanger	0.437	70.47	0.55	12.1	80.93	3016 (5.52)	21.63
HPRD-random	0.394	77.26	6.25	12.5	76.11	1598 (6.78)	38.27
MDC	0.390	57.99	1.48	12.7	65.22	51 (3.85)	23.01
Random	0.335	52.94	0.01	12.2	58.40	251 (1.43)	5.63

Table 3.5: Quality assessment using functional GO similarity and structural domain interactions. The human interaction datasets are ranked by the average *BPscore*. For each dataset, the percentage of interactions with biological process (BP) terms assigned to both interacting proteins is given next to the percentage of homodimers, the fraction of protein self-interactions. The average information content is calculated from the information content of the BP protein annotations. The percentage of DDI-validated interactions using iPfam relates to the fraction of PPIs in which both proteins have Pfam domain assignments. The rightmost column shows the overlap size of the subset of PPIs with a $BPscore \geq 0.8$ and the subset of all DDI-validated PPIs, relative to the size of the union of both subsets.

von Mering et al., 2002; Sprinzak et al., 2003; Bader et al., 2004; Reguly et al., 2006). A previous study showed that the application of similarity measures based on the biological processes in which proteins are involved can be used to validate PPIs (Guo et al., 2006). The *BPscore* measure, which uses the Gene Ontology (GO) annotation (Harris et al., 2004) to calculate the similarity of biological processes annotated to interacting proteins (Schlicker et al., 2006), was used.

For each dataset was calculated: (*i*) the average *BPscore*, (*ii*) the fraction of PPIs in which both proteins are annotated with biological process terms, (*iii*) the fraction of homodimers, and (*iv*) the average information content of the BP protein annotations (Table 3.5). The values (*ii*)-(*iv*) were included to identify possible bias that could affect the *BPscore*. The fraction (*ii*) of annotated PPIs varies greatly between datasets from 53.05% to 94.64%. HPRD-SS is enriched with well-annotated pairs of interacting proteins while the Y2H datasets reach much lower values (Table 3.5). However, one should bear in mind that Y2H results discover new interactions between proteins that are not yet as well studied as in HPRD-SS. The values for the low numbers of homodimers (*iii*) and for the average information content of each dataset (*iv*) negate a possibly biased

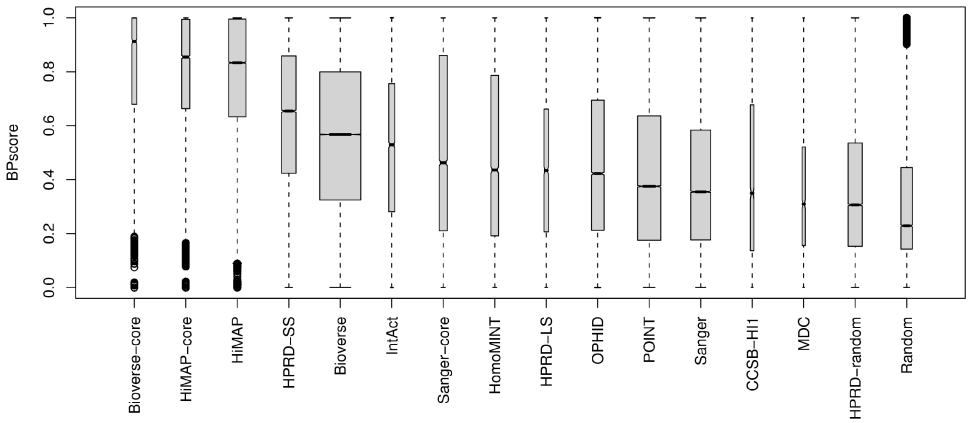


Fig. 3.3: Dataset comparison using boxplots based on GO biological processes annotated to interacting proteins. The datasets are ordered by the *BPscore* median from left to right. The area of each box is proportional to the size of the corresponding dataset.

BPscore (Table 3.5). In particular, the average *BPscore* values ignoring homodimers are only slightly smaller than if homodimers are included. For instance, the average *BPscore* of the Bioverse-core dataset containing the largest amount of homodimers (411 out of 3,266) is 0.774 without homodimers compared to the original value of 0.814. Altogether, there appears to be no overall difference in the BP annotation quality of homodimers and other protein interactions that may significantly affect the average *BPscore*.

In Table 3.5, the highest *BPscore* averages are found for Bioverse-core (0.814), HiMAP-core (0.793), HiMAP (0.785), and HPRD-SS (0.634). The lowest scores are assigned to Random (0.335), MDC (0.390), and HPRD-random (0.394). The predicted datasets HomoMINT, OPHID, POINT, and Sanger reach values between 0.437 and 0.525, which are similar to the average *BPscore* obtained for CCSB-HI1 (0.464) and HPRD-LS (0.488). The average *BPscore* increases in the consensus sets from 0.533 ConSet2, 0.535 ConSet3, 0.538 ConSet4, 0.614 ConSet5 and 0.666 ConSet6. However, the latter value is not higher than the average *BPscore* of Bioverse-core or HiMAP, and the size of ConSet6 is very small (484 interactions).

The top-ranking datasets, Bioverse-core, HiMAP-core, HiMAP, and HPRD-SS are enriched by high-scoring interactions as can be seen in the *BPscore* boxplots (Fig. 3.3) and the *BPscore* distributions (Fig. 3.4). The predicted datasets Bioverse, HomoMINT, OPHID, POINT, Sanger, and Sanger-core show bimodal distributions with frequent *BPscore* values in the range of 0.1-0.2 and of 0.9-1.0. This pattern is similar to the one obtained from the experimental CCSB-HI1

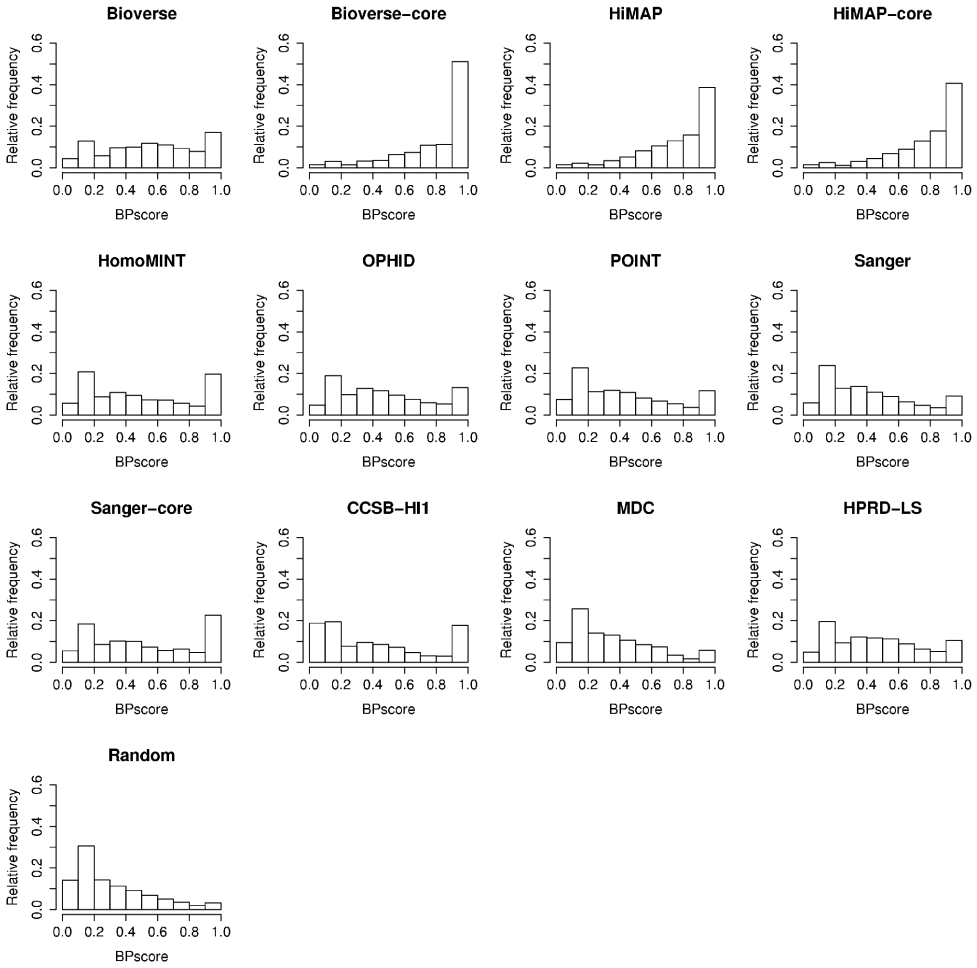


Fig. 3.4: Dataset comparison using histograms of the *BPscore* distribution. The *BPscore* similarity values based on the biological processes annotated to interacting proteins are binned in 0.1-steps.

dataset and for HPRD-LS. (Fig. 3.3 and Fig. 3.4). Remarkably, the *BPscore* distribution of the MDC dataset is similar to those of the random datasets, but slightly enriched with high *BPscore* values.

Furthermore, to compare the functional similarity of interacting proteins, not only the quantitative measure *BPscore* was used but also a classification approach in which the interacting proteins are categorized according to the biological processes annotated on GO. To this end, the frequency of PPIs was analyzed with respect to 20 relevant biological processes from top levels of the GO hierarchy (von Mering et al., 2002; Ge et al., 2001). In the case of datasets with many accurate PPIs, a high density of interactions with both proteins in the same GO category should be observed particularly along the diagonal of the computed 2D histograms (Fig. 3.5). The MDC dataset is the only dataset that does not

clearly demonstrate this expectation. Apart from that, Fig. 3.5 visualizes apparent bias and peculiarities of each human interaction dataset in functional terms. For example, the categories 'transport' and 'secretion' as well as 'biosynthesis' and 'amino acid and derivative metabolism' represent commonly related biological processes except for the datasets MDC, HPRD-LS and IntAct. Among others, the association of the categories 'cell differentiation' with 'development' as well as 'cell communication' with 'behavior' is present in all predicted datasets and in the literature-curated datasets HPRD-SS and IntAct, but not in the Y2H screens CCSB-HI1 and MDC or in HPRD-LS. Also, some predicted datasets do not show certain functional relationships, for instance, most protein interactions between 'electron transport' and 'metabolism' are absent in the HomoMINT dataset, and the categories 'regulation of biological process' and 'cell death' are not linked in the datasets HomoMINT, POINT, and Sanger.

Structural domain interactions

Protein-protein interactions may be caused either by the physical contact of domain surfaces or by short peptides binding to domains (Albrecht et al., 2005; Pawson & Nash, 2003). After decomposing interacting proteins into their consitutent Pfam domains, those interactions explained by at least one of the 3,020 iPfam (Finn et al., 2005) domain-domain interactions (DDI) were considered as 'DDI-validated'. Each PPI was classified as validated, non-validated, or impossible to evaluate because at least one of the two interacting proteins does not contain known domains. However, it must be pointed out that this validation method is only a simplistic measure and may result in incorrect classifications in several cases. For instance, if the complex structure of two interacting domains is not known yet and thus is missing in iPfam, the corresponding PPI would be erroneously we assigned to the non-validated class. Another case of misclassification occurs when an interaction that does not take place in vivo is validated. For example, interactions between proteins that are never expressed simultaneously or present in different cellular locations will be validated if they contain domains interacting in iPfam.

Despite these shortcomings, the evaluation of PPIs based on iPfam DDIs yields results comparable to those obtained by the computation of functional similarity based on GO (Table 3.5 and Fig. 3.6). In particular, datasets with a high average *BPscore* also have a large proportion of DDI-validated PPIs as follows: Bioverse-core with *BPscore* 0.814 and 70.0% of DDI-validated interactions, HiMAP-core with *BPscore* 0.793 and 48.2% of DDI-validated interactions, HiMAP with *BP-*

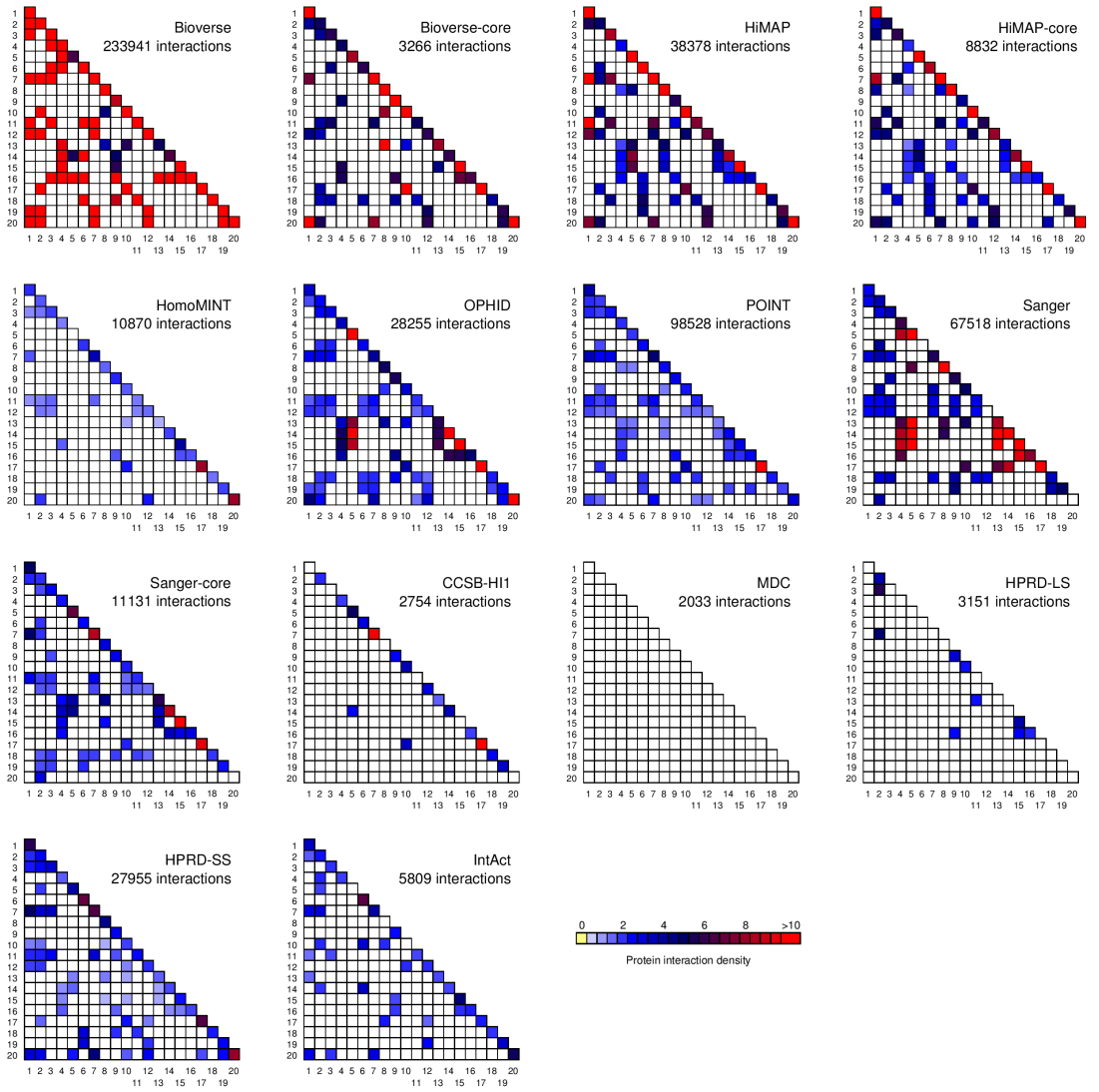


Fig. 3.5: 2-D histograms of the distribution of PPIs according to the biological processes annotated to the interacting proteins in each human interaction dataset. Each human interaction dataset is depicted by a triangle matrix whose axes represent top levels of the GO hierarchy. The dot color in the histograms reflects the protein interaction density that is the ratio of the number of PPIs assigned to the respective matrix cell divided by the total number of PPIs possibly formed; the total number of possible PPIs was derived by counting the members of the respective GO category, and the density was normalized to 1,000 possible PPIs. The protein interaction density is not shown if the observed number of PPIs is non-significant (p-value ≥ 0.01, using Fisher's exact test as in case of the overlap computation). The numbers along the axes represent the following GO categories: 1: cellular process; 2: cell communication; 3: cell differentiation; 4: cellular physiological process; 5: amino acid and derivative metabolism; 6: cell death; 7: cell motility; 8: electron transport; 9: nucleobase, nucleoside, nucleotide and nucleic acid metabolism; 10: transport; 11: development; 12: physiological process; 13: metabolism; 14: biosynthesis; 15: catabolism; 16: macromolecule metabolism; 17: secretion; 18: regulation of biological process; 19: response to stimulus; 20: behavior.

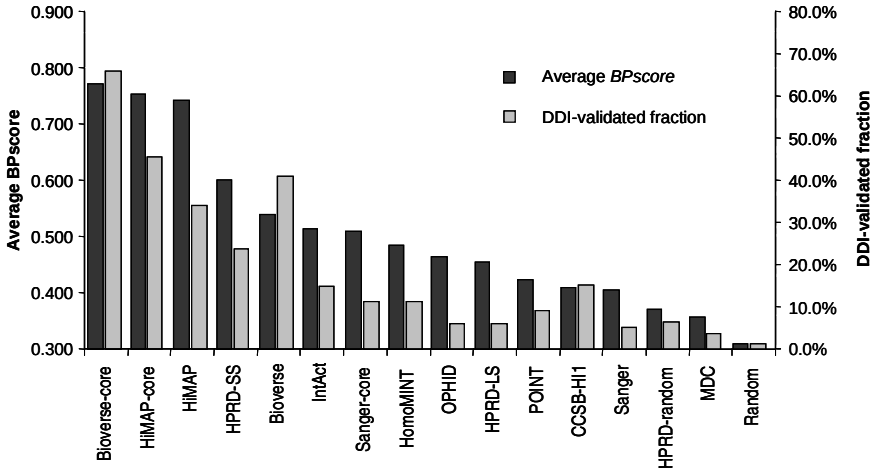


Fig. 3.6: Comparison of the average functional GO similarity *BPscore* values and the percentage of validated structural domain-domain interactions in relation to the overall number of protein-protein interactions with Pfam domain assignments as shown in Table 3.5.

score 0.785 and 36.2% of DDI-validated interactions. The high values obtained for the HiMAP datasets are partially due to the fact that they have been compiled by deriving putative DDIs from PPIs in HPRD first and then using these DDIs to predict other PPIs. High scores for the Bioverse datasets may be reached because they additionally include known interactions from 3D complex structures, which also results in a large number of DDI-validated interactions. Lower values are obtained for the other predicted datasets of PPIs and the Y2H screens. For example, only 16.0% and 3.8% of the PPIs in CCSB-HI1 (250 of 1,558) and MDC (51 of 1,326), respectively, could be DDI-validated, and the average *BPscore* values were 0.464 and 0.390, respectively. Moreover, the fraction of interactions DDI-validated of MDC is close to the randomized datasets HPRD-random (*BPscore* 0.394 and 6.8%) and Random (*BPscore* 0.335 and 1.4%).

For each human dataset, the overlap of subsets of PPIs with high *BPscore* values not smaller than 0.8 with subsets containing all DDI-validated PPIs was examined (Table 3.5). Regarding Bioverse-core, HiMAP-core, and the Y2H dataset CCSB-HI1, significant overlaps with 51,9%, 42.7%, and 37.2%, respectively, are observed. These overlaps may primarily consist of very reliable PPIs as judged by both quality measures *BPscore* and DDI-validation.

Comparison with reference sets

Even though the analyses so far have pointed to the limited quality of experimental protein interaction datasets, both HPRD-SS and the combined Y2H datasets

	Computed ratios			Number of PPIs in overlap		Average *BPscore* of PPIs in overlap		DDI-validated PPIs in overlap in percent	
Dataset	**TPR**	**FPR**	**LR**	**PRS (%)**	**NRS (%)**	**PRS**	**NRS**	**PRS**	**NRS**
Use of HPRD dataset as PRS set									
Bioverse-core	0.0390	1.80E-06	21405.0	1088 (33.31)	29 (0.89)	0.770	0.554	60.33	24.14
HPRD-SS	1.0000	6.60E-05	15189.9	27955 (100.0)	1050 (3.76)	0.620	0.471	25.22	14.11
HiMAP-core	0.0270	3.60E-06	7544.9	767 (8.68)	58 (0.66)	0.825	0.568	52.05	25.00
IntAct	0.0550	1.50E-05	3762.1	1543 (26.56)	234 (4.03)	0.676	0.369	34.66	6.22
HiMAP	0.0520	3.50E-05	1476.1	1454 (3.79)	562 (1.46)	0.818	0.573	49.38	8.85
HomoMINT	0.0290	3.30E-05	867.7	806 (7.41)	530 (4.88)	0.746	0.302	40.92	0.88
OPHID	0.0470	7.80E-05	597.2	1301 (4.60)	1243 (4.40)	0.752	0.318	40.43	3.17
Sanger-core	0.0190	3.30E-05	569.5	530 (4.76)	531 (4.77)	0.818	0.249	50.30	0.79
CCSB-HI1	0.0052	9.30E-06	559.1	146 (5.30)	149 (5.41)	0.829	0.247	62.40	0.00
HPRD-LS	0.0044	1.30E-05	329.1	124 (3.94)	215 (6.82)	0.717	0.361	26.61	4.64
POINT	0.1300	4.30E-04	300.4	3616 (3.67)	6867 (6.97)	0.740	0.302	42.59	1.34
Bioverse	0.1600	7.20E-04	225.6	4552 (1.95)	11512 (4.92)	0.738	0.463	64.18	33.49
MDC	0.0015	8.80E-06	165.9	41 (2.02)	141 (6.94)	0.809	0.296	66.67	0.00
Sanger	0.0290	2.60E-04	111.0	816 (1.21)	4196 (6.21)	0.802	0.273	44.52	0.97
HPRD-random	0.0160	2.50E-04	63.0	444 (1.43)	4021 (12.99)	0.963	0.294	79.56	0.53
Random	0.0001	1.70E-04	0.6	3 (0.01)	2702 (9.01)	0.494	0.272	0.00	0.40
Use of combined Y2H datasets as PRS set									
CCSB-HI1	0.5800	9.30E-06	61802.2	2754 (100.00)	149 (5.41)	0.416	0.247	16.05	0.00
MDC	0.4300	8.80E-06	48210.8	2033 (100.00)	141 (6.94)	0.362	0.296	3.85	0.00
Bioverse-core	0.0075	1.80E-06	4150.8	36 (1.10)	29 (0.89)	0.891	0.554	88.57	24.14
HiMAP-core	0.0120	3.60E-06	3286.0	57 (0.65)	58 (0.66)	0.835	0.568	53.85	25.00
IntAct	0.0130	1.50E-05	857.4	60 (1.03)	234 (4.03)	0.752	0.369	50.00	6.22
HPRD-SS	0.0370	6.60E-05	560.5	176 (0.63)	1050 (3.76)	0.820	0.471	62.67	14.11
HomoMINT	0.0150	3.30E-05	460.6	73 (0.67)	530 (4.88)	0.825	0.302	53.33	0.88
HPRD-LS	0.0060	1.30E-05	451.0	29 (0.92)	215 (6.82)	0.553	0.361	30.43	4.64
HiMAP	0.0140	3.50E-05	410.5	69 (0.18)	562 (1.46)	0.817	0.573	53.23	8.85
Sanger-core	0.0130	3.30E-05	390.4	62 (0.56)	531 (4.77)	0.806	0.249	57.69	0.79
OPHID	0.0099	7.80E-05	126.4	47 (0.17)	1243 (4.40)	0.761	0.318	50.00	3.17
POINT	0.0390	4.30E-04	90.1	185 (0.19)	6867 (6.97)	0.84	0.302	59.62	1.34
Sanger	0.0180	2.60E-04	69.3	87 (0.13)	4196 (6.21)	0.802	0.273	54.17	0.97
Bioverse	0.0290	7.20E-04	40.7	140 (0.06)	11512 (4.92)	0.799	0.463	73.68	33.49
HPRD-random	0.0052	2.50E-04	20.8	25 (0.08)	4021 (12.99)	0.935	0.294	82.61	0.53
Random	0.0004	1.70E-04	2.5	2 (0.01)	2702 (9.01)	0.607	0.272	0.00	0.40

Table 3.6: Quality assessment using likelihood ratios. The human interaction datasets are ranked by decreasing likelihood ratios (LR). The ratios TPR, FPR, and LR are computed using the positive reference set (PRS) and the negative reference set (NRS). The PRS set consists either of HPRD-SS or of the combined Y2H datasets. The number of PPIs in the overlap of the respective dataset with the PRS or NRS sets, their average *BPscore*, and the percentage of DDI-validated PPIs are also listed.

CCSB-HI1 and MDC were used as substitutes of an ideal positive reference set (PRS), which is not available yet for the human interactome. Also, a negative reference set (NRS) was constructed from all possible 15,949,400 interactions between proteins annotated to localize in the cell nucleus with proteins to localize in the membrane (Jansen et al., 2003). A likelihood ratio (LR) was calculated for all datasets (Table 3.6). Additionally, the LR was plotted versus the number of interactions in each dataset (Fig. 3.7).

Higher LRs indicate enrichment of a dataset with true positives in relation to false positives, but the derived LR values can be taken only as a relative measure because the proportion of true positives in the PRS set is unknown. As can

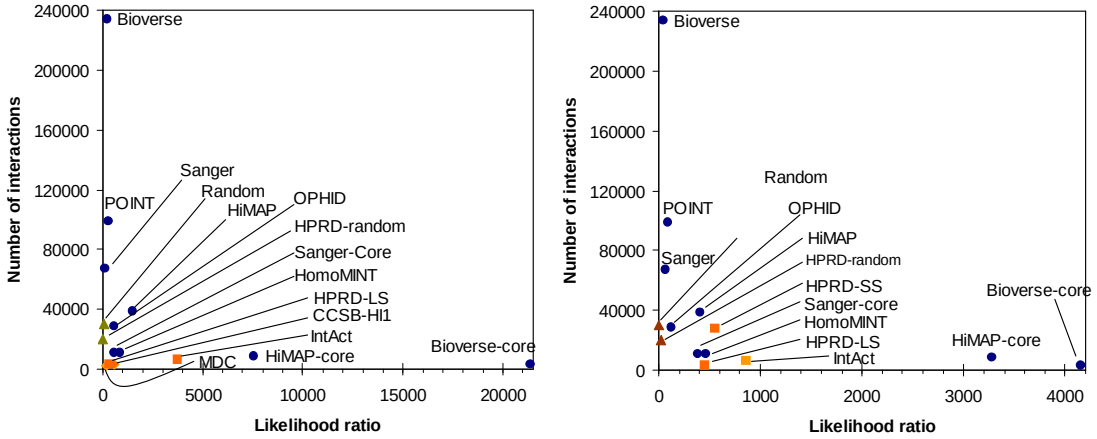


Fig. 3.7: Plots of the likelihood ratio (LR) vs. the number of interactions: (A) LR estimated using HPRD-SS as PRS set, and (B) LR estimated using the combined Y2H datasets as PRS set.

be seen in Table 3.6, the rankings of the datasets by LR using either HPRD-SS or the combined Y2H datasets are similar to each other and also to the rankings based on the other applied quality measures. Like with the other rankings, Bioverse-core, HiMAP-core, and HiMAP rank top and Sanger, POINT, CCSB-HI1, MDC, and the randomized datasets rank at the bottom. The overlaps of human PPI datasets with PRS sets exhibit both a high average *BPscore* and an elevated number of DDI-validated PPIs when compared to the overall values of each dataset (Table 3.5 and Table 3.6). Remarkably, these overlap values are higher than the corresponding values for the PRS sets of HPRD-SS and the combined Y2H datasets. This means that the PPIs in the overlaps seem to constitute high-quality subsets of the respective PRS sets. In contrast, PPIs in the NRS overlaps have a lower average *BPscore* and fewer DDI-verified PPIs than in the PRS overlaps.

Bioverse-core and HiMAP-core have outstanding LR based on both HPRD-SS and the combined Y2H datasets as PRS sets although they predict only 3,266 and 8,832 respectively (Table 3.6 and Fig. 3.7). In contrast, the complete Bioverse dataset ranks drastically lower. Bioverse has an above-average *BPscore* (0.575) and a high number of DDI-validated interactions (43.5%) with respect to the other datasets, but it is ranked by LR near the bottom due to an exceptionally high FPR. This may be caused by the primarily aim of Bioverse which is the generation of a huge network containing many proteins and interactions to derive functional predictions(McDermott et al., 2005). Nevertheless, the results show that the overlap of Bioverse and the NRS set has an average *BPscore* of 0.463 and 33.5% of DDI-validated PPIs. Therefore, Bioverse predicts DDI-validated protein interactions with high functional similarity, but many of those interactions may

Dataset	Number of interactions	Average *BPscore*	DDI-validated PPIs in percent	LR	PRS set overlap (%)	NRS set overlap (%)
IntAct (PA)	109	0.811	28.99	—	4 (3.67)	0 (0.00)
IntAct (X-ray)	160	0.777	85.29	11702.9	7 (4.38)	2 (1.25)
HPRD (3 pub.)	861	0.668	41.44	3343.7	16 (1.86)	16 (1.86)
HPRD ($\geq$ 4 pub.)	497	0.668	48.31	1308.4	9 (1.81)	23 (4.63)
IntAct ($\geq$ 4 pub.)	93	0.666	41.03	8359.2	5 (5.38)	2 (2.15)
IntAct (other)	605	0.660	31.88	3600.9	14 (2.31)	13 (2.15)
HPRD (2 pub.)	3048	0.654	33.10	1382.1	31 (1.02)	75 (2.46)
IntAct (Co-IP)	920	0.647	23.36	1800.4	14 (1.52)	26 (2.83)
IntAct (3 pub.)	192	0.637	20.25	6130.1	11 (5.73)	6 (3.13)
IntAct (2 pub.)	582	0.637	27.72	1671.8	7 (1.20)	14 (2.41)
HPRD (*in vivo*)	17417	0.611	22.84	462.2	98 (0.56)	709 (4.07)
HPRD (*in vitro*)	19616	0.603	25.63	649.7	130 (0.66)	669 (3.41)
HPRD (1 pub.)	26550	0.584	21.23	418.7	144 (0.54)	1150 (4.33)
HPRD (Y2H)	7964	0.554	17.17	1137.9	146 (1.83)	429 (5.39)
IntAct (1 pub.)	4942	0.515	13.64	583.6	37 (0.75)	212 (4.29)
IntAct (Y2H)	2289	0.493	14.69	1005.6	40 (1.75)	133 (5.81)
IntAct (TAP)	1993	0.457	5.25	45.2	1 (0.05)	74 (3.71)

Table 3.7: Comparison of HPRD and IntAct by number of publications and experimental technique. A subset of protein interactions reported in exactly n publications is denoted by 'n pub.' HPRD classifies experiments into three categories: *in vivo*, *in vitro*, and *yeast two-hybrid*. The IntAct classification of experimental techniques is based on a controlled vocabulary. Here, we regard only the most common experimental techniques frequently found in IntAct: yeast two-hybrid (Y2H), tandem affinity purification (TAP), co-immunoprecipitation (Co-IP), X-ray crystallography (X-ray), and protein array (PA). All other techniques are labeled 'other'. The datasets are ordered by decreasing *BPscore*.

be false, which suggests that Bioverse overpredicts considerably. This finding also supports the approach introduced in this study of comparing human interaction datasets with distinct quality measures in order to identify unfavorable biases of different kind in the datasets; the use of functional similarity or DDI validation alone for the estimation of data quality could lead to misleading results.

When benchmarked against HPRD-SS, the experimental Y2H datasets CCSB-HI1 and MDC as well as HPRD-LS rank in the lower half of the list of all datasets due to a low number of PRS matches and a high number of NRS matches. This is in agreement with the rankings obtained by the preceding assessments. Moreover, the *BPscore*s and the DDI-validation of the Y2H screens (Table 3.5) are similar to those of the Y2H interactions in HPRD and IntAct (Table 3.7). Notably, the CCSB-HI1 and the MDC interaction sets have similar FPR, but CCSB-HI1 has about four times as many PPIs in the PRS set derived from HPRD-SS than MDC. This reflects the trend found in the other quality assessments; the CCSB-HI1 dataset have scores that are better than the MDC dataset scores. Interestingly, when benchmarked against the combined Y2H datasets, the LR values of HPRD-SS (560.5) and IntAct (857.4) are closer to the LR values of some of the predicted datasets, particularly HomoMINT (460.6). In contrast, much higher values are obtained for the LRs of Bioverse-core (4,150.8) and HiMAP-core (3,286.0). This

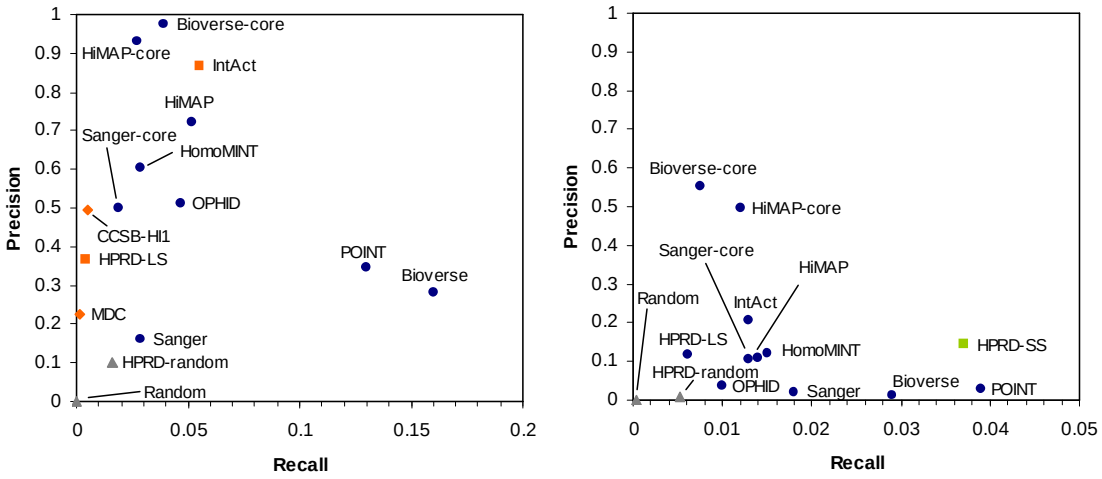


Fig. 3.8: Recall vs. precision plots using (let) HPRD or (right) the combined Y2H datasets as PRS set. While recall equals the computed true positive rate (TPR = $|E_i \cap$ PRS$|$ / $|$PRS$|$), precision is calculated by the following formula: $|E_i \cap \text{PRS}|$ / ($|E_i \cap$ PRS$| + |E_i \cap \text{NRS}|$).

middle rank for HPRD-SS and IntAct agrees with the rankings obtained by the other assessments described above. Since three different quality measures give similar results, this raises the question whether the literature-curated datasets contain a surprisingly large fraction of false positives (Rual et al., 2005). A more detailed analysis of those datasets indicates (Table 3.7) that protein interactions supported by two or more publications achieve higher LR (above 1,300) mainly due to a larger fraction of interactions that overlaps with the combined Y2H datasets used as PRS. Interactions derived from X-ray crystallography, as annotated in IntAct, have one of the highest LR (11,702) as well as a high average *BP-score* (0.777) and a large number of DDI-validated interactions (85.29%). Those values are comparable to the respective values of the top-ranking core datasets in Table 3.5.

3.5.4 Recall and precision analysis

Precision values obtained by using the combined Y2H datasets CCSB-HI1 and MDC as PRS appear much lower than the corresponding values obtained using HPRD-SS as PRS, which is probably due to a considerable rate of false positives in the Y2H screens (Fig. 3.8). In the recall vs. precision plot using HPRD-SS as PRS, both HiMAP and OPHID datasets have much higher precision and recall than datasets adjacent to them using the combined Y2H datasets as PRS. In the case of HiMAP, it may be biased towards HPRD-SS because a previous release of HPRD was originally used to evaluate the predicted PPIs of HiMAP. Remarkably,

Dataset	Average number of neighbors	Maximum number of neighbors	γ of degree distribution	Network diameter	Average shortest path length	Average clustering coefficient	γ of clustering coefficient distribution
Bioverse	60.24	842	-1.1887	10	3.5035	0.4801	-0.1845
Bioverse-core	4.67	34	-1.7635	17	6.3159	0.5029	0.2160
HiMAP	13.26	159	-1.7441	18	5.1591	0.4401	-0.0965
HiMAP-core	6.09	44	-1.7982	26	9.3950	0.3156	0.1253
HomoMINT	4.95	68	-2.0799	12	4.9153	0.0650	-0.4486
OPHID	12.39	192	-1.4260	18	4.5375	0.1885	0.0904
POINT	16.26	522	-1.6927	10	3.5284	0.0889	-0.3508
Sanger	22.69	365	-1.4090	10	3.8715	0.2342	0.0039
Sanger-core	5.87	75	-1.8402	20	6.4511	0.1861	0.3704
CCSB-HI1	3.43	129	-1.5637	12	4.3581	0.0626	-0.7932
MDC	3.58	95	-1.5149	12	4.6248	0.0205	-0.8197
HPRD-LS	3.13	213	-1.3000	10	4.4327	0.0602	-1.1700
HPRD-SS	6.78	202	-1.8420	15	4.4627	0.1276	-0.4830
IntAct	3.83	181	-1.4450	18	5.1542	0.1022	-0.7890
Random	11.99	30	—	6	3.6986	0.0026	—

Table 3.8: Topological network parameters for each human protein interaction dataset. The degree and clustering coefficient distributions are fitted to power laws with exponents γ.

the precision of the predicted datasets Bioverse-core and HiMAP-core is larger than that of manually curated dataset HPRD-SS in the recall vs. precision plot using the combined Y2H datasets as PRS, and the precision of the predicted datasets Sanger-core, HiMAP, and HomoMINT are very close to the precision of HPRD-SS. These interesting results suggest that predicted PPIs can be quite reliable. The precisions of the two Y2H datasets using HPRD-SS as PRS are significantly different (CCSB-HI1 with 0.389 versus MDC with 0.145).

3.5.5 Topological network analysis

Not only quality measures as detailed above, but also topological parameters can be used to uncover potential bias in the networks formed by the different PPI datasets. The degree distribution, diameter, average shortest path length, and clustering coefficient (Shannon et al., 2003; Barabási & Oltvai, 2004) were computed for all datasets (Table 3.8 and Fig. 3.9) except for HPRD-random, which has the same topology as the complete HPRD (HPRD-SS and HPRD-LS) by definition. The degree of a protein is defined as the number of its interaction partners. All networks (except for the random network) fit a scale-free degree distribution (Barabasi & Albert, 1999), which means that the probability $P(k)$ of proteins with k interactions decays as a power law: $P(k)\ k^{\lambda}$. In the analysis, the value of the exponent λ lies between -1.1887 and -2.0799 for all (non-randomized) networks. However, the average number of neighbors varies considerably between the datasets from 3.13 for HPRD-LS to 60.24 for Bioverse. The average clustering

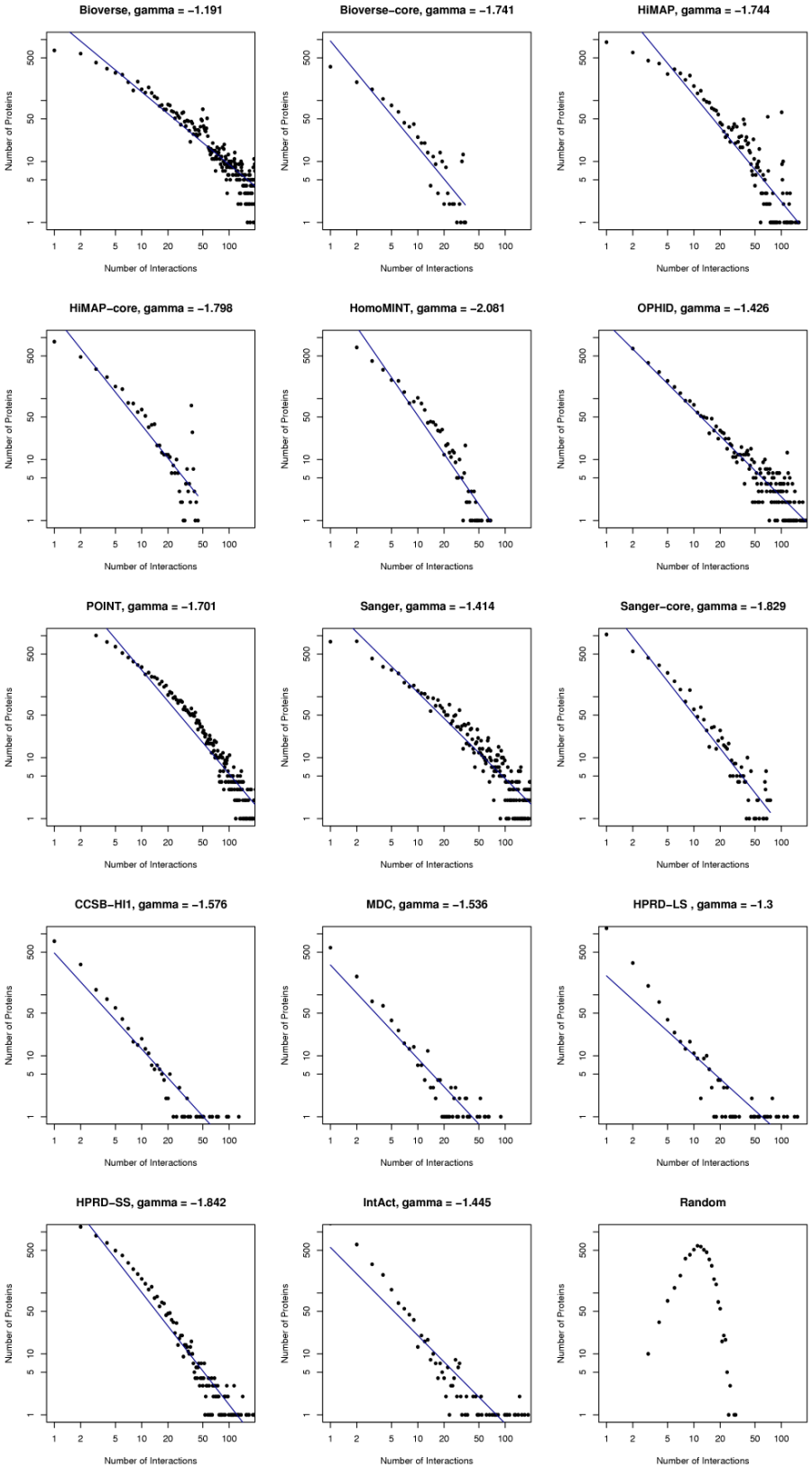


Fig. 3.9: Degree distributions together with the exponent λ of the fitted power law.

coefficient, a measure of interaction density, also exhibits significant differences ranging from 0.0205 of MDC to 0.5029 of Bioverse-core.

To highlight the numeric differences the degree distributions are depicted not only as plots of the number of interactions versus the number of proteins, but also as 2D histograms, in which each axis represents the number of neighbors for one of the two interacting proteins (Fig. 3.10). The 2D histograms particularly allow for the visual identification of further bias in the datasets. The HPRD-SS histogram displays a high density of interactions between proteins with 1 to 10 neighbors, and proteins with a large numbers of neighbors tend to interact preferentially with proteins with few neighbors as expected from a scale-free network topology. In detail, the CCSB-HI1, HomoMINT, HPRD-LS, HPRD-SS, IntAct, and MDC display similar histograms (see Fig. 3.9 and Fig. 3.10). The HiMAP histogram also resembles the HPRD-SS histogram except for the fact that interactions between proteins of high degree are surprisingly frequent. For example, a cluster of 64 fully connected proteins was identified, each protein having 101 interaction partners for a total of 2,016 interactions). All these proteins are members of the potassium channel family and their amino acid sequences are closely related.

In general, the 2D histograms of the predicted datasets Bioverse-core, HiMAP-core, HiMAP, OPHID, POINT, Sanger, and Sanger-core show an unexpected abundance of interacting proteins with high degrees in contrast to the experimental datasets CCSB-HI1, HPRD-LS, HPRD-SS, IntAct, and MDC. Hence, it appears that, even though the networks of the predicted datasets are generally scale-free, they contain some unfavorable bias towards interacting proteins with numerous neighbors.

The frequencies of interactions in terms of protein length was analyzed in addition to the computation of network parameters, (Fig. 3.11). However, no particular bias in the datasets towards certain protein lengths was found. The only exception is the HiMAP dataset that is inexplicably enriched with proteins whose lengths exceed 750 amino acids.

3.5.6 Shared neighbors

As part of the topological analysis a new a graph measure called shared-neighbors (*SN*) was defined and integrated as part of the *NetworkAnalyzer* (Assenov et al., 2008) plug-in for Cytoscape. The *SN* measure represents the count of the number of directly connected nodes in common between any two nodes in a network. Formally, for a network G with node set V and edge set $E \subseteq V \times V$, the shared

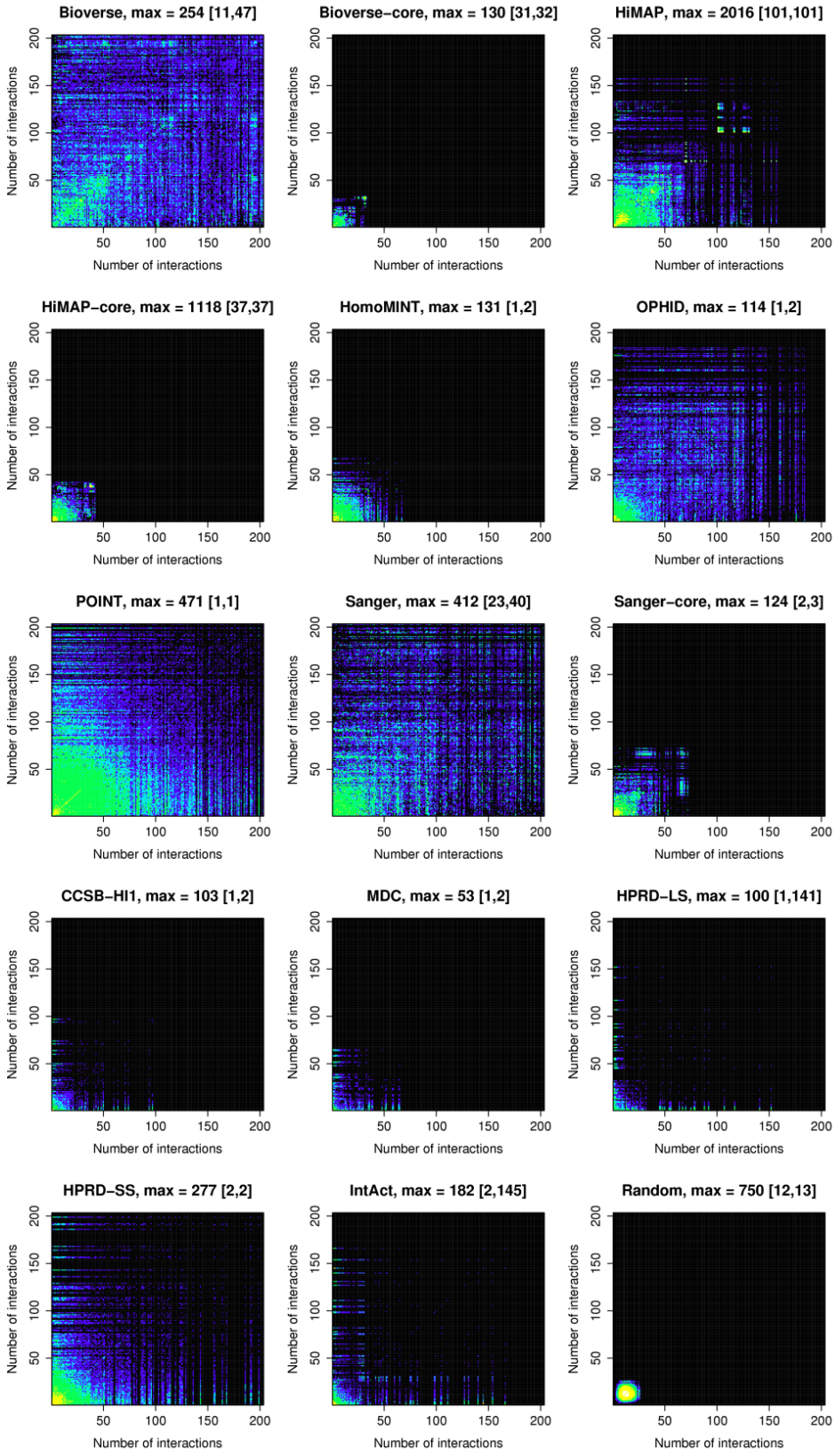


Fig. 3.10: 2-D histograms of degree frequency. The dot colors in the histograms reflect the number of occurrences of two interacting proteins with specific degrees (number of interactions). The maximum number N of occurrences in each dataset is given above the histograms; the notation max=N [X,Y] refers to two interacting proteins of degrees X and Y.

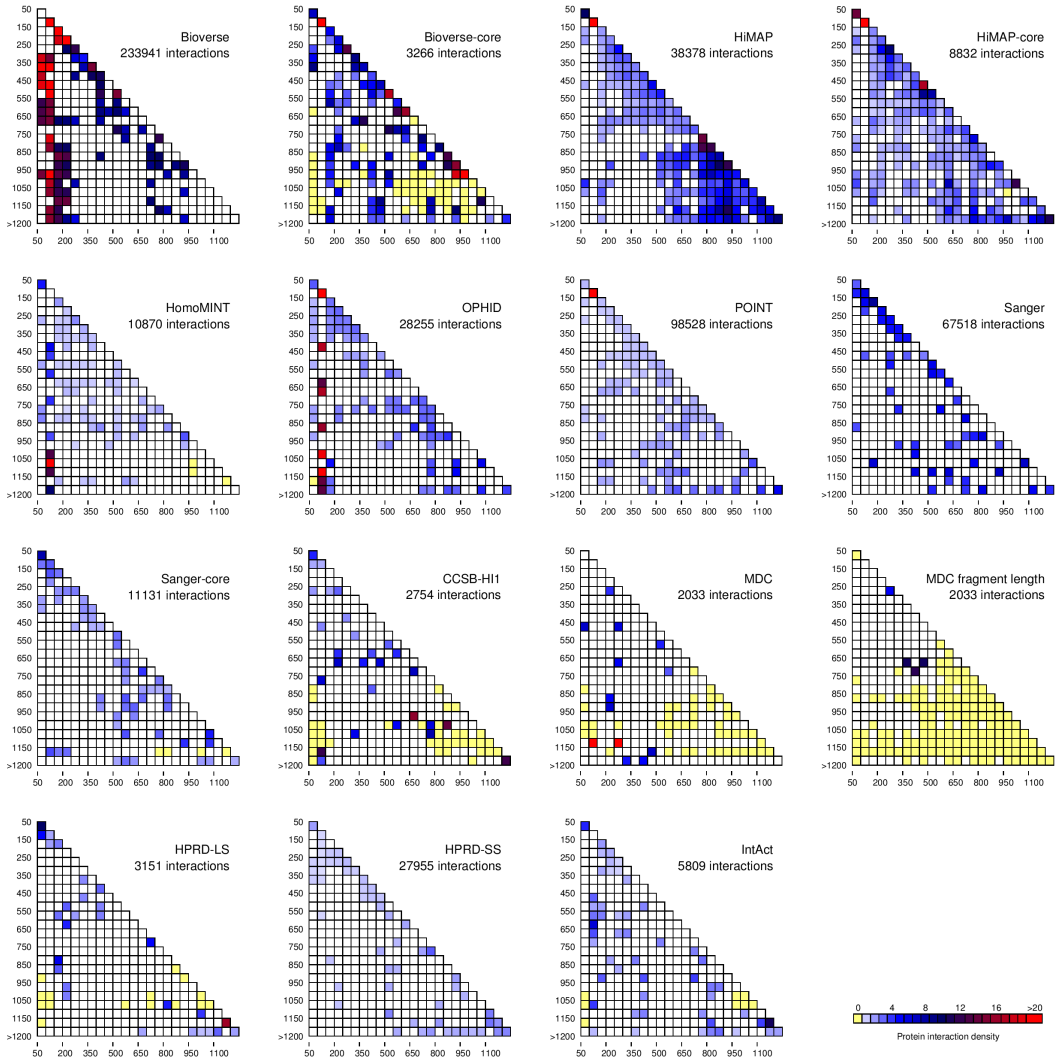


Fig. 3.11: 2D histograms of the distribution of PPIs according to the length of interacting proteins binned in steps of 50 amino acids. Each human interaction dataset is depicted by a triangle matrix whose axes represent the sequence lengths of interacting proteins. The dot color in the histograms reflects the protein interaction density that is the ratio of the number of PPIs assigned to the respective matrix cell divided by the total number of PPIs possibly formed; the total number of possible PPIs was derived by counting the members of the respective protein length bin, and the density was normalized to 1,000 possible PPIs. The protein interaction density is not shown if the observed number of PPIs is non-significant (p-value ≥ 0.01, using Fisher's exact test as in case of the overlap computation). The matrix entitled 'MDC fragment length' was derived using the actual lengths of the protein fragments as used in the Y2H screen in contrast to the matrix 'MDC' whose proteins lengths belong to complete protein sequences as in case of all other datasets.

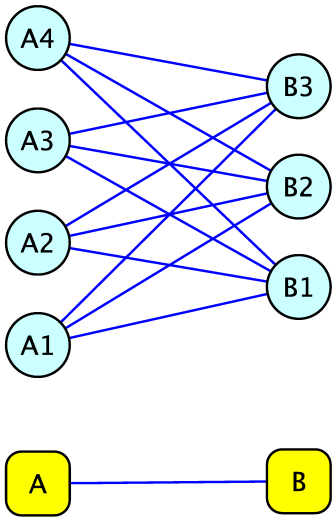


Fig. 3.12: Interolog mapping when multiple homologs of the original protein exist. The nodes A and B represent proteins that are known to interact in one species (e.g. *Caenorhabditis elegans*). The nodes A1-A4 are homologs of A, and the nodes B1-B3 homologs of B in other species (e.g. *Homo sapiens*). If all possible combinations of interacting homologs are considered, any two homologs of A will share all homologs of B as interacting neighbors and vice versa.

neighbors $SN(v, w)$ for two nodes $v, w \in V$ is given by

$$SN(v, w) = |\{i \in V | (i, v) \in E \ \ \& \ \ (i, w) \in E\}|$$

SN can also be described as the number of paths of length two that connect a pair of nodes.

In a molecular evolution context, pairs of proteins that share several neighbors may arise by gene duplication events (Wagner, 2001; Barabási & Oltvai, 2004). Before further divergence of the duplicated gene product, it would interact with the same neighbors of the original gene product. After divergence, only some of the neighbors may still be shared by both the duplicate and the original gene products. Using the SN measure together with sequence identity can help to uncover such evolutionary events. The measure can also help to detect funtionally related proteins because it is known that they tend to share common interaction partners.

In this analysis, the SN measure was primarily used to detect possible bias in the topology of predicted protein interaction networks derived by the interolog mapping method. This mapping method might result in an erroneously predicted combinatorial expansion of new interactions if several homologs are found for each of the proteins involved in the original interaction (Fig. 3.12). For example, if

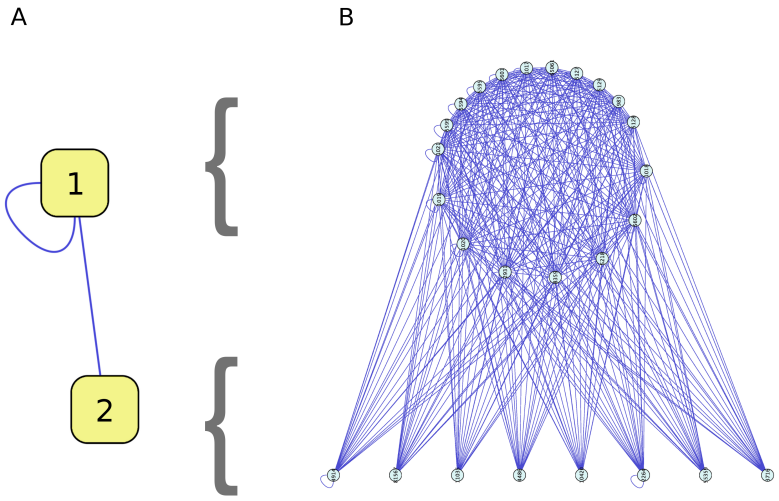


Fig. 3.13: Combinatorial expansion due to interologs mapping. Two original interactions (A) were mapped in Bioverse to 317 PPIs between 27 proteins (B). In (B), the circle on top consists of cyclin-dependent kinases derived by (1), while the proteins on the bottom are tyrosine kinases derived from protein (2). This dense cluster proteins was revealed by the shared neighbors analysis.

some original interaction exists between proteins A and B with 4 and 3 homologs, respectively, 12 new interactions could be derived (Fig. 3.2). Notably, any two proteins homologous to A would share all proteins homologous to B and vice versa.

A special case occurs if the original interaction is a homodimer represented as self-loop. The interolog mapping of this self-loop may result in many interactions between all homologs because any pair of interacting proteins will share all homologs as interacting neighbors (Fig. 3.13). The detection of the shared neighbors for every pair of nodes in a predicted network would indicate whether a serious problem with the network exists. Another approach to detect the same problem would be the determination of all paralogous proteins in the network and the subsequent check if they share the same neighbors. Although this alternative method is possibly more precise, it would require additional information not contained in the network itself. In this context, it is of interest that the Bioverse dataset has been derived by allowing the mapping of even low-similarity proteins with the aim of providing a very large dataset for further functional analyses (McDermott et al., 2005). Accordingly, Bioverse overpredicts many PPIs and contains a large number of false positives as reflected by Fig. 3.10 and revealed as well by the quality assessment using the likelihood ratio.

The frequencies of shared neighbors are plotted in Fig. 3.14. In case of the largest dataset Bioverse, there are pairs of proteins that share hundreds of neigh-

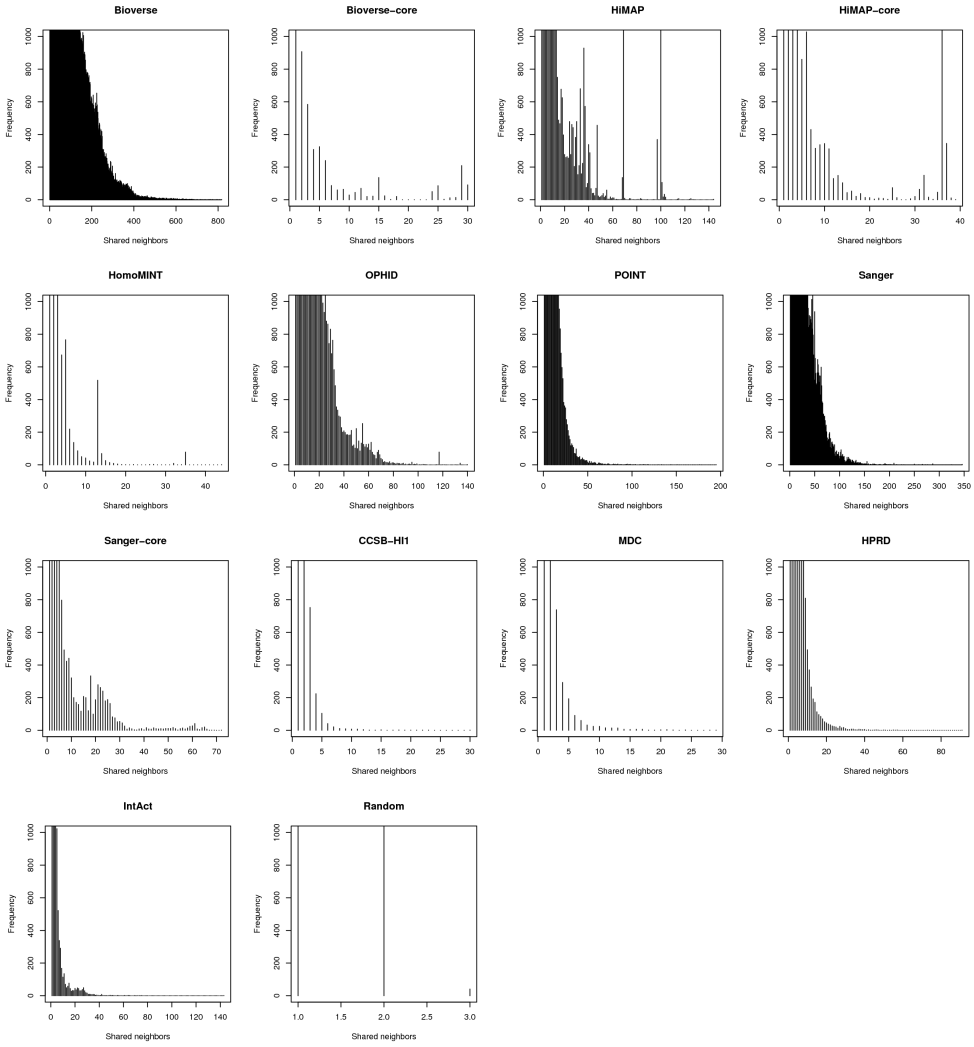


Fig. 3.14: Frequency distribution of shared neighbors numbers. The histograms highlight some biases in the predicted datasets which tend to have elevated frequencies for certain numbers shared neighbors. For instance, in the case of HiMAP there are two bars having a frequency over 1,000 that correspond to 69 and 97 shared neighbors.

bors (up to 817 neighbors) and thousands of pairs that share between 1 to 150 neighbors. Thus, Bioverse tends to overpredict due to low-similarity interolog mappings as the one exemplified in Fig. 3.13. In comparison pairs of proteins sharing more than 50 neighbors are rare for the manually curated datasets HPRD and IntAct and for the Y2H datasets, CCSB-HI1 and MDC, where a pair of proteins rarely shares more than 20 neighbors.

Apart from Bioverse, other predicted datasets are found to contain large amounts of shared proteins compared to the experimental datasets (Fig. 3.15).

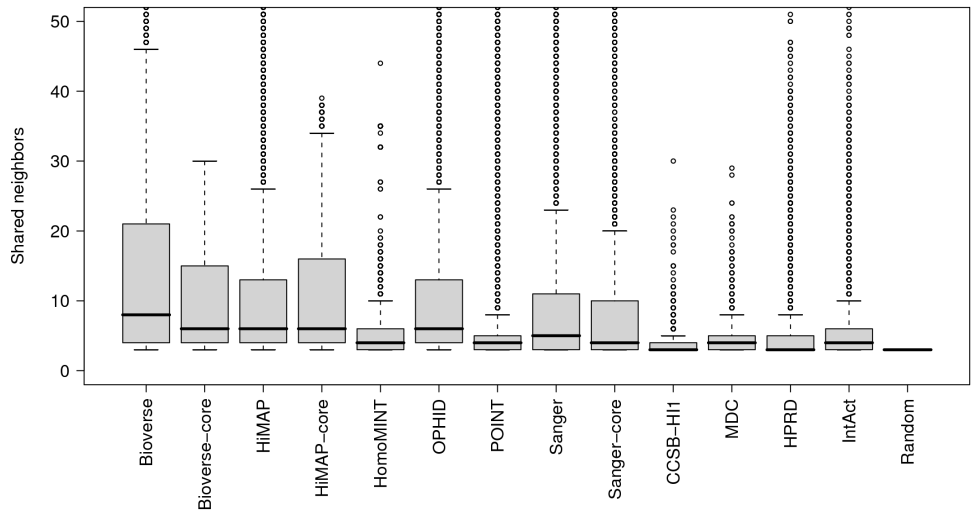


Fig. 3.15: Distributions of the number of shared neighbors in the different datasets. Only numbers of shared neighbors larger than two were included.

Large average neighborhood sizes also occur in the core datasets, but the maximum values are comparable to those in the Y2H datasets. Interestingly, HomoMINT and POINT are the datasets that most closely resemble the experimental data.

3.5.7 Comparison with manually curated datasets

To analyze the reliability of the protein interactions in the literature-curated datasets further, the interactions in HPRD and IntAct were subdivided by the number of publications reporting them and by the experimental technique (Table 3.7). The *in vivo* and *in vitro* classifications of HPRD obtain similar scores in all assessments, indicating that their reliabilities are similar. It is also apparent that the more publications support a protein interaction, the higher are its scores. Interestingly, PPIs derived from protein arrays have the highest *BPscore* (0.811) and do not overlap with the NRS. However, this could be misleading because all those PPIs come from the same publication. Moreover, as expected, X-ray crystallography returns a very high number of DDI-validated PPIs (85.29%). In contrast, the protein interactions derived from tandem affinity purification (TAP) have the lowest overlap with the combined Y2H datasets (only 1 interaction), and the number of DDI-validated interactions is the smallest (5.25%). Furthermore, the Y2H interactions contained in HPRD and IntAct have a *BPscore* and fraction of DDI-validated interactions similar to that of the Y2H dataset CCSB-HI1.

Datasets such as HPRD *in vivo* and *in vitro* listed in Table 3.7 have higher *BP-score* and number of DDI-validated interactions, but lower LR (using the combined CCSB-HI1 and MDC datasets as PRS) than those Y2H interactions in HPRD and IntAct. This might be explained by the idea that Y2H screens can detect interactions not found by other methods.

3.5.8 Predicted interactions based on high-throughput data

PPIs in predicted human datasets are primarily derived from interologs using high-throughput data (Table 3.2). For instance, the DIP database, used by Bioverse and POINT, contains 80% of PPIs detected by high-throughput experiments. A similar portion is contained in MINT causing that only 6% of the PPIs in HomoMINT are derived from small-scale experiments. The assessments also show that predicted datasets such as Sanger derived solely from high-throughput experiments perform similar to Y2H screens. Other predictions such as HomoMINT, OPHID, and POINT that utilized many high-throughput interologs and relatively few from small-scale experiments score only slightly better. The Sanger-core dataset, which is based on interologs reported in more than one publication, achieves higher assessment scores than the Sanger dataset. However, the Sanger-core values of *BPscore*, DDI-validation, and LR assessment are still similar to those of HomoMINT, OPHID, and POINT. The HiMAP datasets, which do not only rely on interologs, achieve better performance are revealed by the assessment results. The outstanding scores of Bioverse-core may be due to the inclusion of PPIs from X-ray crystallography and, in contrast to Bioverse, due to the application of a stringent sequence similarity threshold for establishing orthology. Therefore, the results reported here suggest that predictions based on interolog mapping can be as good as the original data used to derive them and even better if appropriate filters and methods are additionally employed.

3.6 Summary

The quality of several comprehensive human protein interaction datasets was compared and assessed based on different criteria. This quantitative and qualitative analysis included six predicted datasets and three high-confidence core subsets, two literature-curated datasets, and two high-throughput Y2H datasets. This analysis was based on the functional similarity of interacting proteins, the validation of PPIs using structurally known protein domain interactions, and the

evaluation of the contents of the datasets against positive and negative reference sets of PPIs. Also, the different interaction sets were ranked based on the scores assigned by the applied quality measures. Finally, the interaction network topologies resulting from the datasets was investigated.

In summary, the findings reported here indicate that the datasets Bioverse-core and HiMAP-core of predicted PPIs contain high-quality data in comparison to other predicted datasets. The proposed assessments also support the view that many PPIs of predicted datasets appear to be at least as reliable as the results of Y2H screens. Therefore, it is useful to combine predicted and experimental datasets to increase the coverage of the human interactome. This is particularly important if one wants to use human networks in the context of diseases. However, it is important to keep in mind that each dataset, even the literature-curated datasets, (i) may have been optimized for one or the other quality measure, (ii) contains a significant amount of low-quality data, and (iii) seems to be biased towards certain biological functions. For instance, DDI-validation may be biased because Bioverse utilized structural information from PDB. Also, HiMAP inferred PPIs from predicted DDIs and used GO-based functional similarity as well as an older version of HPRD (August 2004, 17,462 interactions).

Presumably, the observed diversity of the datasets is mainly due to the distinct data sources used for the predictions and details of the interolog mapping procedure. Nevertheless, the detected low overlap of datasets means that the datasets can complement each other well. To this end, it would be helpful that confidence values are assigned to PPIs and that interaction datasets are carefully described with methodological details and the original data sources. Moreover, it was found that the examination of network topologies aids in the identification of further bias and artifacts produced by the prediction methods. However, the discovered topological diversity also advises caution against possible misinterpretations when using topological parameters derived from the current human datasets in biological applications.

4

Analysis of Human Scaffold Proteins

This chapter describes the use of the integrative data warehouse for the computational identification of scaffold proteins (Ramírez & Albrecht, 2010), an important type of gene products involved in signaling cascades whose function is the organization of higher order complexes through protein–protein interactions. The published results represent the first attempt to estimate the number and abundance of these proteins in the human proteome (Alexa et al., 2010). Before this study, scaffold proteins were only identified fortuitously through direct experimentation. The identification of scaffold proteins was possible thanks to the cross linking of protein function, biological processes and protein-protein interactions from different public sources. This approach proves the advantages of local data storage to cross link diverse sources of information.

4.1 Introduction

Signaling cascades determine how the cells respond to changes in their external and internal environment. A multitude of signaling proteins with a broad substrate specificity are in charge of mediating the signaling process. Thus, it is of immense importance to understand how the cell achieves efficiency and accuracy in signaling (Buday & Tompa, 2010). Scaffold proteins are a recently described category of molecules influencing cellular signaling first described in yeast by Choi et al. (1994). Scaffold proteins bind to multiple enzyme or receptor proteins although they itself are devoid of any catalytic activity. Their main role is to colocalize several members of a catalytic pathway to specific areas of the cell (Shaw & Filbert, 2009) and to permit a better fine tuning of regulatory processes through the coordination of positive and negative feedbacks. Because scaffold proteins separate catalysis from molecular recognition, catalytic molecules can

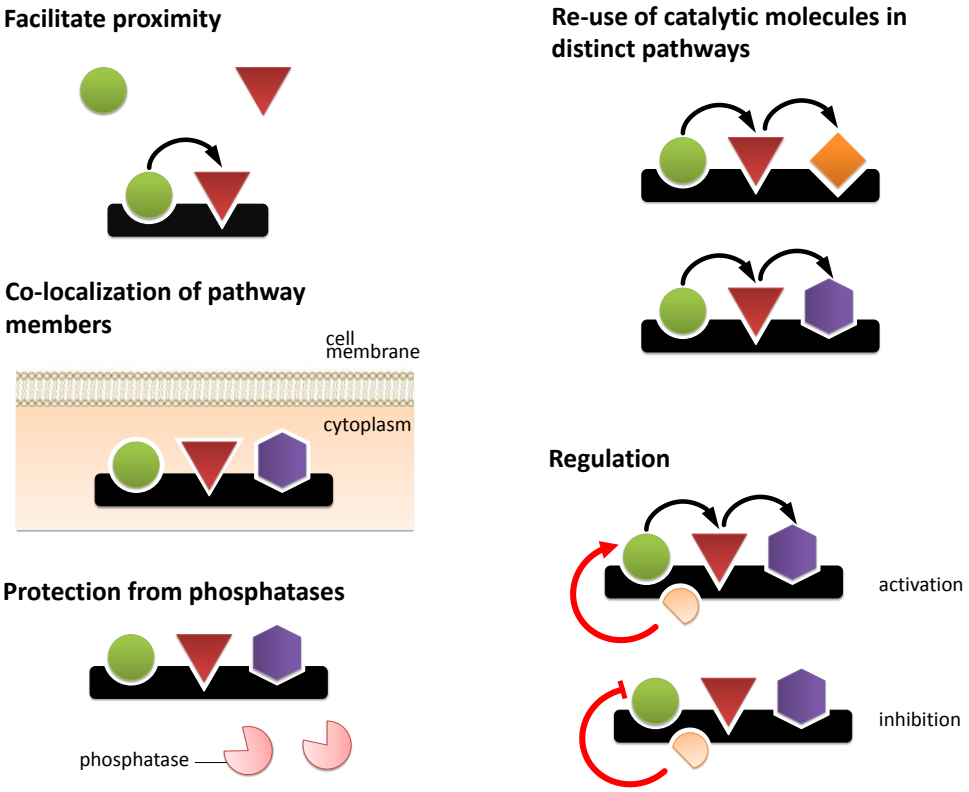


Fig. 4.1: Scaffold proteins bind to multiple signaling molecules simultaneously. They increase the efficacy of a signaling pathway and help to localize signalling molecules to a specific cell compartment. Scaffold proteins also regulate signal transduction by coordinating positive and negative feedback signals, and insulate binding partners from competing proteins. Image inspired by Shaw & Filbert (2009); Zeke et al. (2009).

be re-used in distinct pathways by coupling with different scaffold proteins while avoiding cross-signaling (Bhattacharyya et al., 2006). Locasale et al. (2007) have also indicated that scaffold proteins protect binding signalling molecules from inactivation by phosphatases or from degradation (Fig.4.1).

Although scaffold proteins are considered fundamental to signaling process, their identification by sequence similarity methods have proven to be difficult they appear not to be evolutionarily related. Instead, it is supposed that scaffold proteins have originated independently several times during the evolution of signalling systems and, as consequence, share little sequence identity (Zeke et al., 2009). For this reason, scaffold proteins are mainly discovered fortuitously while studying the function of well-known signaling proteins. To faciliate the discovery of scaffold proteins Zeke et al. (2009) proposed an interaction-based definition to systematically identify potential scaffold candidates in interactomes. This definition is based on three common properties of scaffold proteins: (*i*) lack of intrinsic

catalytic activity relevant for signaling, (*ii*) direct interaction with at least two signaling proteins possessing *catalytic or receptor activity* (referenced here as CRPs), and (*iii*) direct or indirect interaction of at least two CRPs with each other. Because some scaffold proteins have properties (*ii*) and (*iii*) while having some catalytic activity, the term *classical* scaffold proteins is used for those cases having property (*i*) as well.

To explore this definition a comprehensive search for scaffold candidates using the data warehouse was implemented. As a result, a reliable set of 250 candidate scaffold proteins was identified. This proves that current public databases can be efficiently mined using data warehouses to extract useful information otherwise difficult to obtain, as in the case of scaffold proteins.

4.2 Computational identification of scaffold proteins

Using the Gene Ontology (GO) annotations as well as UniProtKB keywords, 3,185 human intracellular proteins involved with signal transduction were initially selected. From this list of signaling proteins, those fulfilling criterion (*i*) were further selected. Besides a lack of catalytic activity, the first criterion was extended to exclude proteins known to bind nucleic acids, related to translation, having receptor activity, being GTPase regulators or that are chaperons. Using this extended criteria (*i*) a list of 649 proteins, referred here as *special signaling proteins* (SSPs), was obtained.

Out of 9,814 proteins with at least two interacting proteins in the integrative data warehouse, 282 interact with at least two CRPs according to property (*ii*). Applying property (*iii*) reduces this number further to a final set of 250 scaffold candidates. A high-confidence subset of 130 candidates was obtained by only considering reliable SSP-CRP interactions reported in at least two scientific publications listed in PubMed. This threshold was based on the conclusions obtained in the previous chapter. Self-interactions were ignored and indirect interactions between two proteins were assumed to occur if both proteins bind another CRP.

Each scaffold candidate is found to interact on average with 9 CRPs and with 32 proteins overall. The high mean number of scaffold candidate binding partners and the limited number of binding sites available for any protein suggests that many scaffold candidates form various interaction platforms depending on cell type, tissue specificity, location and time (Bhattacharyya et al., 2006). Alternative

protein description	gene symbol	disease
cAMP-dependent protein kinase type I-alpha regulatory subunit	PRKAR1A	Intracardiac myxoma, primary pigmented nodular adrenocortical disease 1 (PPNAD1), papillary thyroid carcinoma, type 1 carney complex (CNC1)
Axin-1	AXIN1	Caudal duplication anomaly, hepatocellular carcinoma
SH3 domain-binding protein 2	**SH3BP2**	Cherubism
Keratin, type I cytoskeletal 18	KRT18	Familial cirrhosis
Alpha-synuclein	SNCA	Autosomal dominant lewy body parkinson disease 4 (PARK4), familial parkinson disease type 1(PARK1), lewy body dementia (DLB)
Suppressor of cytokine signaling 3	SOCS3	Atopic dermatitis 4 (ATOD4)
Insulin receptor substrate 1	IRS1	Noninsulin-dependent diabetes mellitus (NIDDM)
C-jun-amino-terminal kinase-interacting protein 1	**MAPK8IP1**	Noninsulin-dependent diabetes mellitus (NIDDM)
NF-kappa-B inhibitor alpha	NFKBIA	Ectodermal dysplasia with t-cell immunodeficiency
CD2-associated protein	CD2AP	Focal segmental glomerulosclerosis 3 (FSGS3)
Huntingtin	**HTT**	Huntington disease (HD)
Nucleotide-binding oligomerization domain-containing protein 2	NOD2	Blau syndrome, early-onset sarcoidosis, inflammatory bowel disease 1 (IBD1), susceptibility to psoriatic arthritis, susceptibility to sarcoidosis 1 (SS1)
Caveolin-1	**CAV1**	Congenital generalized lipodystrophy type 3 (CGL3)
Adenomatous polyposis coli protein	APC	Adenomatous polyposis of the colon (APC), colorectal cancer (CRC), gastric cancer, hereditary desmoid disease, medulloblastoma (MDB), mismatch repair cancer syndrome
14-3-3 protein epsilon	YWHAE	Miller-dieker lissencephaly syndrome (MDLS)
Na(+)/H(+) exchange regulatory cofactor NHE-RF1	SLC9A3R1	Hypophosphatemic nephrolithiasis/ osteoporosis 2
Nephrin	NPHS1	Congenital nephrosis finnish type 1 (NPHS1)
Sequestosome-1	SQSTM1	Paget disease of bone; pdb
Gap junction alpha-1 protein	GJA1	Atrioventricular septal defect (AVSD), autosomal recessive oculodentodigital dysplasia, Hallermann-Streiff syndrome (HSS), hypoplastic left heart syndrome, oculodentodigital dysplasia (ODDD), syndactyly type 3
Hamartin	TSC1	Focal cortical dysplasia of taylor (FCDT), lymphangioleiomyomatosis (LAM), tuberous sclerosis (TS)

Table 4.1: High confidence scaffold candidates associated with inherited diseases (OMIM). Proteins referred to scaffold or adaptor in the scientific literature are highlighted in bold. Those highlighted proteins were found during the posterior validation of the scaffold candidates.

splice variants and their interaction patterns may also play an important role and might have to be distinguished in the future to characterize scaffold candidate complexes further. Notably, 35 scaffold candidates (14%) of all identified 250 scaffold candidates are already known to be contained in protein complexes, and at least 184 scaffold candidates interact directly with each other, on average with 3 other scaffold candidates, pointing to the possible formation of larger supramolecular scaffold complexes (Fig. 4.3).

The obtained set of scaffold candidates is statistically enriched for specific GO-based molecular functions, biological processes, and cellular components (Fig. 4.2), some of which may be used to categorize scaffold candidates fur-

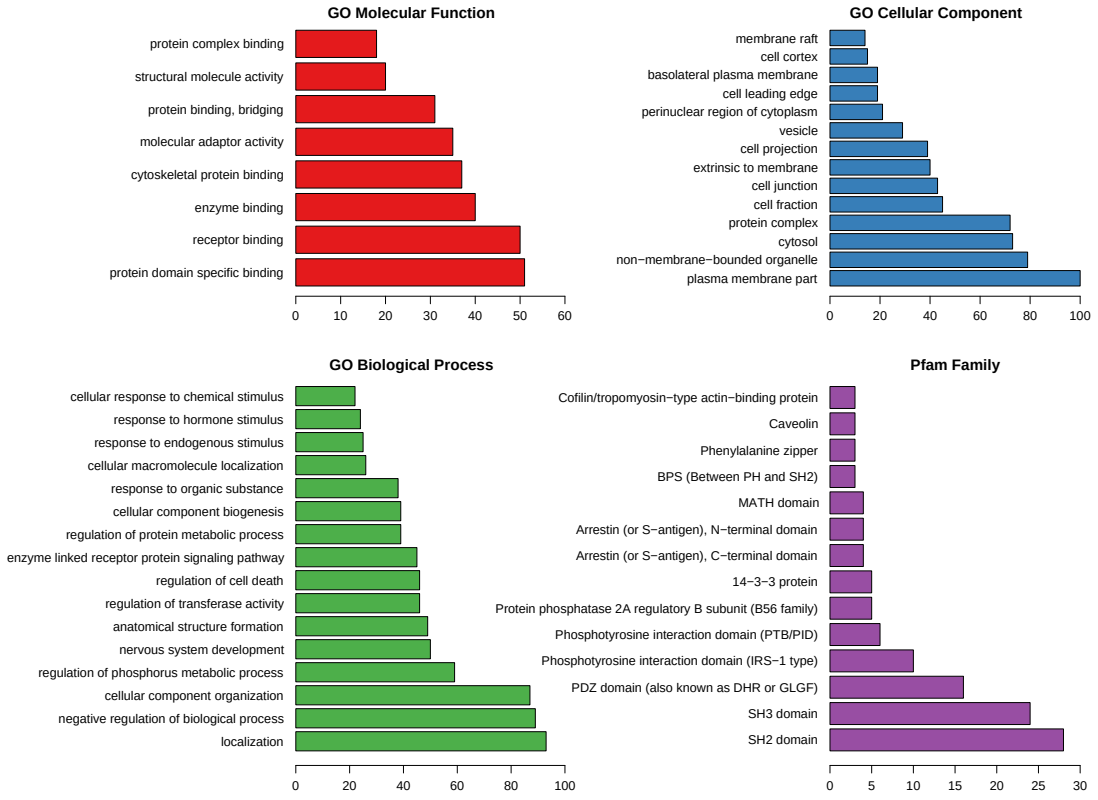


Fig. 4.2: Enrichment of Gene Ontology annotations and Pfam families. The four histograms show significantly enriched Gene Ontology annotations and Pfam domain families for the 250 potential scaffold proteins in comparison to all human signaling proteins found in this study (p-value < 0.001). The x-axis represents the absolute number of potential scaffold proteins belonging to the respective category.

ther. For instance, various metabolic and regulatory processes are significantly overrepresented in the scaffold candidate set compared to all signaling proteins. Interestingly, 30 scaffold candidates are annotated as having 'molecular adaptor activity', out of 52 human proteins having such annotation. Indeed, GO defines adaptor proteins as molecules having binding activity 'that brings together two or more molecules, permitting those molecules to function in a coordinated way' and may be regarded as a subset of scaffold proteins (Zeke et al., 2009). Further scaffold candidate characterization may also consider the protein domain composition because scaffold candidates and their interaction partners are frequently enriched with certain Pfam domain families like PH as well as PDZ and SH2/SH3 used for signal transduction (Fig. 4.2) (Schelhorn et al., 2008; Pawson & Nash, 2003). In table 4.1 20 high-confidence scaffold candidates that are associated with inherited diseases are listed. In seven cases the involved proteins were associated

scaffold protein	identified in the study?	reason	criteria not fulfilled
KSR1	no	annotated as catalytic	(i)
KSR2	no	annotated as catalytic	(i)
JIP1	yes, high-confidence		
JIP2	yes, high-confidence		
JIP3	yes, high-confidence		
JIP4	no	not annotated as signaling protein	(i)
β-arrestin 2	yes, high-confidence		
Paxillin	yes		
Gab1	yes, high-confidence		
Gab2	no	protein without annotations	(i)
Gab3	no	protein without annotations	(i)
PSD95	yes, high-confidence		
Homer1	yes, high-confidence		
Homer2	yes		
Homer3	no	catalytic binding partners not known to interact.	(iii)
mAKAP	no	not annotated as signaling protein	(i)
AKP79	yes, high-confidence		
RACK1	yes, high-confidence		

Table 4.2: Classical scaffold proteins used as validation set. The list, introduced by Zeke et al. (2009), contained the best known human scaffold proteins at the time of the publication. The column 'identified in study?' shows whether the known scaffold was found in the computational search. High-confidence refers to the list of 130 scaffold candidates found by using a more stringent criteria for reliable protein-protein interactions. Out of 18 known scaffold proteins, 10 were found in our study, eight of those in the high-confidence set.

with more than one disease. The gene product GJA1, for example, is associated to six different conditions.

4.3 Validation of scaffold candidates

The computational results were compared with a list of 18 human well studied classical scaffold proteins published by (Zeke et al., 2009). The interactome search found 10 of the 18 examples, eight of them in the high-confidence set. The remaining eight proteins are missing for different reasons: Homer3 does not have property (iii) which may be just due to insufficient interaction data in the curated databases; the other seven scaffold proteins could not be found in the search because the required GO annotation or UniProtKB keyword was not present in case of property (i). This is not surprising because the coverage of protein annotations is not complete (Fig. 5.4**A**). In detail, KSR1/2 are annotated with a

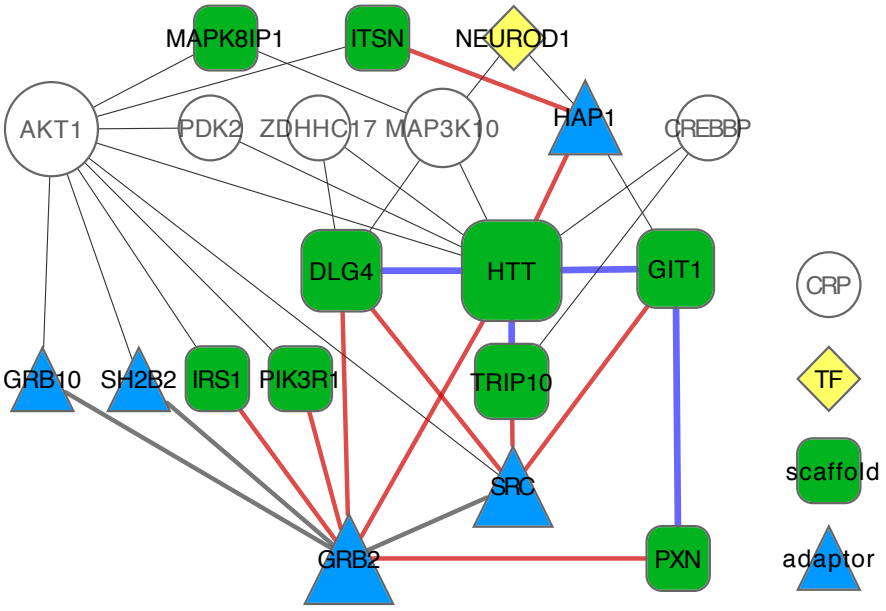


Fig. 4.3: Huntingtin (Htt) interaction network with scaffold and adaptor proteins. Htt binds to a number of scaffold and adaptor proteins forming supramolecular scaffold complexes.

debatable kinase activity (Kolesnick & Xing, 2004), JIP4 and RACK1 are not annotated as being involved in signaling transduction, mAKAP is involved in protein targeting, but not in signal transduction according to GO, and GAB2 as well as GAB3 are not annotated at all (Table 4.2). All of the validated scaffold candidates, except of GAB1 are annotated as membrane proteins, which agrees well with the fact that 133 scaffold candidates (53%) are located to membranes according to GO.

PubMed was searched for published abstracts containing the scaffold candidate name, symbol or some synonym together with the words "scaffold*" or "adaptor". For 208 scaffold candidates (83%), at least one matching abstract could be retrieved. Manual inspection of the abstracts for 50 of those candidates (20%) confirmed that many of them were referred to as scaffold or adaptor proteins in the literature for over 70% of the studied cases (Table 4.3 and supplementary Table A).

symbol	type	pubmed id	text snippet
AKAP12	scaffold	17686059	A-kinase anchor protein 12 (**AKAP12**) is a **scaffold** protein that participates in mitotic regulation and others signalling processes and probably exerts tumour suppressor function.
	scaffold	16442664	A-kinase anchoring proteins (**AKAPs**) define an expanding group of **scaffold** proteins that display a signature binding site for the RI/RII subunit of protein kinase A.
	scaffold	17626016	SSeCKS (Src-suppressed C kinase substrate), also called gravin/**AKAP12**, is a large **scaffolding** protein with metastasis suppressor activity.
		and others	
EFS (SYN)	scaffold	11867627	Instead, the pathway involved relies on increased tyrosine phosphorylation of, and recruitment of Crk to, the SRC substrate **SIN**/**EFS**. The latter is a **scaffolding** protein structurally similar to the SRC substrate Cas
	scaffold	18256281	For over a decade, p130Cas/BCAR1, HEF1/NEDD9/Cas-L, and **EFS**/**SIN** have defined the Cas (Crk-associated substrate) **scaffolding** protein family.
HTT	scaffold	12881483	We propose that **HTT**, together with HAP1, may function as a **scaffold** for the activation of ND by MLK2.
	scaffold	19269181	Overall, the predicted structure of **huntingtin** is consistent with a cellular role as a **scaffold** protein.
	scaffold	19429504	Unexpectedly, the faulty gene product, mutant **huntingtin** (mtHtt), is an extremely large protein of 350 kDa and might act as a **scaffold** protein regulating vesicle and organelle trafficking and signaling pathways.
INADL (PATJ)	scaffold	17234746	Here we report, using a two-hybrid assay, a direct molecular interaction between **TSC2** C-terminal part and PDZ 2 and 3 of **PATJ**, a **scaffold** member of the Crumbs 3 (CRB 3) complex in human intestinal epithelial cells, Caco2.
	scaffold	16697075	One evolutionarily conserved protein complex, which can be found both in Drosophila and mammalian epithelial cells, is composed of the transmembrane protein Crumbs/Crb3 and the **scaffolding** proteins Stardust/Pals1 and D**PATJ**/**PATJ**, respectively, and localise
	scaffold	15863617	A unified assembly mode revealed by the structures of tetrameric L27 domain complexes formed by mLin-2/mLin-7 and **PATJ**/Pals1 **scaffold** proteins. AXIN1 almost all references mention it as a Scaffold
NPHS1 (Nephrin)	scaffold	18480178	**Nephrin**, an essential adhesion and **scaffolding** molecule expressed in podocytes, emerged in this screen
	scaffold	19443634	Within the glomerulus, the **scaffolding** protein **nephrin** bridges the actin-rich foot processes that extend from adjacent podocytes to form the slit diaphragm
PIK3R1 (p85)	scaffold	17024187	We show that **p85** acts as a **scaffold** to bind Cdc42 and septin 2 simultaneously. **p85** is thus involved in the spatial control of cytosolic division through regulation of Cdc42 and septin 2, in a PI3K-activity independent manner.
TANK	scaffold	18353649	Recent data provide insight into the requirement for **scaffold** proteins in complex assembly; NF-kappaB essential modulator coordinates some IKK complexes, whereas **TANK**, NF-kappaB-activating kinase-associated protein 1 (NAP1) or similar to NAP1 TBK1 adaptor
	scaffold	17823124	we have identified **TANK** as a **scaffold** protein that assembles some but not all IRF3/7-phosphorylating TBK1-IKKepsilon complexes

Table 4.3: Literature review of seven scaffold protein candidates obtained by our interactome search. The complete table containing the literature review of all candidates manually inspected appears in appendix A. Symbol: NCBI gene symbol; type: the protein is refereed to as 'scaffold', 'adaptor' or both in the literature; pubmed id: NCBI PubMed identifier; text snippet: extract from the abstract where a description of the scaffold candidate is found.

4.4 Huntingtin as scaffold protein

Huntingtin (Htt) is a multifunctional human protein involved in diverse cellular processes such as synaptic signaling, transcriptional regulation, anti-apoptotic activity, and vesicular trafficking (Caviston & Holzbaur, 2009; Imarisio et al., 2008). A pathogenic expansion of the polyglutamine repeat region in the Htt sequence is implicated in the neurodegenerative disorder Huntington's disease. Htt was identified as scaffold candidate in the computational search, and further literature review confirmed this finding. For instance, Htt is reported to act as scaffold protein by mediating the complex formation of the mitogen-activated protein kinase kinase kinase 10 (MAP3K10), a JNK signaling pathway protein, and NeuroD, a transcription factor (Marcora et al., 2003). In this protein complex, Htt binds indirectly to NeuroD via the huntingtin-associated protein 1 (HAP1). Furthermore, Htt functions as scaffold when coordinating the binding of motor proteins to vesicular cargo (Caviston & Holzbaur, 2009). Generally, Htt is known to interact with dozens of proteins and at least ten CRPs including MAP3K10. Thus, Htt may play a pronounced role as scaffolding protein in multiple signaling pathways. Interestingly, Htt also associates with other signaling scaffold proteins like DLG4, GIT1, ITSN and TRIP10 and the adaptor GRB2, the latter of which interacts indirectly with other scaffold and adaptor proteins (Fig. 4.3). This observation may particularly point to the formation of supramolecular scaffold complexes containing Htt.

4.5 Summary

Using the definition proposed by Zeke et al., hundreds of human scaffold candidates were discoverd whose functional properties agree well with known scaffold proteins. Nevertheless, experimental verification and manual curation is still needed. Based on the analysis, is estimated that the false positive and negative rates of our results are below 40–50%, but exact rates are difficult to obtain. Considering these rates, $\sim$ 300 proteins may be a first rough estimate of the overall number of scaffold proteins in the human proteome.

Further refinements of the molecular characteristics of scaffold proteins as well as qualitative and quantitative advances in gene and protein function annotation may result in an even more reliable list of scaffold candidates. However, an important note of caution is that the overall amount of currently available protein interaction data for human is still small and of differing quality (Zeke et al.,

2009; Ramírez et al., 2007), which currently limits complete interaction-based searches for scaffold candidates. Apart from that, while the computational search used pairwise protein interactions because of the original definition given by Zeke et al., whole scaffold complexes including stoichiometric data may also be identified experimentally by recent mass spectrometry-based techniques or found in existing datasets of protein complexes using a definition similar to that of scaffold candidates.

5

Novel Search Method for the Discovery of Functional Relationships

This chapter describes a novel search method for the discovery of biological relationships based on the similarity of gene and protein annotations (**?**). This method allows to quickly scan the data warehouse for genes or proteins that are similarly annotated.

By using this novel method, called *BioSim*, the warehouse can now be interrogated in novel ways, for example to identify new gene-disease associations based on known associations as will be shown in this chapter.

The BioSim method also facilitates the development of derived applications such as the clustering of genes and proteins by function or the assessment of protein-protein interactions which are discussed in the next chapter. These applications have a growing demand caused by the popularization of high-throughput techniques in order to analyze and prioritize the resulting lists of genes and proteins.

5.1 Introduction

Similarity search plays an important role in biological, pharmaceutical, and medical investigations. For instance, the introduction of the BLAST algorithm by Altschul et al. (1990) to search for similar sequences is considered a milestone in genomics (Bahcall, 2007), and similarity search methods to mine databases of three-dimensional molecule conformations have been important for drug discovery (Willett et al., 1998). Presently, the growing availability of annotations that characterize genes and proteins (Reeves et al., 2008) opens the new possibility to find biological relationships by similarity searches based on function, domain composition, disease association, tissue expression, etc. For example, the identification of similarly annotated genes and proteins can reveal new gene-disease

associations (Aerts et al., 2006), suggest novel protein functions (Friedberg, 2006), and indicate new drug targets (Chan et al., 2010).

In general, similarity searches compute pairwise similarities of a query with the entities in a database to obtain a ranked list of high-scoring similarities. In particular, a number of methods have been proposed for the quantification of pairwise similarities of gene and protein annotations. Most of those functional similarity methods are based on Gene Ontology (GO) annotations (Chabalier et al., 2007; del Pozo et al., 2008; Speer et al., 2004; Sevilla et al., 2005; Lord et al., 2003; Popescu et al., 2006; Schlicker et al., 2006; Lerman & Shakhnovich, 2007; Mistry & Pavlidis, 2008; Pesquita et al., 2008; Benabderrahmane et al., 2010). However, the last years have shown a dramatic growth in datasets that result from high-throughput experiments and computational work and yield annotation sources that provide manifold information about, for instance, protein interactions, signaling circuits, metabolic pathways, cellular localization, tissue expression, disease associations, and protein domain architecture. Currently, only one similarity search method explicitly takes multiple annotation sources into account, namely, the *kappa coefficient* used by the DAVID Gene Functional Classification Tool (Huang et al., 2007). In contrast, the integration of multiple annotation sources into a network structure is often applied in the context of gene function prediction (Huttenhower et al., 2009; Warde-Farley et al., 2010; Wang & Marcotte, 2010).

When developing efficient methods for searching through gene and protein annotation data, a particular task is the construction of data structures that represent the annotations. Most methods rely on the graph structure of GO to estimate quantitative semantic relationships among the gene/protein annotations (Pesquita et al., 2009). However, the GO structure limits the inclusions of non-ontological (i.e., non-GO) annotations into methods. A flattened representation of the GO hierarchy solves this problem by storing the annotations as Boolean arrays in which the presence and absence of annotations is recorded (Huang et al., 2007). This representation implicitly contains the ontological relations and allows for the inclusion of non-ontological annotations as part of the array. This avoids the inference of relationships through the hierarchical structure of GO. GO-based similarity methods that use this data structure are *COS* (Chabalier et al., 2007), *simGIC* (Pesquita et al., 2008) and *TO* (Mistry & Pavlidis, 2008). Although these methods do not consider annotation sources other than GO, they achieve better performance than methods such as those of Resnik (1999) and Lin (1998) that depend on the GO graph structure.

In this chapter, we will introduce the new method *BioSim* for similarity searches based on diverse annotation sources of gene and protein function and extend the existing methods *COS*, *kappa coefficient*, *simGIC*, and *TO* to utilize annotations not only from GO, but from 22 major biological databases for human genes and proteins. We will also compare the performance of *BioSim* with the other methods in different benchmarks.

5.2 Data sources

22 biological databases integrated in the data warehouse in 2009 were used for this study. These include functional annotations from all three GO categories (MF, molecular function; BP, biological process; CC, cellular component) and from the UniProtKB controlled vocabulary of keywords. The data warehouse also contains clusters of similar sequences from Ensembl protein families and from UniRef90; protein domain architectures from Pfam and InterPro; metabolic and signaling pathways from HumanCyc, KEGG, and Reactome; protein interactions and protein complexes from CORUM, DIP, HiMAP, HPRD, IntAct, MINT, PDB, and STRING; disease associations from OMIM; enzyme classifications from the Enzyme nomenclature database; gene expression data for different tissues and cell lines from the Novartis Gene Atlas; Mammalian Phenotype Ontology annotations of human genes as provided by the Mouse Genome Database; and orthologs of protein sequences from OrthoMCL.

From the annotation sources, the functionally relevant features associated with individual genes and proteins were extracted. In the following, these features are referred as *annotation terms*, which correspond, for example, to a specific molecular function (e.g. oxidoreductase activity), domain (e.g. SH2) or pathway (e.g. glycolysis) annotated to genes and proteins. To enable comparisons between functional similarity methods using multiple annotation sources and those using only GO annotations, proteins with no available GO annotation were excluded. This resulted in a list of 18,076 protein entries out of 20,177 manually reviewed proteins in UniProtKB release 15.5.

5.3 Functional similarity methods

In the following, A_X and A_Y denote the sets of annotation terms associated with the gene products X and Y, respectively.

BioSim

This novel method, developed during my Ph.D., is defined as follows:

$$\mathrm{BioSim}(X,Y) = \prod_{t \in \{A_X \cap A_Y\}} p(t)$$

Here, $t \in \{A_X \cap A_Y\}$ is the set of common annotation terms between X and Y and $p(t)$ is the probability that both A_X and A_Y contain term t by chance. Since *BioSim* is the product of the probabilities $p(t)$, a score of zero represents the highest similarity and a score of one the lowest. This is in contrast to other methods described below, except *TO*. The probability $p(t)$ is estimated using the cumulative hypergeometric distribution:

$$p(t) = \sum_{k=2}^{D} \frac{\binom{N_t}{k}\binom{N-N_t}{D-k}}{\binom{N}{D}}$$

In this case, the cumulative hypergeometric distribution describes the probability of getting at least 2 two proteins annotated with the same term in a sequence of D draws, without replacement, from a population of N proteins. N_t is the number of proteins annotated with term t and D is the sum of $|A_X|$ and $|A_Y|$. The resulting probability not only depends on the frequency of the annotation term N_t but also on D. This is an important feature of our method to account for the annotation bias that exists for intensively studied genes and proteins. A pair of proteins associated with many annotations terms (large D) has an increased probability $p(t)$ to share the annotation term t (i.e., a decreased functional similarity) in comparison to a pair of proteins associated with few annotations terms (small D).

Term overlap length (TO)

TO represents the number of annotations terms shared by two proteins X and Y (Mistry & Pavlidis, 2008):

$$\mathrm{TO}(X,Y) = |\{A_X \cap A_Y\}|$$

Kappa coefficient (KC)

This method is used in the well-known DAVID Gene Functional Classification Tool (Huang et al., 2007). It computes a normalized difference of the observed number of annotation terms $O(X,Y)$ shared by two proteins X and Y, and the expected number $E(X,Y)$ of shared annotation terms that are randomly chosen (Huang et al., 2007). It is defined as follows:

$$\mathrm{KC}(X,Y) = \frac{O(X,Y) - E(X,Y)}{1 - E(X,Y)}$$

In the following, we describe the *simGIC* and *COS* methods. Unlike the previous methods, they incorporate term weights based on the information content (IC) (Resnik, 1995) of a term t:

$$\mathrm{IC}(t) = -\log \frac{N_t}{N}$$

Here, N_t is the number of proteins annotated with term t and N the total number of proteins in our study.

simGIC

This method introduced by Pesquita et al. (2008) includes the summed information contents of shared vs. all annotated terms for two proteins X and Y:

$$\mathrm{simGIC}(X,Y) = \frac{\sum_{\forall t \in \{A_X \cap A_Y\}} \mathrm{IC}(t)}{\sum_{\forall t \in \{A_X \cup A_Y\}} \mathrm{IC}(t)}$$

Cosine similarity (COS).

This classical method is defined as follows (Salton et al., 1975):

$$\mathrm{COS}(X,Y) = \frac{\vec{A}_X \cdot \vec{A}_Y}{|\vec{A}_X||\vec{A}_Y|}$$

Here, $\vec{A}_X$ and $\vec{A}_Y$ are the annotation vectors of two proteins X and Y, respectively. In each vector, the absence of an annotation term is represented by 0 and

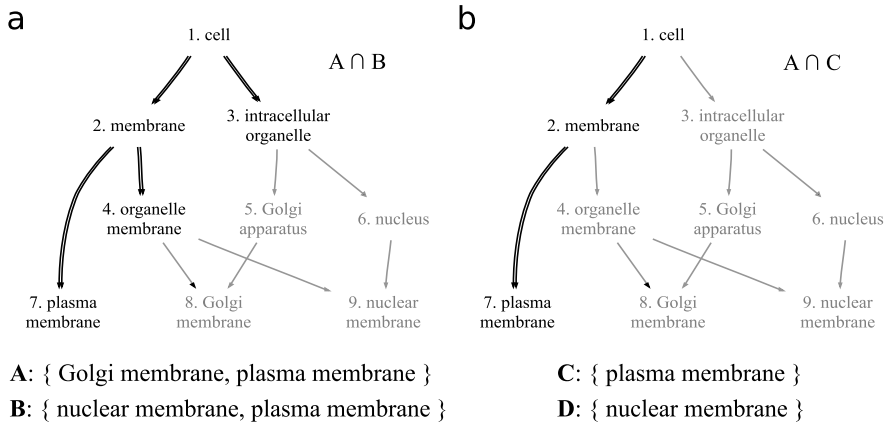


Fig. 5.1: Gene Ontology (GO) cellular component sub-tree (simplified version). Hypothetical annotations of proteins A to D. The graphs highlight the sub-trees shared by A, B (a) and A, C (b).

the presence by IC(t). This method was first used in the context of functional similarity by Chabalier et al. (2007).

5.4 Representation of ontological annotations

Annotations based on ontological or hierarchical structures such as those from the Gene Ontology and Enzyme classifications were converted into a Boolean array following the method used in (Mistry & Pavlidis, 2008; Pesquita et al., 2008; Chabalier et al., 2007). In this method, not only the leaf terms found in the annotation sources are included in the boolean array, but also all ancestors of the term.

To illustrate the approach, consider the proteins A, B, C and D annotated with the following GO terms:

- A:{*Golgi membrane*, *plasma membrane*}
- B:{*nuclear membrane*, *plasma membrane*}
- C:{*plasma membrane*}
- D:{*nuclear membrane*}

A is more similar to B than to the rest because both are annotated as *plasma membrane* proteins, but also share the common ancestor terms *cell, membrane, intracellular organelle, organelle membrane* (Fig. 5.1). If no ancestor terms are taken into account, then A and B will share only the annotation term *plasma membrane*, which is the same term shared by A and C. Furthermore, A and D will not share any term. If ancestor terms are included, A and D will share the four annotations terms *cell, membrane, intracellular organelle, organelle membrane.*

5.5 Evaluation methods

Gold standard

To evaluate the performance of the functional similarity methods, we collected a dataset composed of groups of proteins that are assumed to be functionally related (to a certain extent) and contained in the list of 18,076 proteins with at least one available GO annotation (as described previously). We use this dataset as gold standard in our validation. The protein groups in the dataset were obtained from four benchmark categories that we limited to at most 400 groups per category: (1) 400 groups containing curated protein complexes randomly selected from a total of 2,030 complexes from CORUM; (2) 88 groups of sequence clusters containing closely similar protein sequences based on UniRef90 clusters (sequences of at least 90% identity) and thus with putatively similar functions; (3) 355 groups consisting of reliable interaction partners from a total of 355 proteins with at least two such reliable partners (here, an interaction is reliable if it is reported in at least three different publications); and (4) 400 groups composed of proteins participating in metabolic and signaling pathways from KEGG and Reactome (protein groups were selected randomly from a total of 424 available pathways). Groups of more than 20 proteins were excluded as being too general. The average group size had 6.7 proteins and the overall standard deviation was 4.3. In total, the gold standard consisted of 1,243 groups containing 8,150 proteins (some of the proteins were shared in different groups). In the following, we will refer to those groups as validation groups.

Benchmarking procedures

From each validation group, a query protein was randomly selected and the remaining group members were regarded as gold standard positives. To obtain

ranked lists, pairwise functional similarity scores were computed between the query protein and all other 18,076 protein entries used in our study. The same evaluations were carried out using either only GO annotations or all aforementioned annotation sources (excluding the respective annotation source of the benchmark category).

For baseline comparison, a dataset of 10,000 protein pairs was randomly created and their functional similarity scores were computed using all methods.

To compute a background distribution of sequence similarity we computed the BLAST bit scores (NCBI blastp version 2.2.22) for 100,000 protein pairs randomly drawn from the list of studied proteins. Since the bit score of a protein pair is not symmetric, the average bit score of the pair was used (Pesquita et al., 2008).

Performance measures

The *recall* at a rank k is the number of positives in the k top ranks of the computed ranking list divided by the total number of positives, i.e., the members of the respective validation group. The *average precision* is the mean of the precisions obtained for the ranks of all positives in the ranking list (Buckley & Voorhees, 2000). For example, in case of three positives found at ranks 2, 5, and 10, the average precision would be $(1/2 + 2/5 + 3/10)/3 = 0.4$. The *Precision* at a rank k is the number of positives in the k top ranks divided by k. The *first relevant rank* (FRR) is the best rank of a positive in some ranking list.

Score cut-offs for the functional similarity methods

Using the ranking lists obtained for each validation group, we identified the functional similarity score that yielded 50 false positives. This number is a reasonable threshold suggested by Gribskov & Robinson (1996) for their ROC_{50} method. By averaging these functional similarity scores, suitable score cut-offs were obtained for every similarity method. We refer to these score cut-offs as SC_{50}. The performance curves were generated using the ROCR package (Sing et al., 2005).

Disease associations

For each disease phenotype and each gene not associated with this phenotype, we averaged the computed pairwise functional similarities to every disease gene

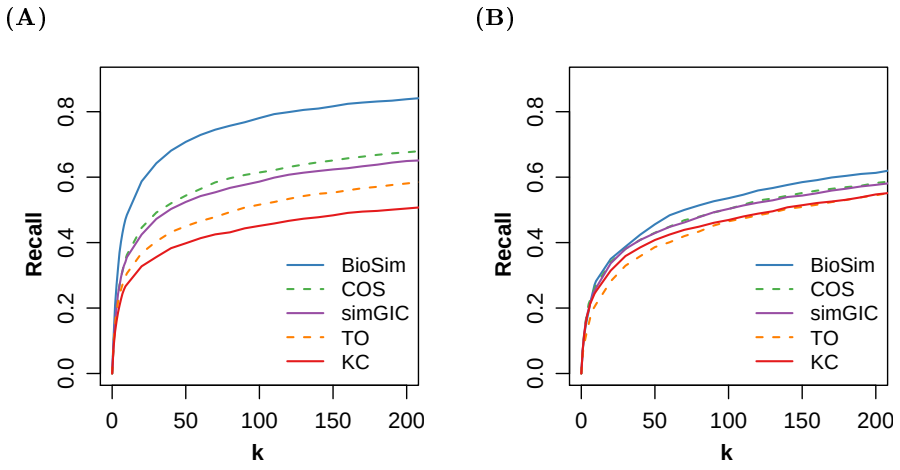


Fig. 5.2: Performance of functional similarity methods. Average recall is plotted for different top ranks k using either multiple annotations sources (**A**) or only GO annotations (**B**). The average values were obtained from benchmarking with 1,243 validation groups. See Fig. 5.3 for details on the performance of the methods in each of the four benchmark categories.

of this phenotype. The averaged scores were used to obtain the final ranking list of unassociated genes.

	curated complexes		protein interactions		similar sequences		pathways		total	
method	**avg. prec.**	**FRR**	**avg. prec.**	**FRR**	**avg. prec.**	**FRR**	**avg. prec.**	**FRR**	**avg. prec.**	**FRR**
Multiple annotation sources										
BioSim	**0.49**	**1**	**0.21**	**5**	**0.55**	2	**0.42**	**1**	**0.39**	**2**
COS	0.36	2	0.12	25	**0.55**	2	0.29	2	0.28	3
simGIC	0.35	2	0.13	22	0.52	1.5	0.30	2	0.28	3
TO	0.24	2	0.15	8	**0.55**	**1**	0.26	2	0.24	3
KC	0.25	3	0.10	83	0.48	1.5	0.23	3	0.21	5
Only GO annotations										
BioSim	**0.25**	6	0.10	38	0.28	6	**0.28**	2	**0.22**	7
COS	0.23	5	**0.11**	54	**0.36**	**2**	**0.28**	2	**0.22**	**5**
simGIC	0.22	**3**	0.10	43	0.34	2.5	**0.28**	2	0.21	**5**
TO	0.19	7	0.09	**27**	0.12	30	0.23	4	0.17	11
KC	0.21	**3**	0.10	58	0.33	3	0.25	2	0.20	**5**

Table 5.1: Performance comparison of functional similarity methods using multiple annotation sources vs. using only GO annotations, over all 1,243 validation groups. In bold, the highest values of each category are highlighted. avg. precision: average precision, FRR: first relevant rank.

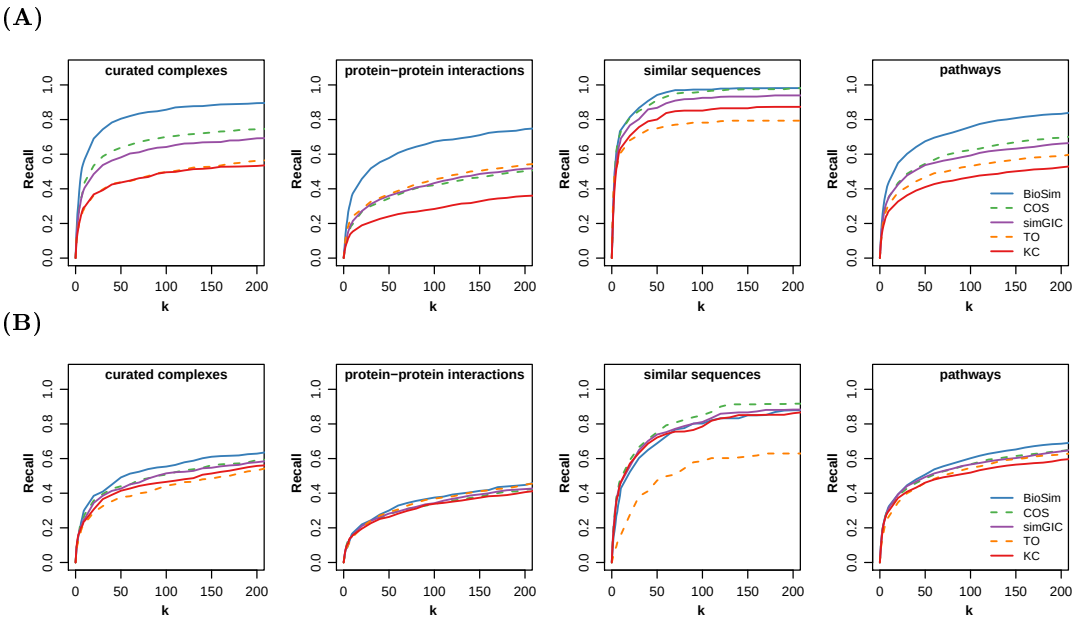


Fig. 5.3: Performance of functional similarity measures by validation group. Average recall plotted for different top ranks k in the four validation groups. (**A**) Results based on multiple annotation sources. (**B**) Corresponding results using only GO annotations.

5.6 Performance of functional similarity methods

The performance of *BioSim* in identifying known functional similarities was compared with that of four other methods: *TO*, *KC*, *simGIC*, and *COS*. Results were averaged over all validation groups. While all methods showed similar performance when using only GO annotations, the performance was improved when considering multiple annotation sources (Fig. 5.2). Notably, *BioSim* achieved better performance than the other methods. For instance, the top twenty hits of *BioSim* had an average *recall* of 0.58. The second best method, *COS*, had an average *recall* of 0.44 (Fig. 5.2**A**). The *average precision* of *BioSim* was 0.39, which was significantly higher than that of the other methods (p-value < 0.01, Wilcoxon signed-rank test). Likewise, *BioSim* had a median value of 2 for the FRR, surpassing the other methods (Table 5.1).

The overall performance of the methods varied for each benchmark category. It was lower for all methods when using the protein-protein interactions category and higher for the sequence clusters category (Fig. 5.3**A** and Table 5.1). The combined average *recall* for all methods was almost one third lower in the protein-protein interactions category than in the sequence clusters category (the respective *recalls* were 0.29 and 0.75). The observed high performance when using the sequence clusters category is due to the tendency of the methods to

rank similar sequences at the top. This can be explained, to some extent, by annotation transfer between homologous protein sequences, by gene annotations that are transferred to all encoded proteins, and by domain annotations that are almost identical for similar sequences. Therefore, the tendency to rank similar sequences at the top reduces the performance of the methods when using benchmark categories different from sequence clusters because gold standard positives are displaced to lower ranks.

5.7 Performance of GO vs. multiple annotation sources

The use of multiple annotation sources improved the performance of four of the five methods although they were not originally developed to handle multiple annotations (in contrast to *BioSim*). Much of this increase seems to be attributable to the availability of more annotation terms per protein. The number of terms annotated to each protein increased from a median of 7.5 GO terms to a median of 15.0 annotation terms when all annotation sources were included (Fig. 5.6). The *TO* method, which counts the number of common terms, but does not account for term specificity, improved its *average precision* from 0.17 to 0.24 when all annotations were used.

Notably, the use of multiple annotation sources does not only increase the number of annotation terms per protein, but also improves the specificity of the annotations. While GO terms annotated to at most four proteins were available for 8,096 proteins, this number doubled to 16,649 proteins in case of multiple annotation sources when not only using GO. The positive effect of the increased annotation specificity on the performance can be observed with the three functional similarity methods *COS*, *simGIC*, and *BioSim*. All three methods weight annotation terms and showed the strongest performance improvement when multiple annotation sources were included.

In particular, *BioSim* was best able to take advantage of the increased number and improved specificity of annotations terms, as shown by the near doubling of its average precision (Table 5.1). In the case of *BioSim*, the functional similarity between two proteins increases if both are annotated with specific terms (terms that are annotated to few proteins) because the corresponding probabilities of the terms are low. Additionally, since *BioSim* computes the product of the probabilities of all terms shared by two proteins, a certain number of even less specific

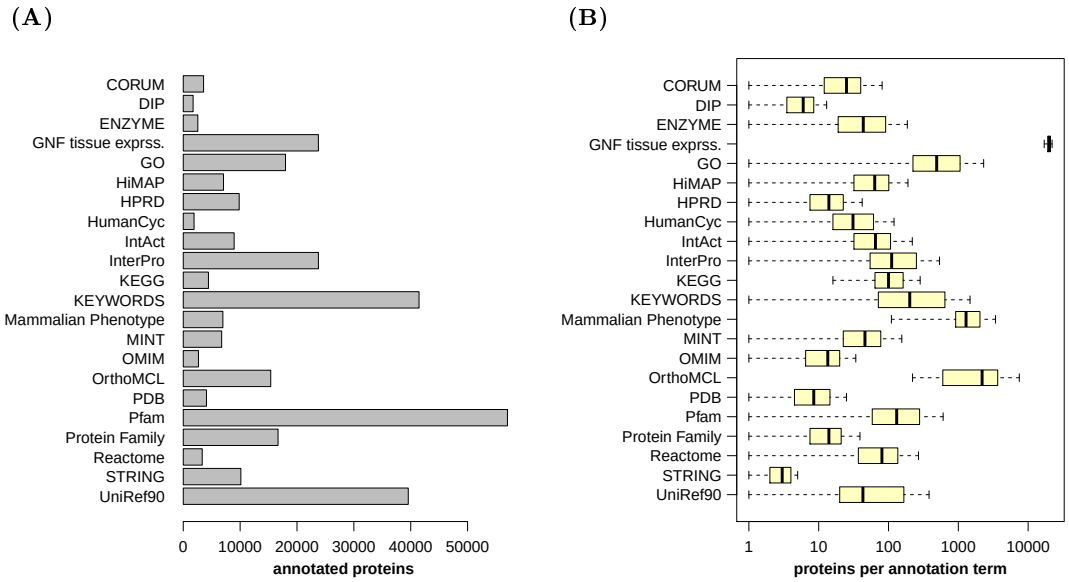


Fig. 5.4: Coverage of data sources. (**A**) Number of proteins annotated in each data source. (**B**) Number of proteins per annotation term in each data source.

terms still increases the overall functional similarity. Annotations from protein-protein interactions, sequence clusters, pathways, and disease associations are normally the most specific and least abundant ones, annotated to no more than a hundred proteins. In contrast, annotations as from cellular localization and tissue expression frequently cover thousands of proteins; and annotations from GO, UniProtKB keywords and protein domains span the whole range from just a few proteins to thousands (Fig. 5.4).

As an example, we looked in detail at one known SNARE protein complex formed by the proteins VAMP2, SNAP25, STX1a and CPLX1. These four proteins are involved in the fusion of neurotransmitter-containing vesicles with the pre-synaptic membrane (McMahon et al., 1995). When *BioSim* was applied using multiple annotation sources to compute the functional similarity of VAMP2 with each of the 18,076 human proteins in our study, SNAP25 achieved the top rank 1 with the strongest functional similarity. The other two complex members STX1a and CPLX1 were found at ranks 3 and 5, respectively. At rank 2 we found PRKD3, a protein that interacts directly with VAMP2, and at rank 4 we found VAMP1 who shares the Synaptobrevin domain with VAMP2. In contrast, when *BioSim* made use of only GO annotations, the rankings of SNAP25, STX1a and CPLX1 decreased to 25, 187, and 805, respectively. Specific annotations, which led to the identification of SNAP25 as functionally similar to VAMP2, included four experimental results that reported the interaction between VAMP2 and STXa1

and several shared pathways in Reactome such as the *proteolytic cleavage of SNARE complex proteins*. Less specific annotations were a shared coiled-coil domain and a similar tissue expression profile. When only GO annotations were taken into account, ICA69 was the protein functionally most similar to VAMP2, primarily, because both proteins are annotated with the term *secretory granule membrane*. This term covers only 25 other proteins, none of which is SNAP25, STX1a or CPLX1. The current knowledge about ICA69 is very limited. It might play a functional role in the transport regulation of insulin secretory granule proteins (Buffa et al., 2008) as well as in neurotransmitter transport as inferred by sequence similarity in UniProtKB. However, ICA69 has not been associated with the fusion of pre-synaptic vesicles.

In general, although GO annotations are expected to improve over time as more information is added, the use of other annotation sources helps to bridge the time until new data is incorporated. Furthermore, useful annotations to derive functional similarities such as protein-protein interactions and disease associations are not part of GO. Moreover, the use of multiple annotation sources can also reduce the impact of incorrect annotations found in biological databases (Schnoes et al., 2009).

5.8 *BioSim* scoring versus other methods

BioSim distinguished functional relationships of gold standard positives from those of randomly paired proteins better than the other methods. Gold standard positives consistently received a *BioSim* score close to 0, while random pairs obtained a score close to 1 (Fig. 5.5**A**). In particular, we plotted *precision* and *recall* averages from our benchmark results for every method at different score cut-offs (Fig. 5.5**B**). We also computed a score cut-off (SC_{50}) that resulted in 50 false negatives on average. The obtained SC_{50} score cut-offs, along with the score range of each method from lowest to highest functional similarity, were: *BioSim*: $\leq 1.18 \times 10^{-9}$ (range $[1;0]$), *TO*: ≥ 115 (range $[0;\infty)$), *KC*: ≥ 0.360 (range $[0;1]$), *simGIC*: ≥ 0.096 (range $[0;1]$), and *COS*: ≥ 0.101 (range $[0;1]$). For *COS* and *simGIC*, the second and third best methods, the SC_{50} score cut-offs were very close to zero, their non-similarity score; the *recall* at the respective SC_{50} cut-off had a median of 0.50 and a distribution covering the whole range (see Fig. 5.5**B**). In other words, for both methods, the SC_{50} cut-off resulted in very different *recalls* in each benchmarking group. The *KC* and *TO* methods had a *recall* median below 0.5 for their respective SC_{50} score cut-offs. In comparison, the *recall* for *BioSim*

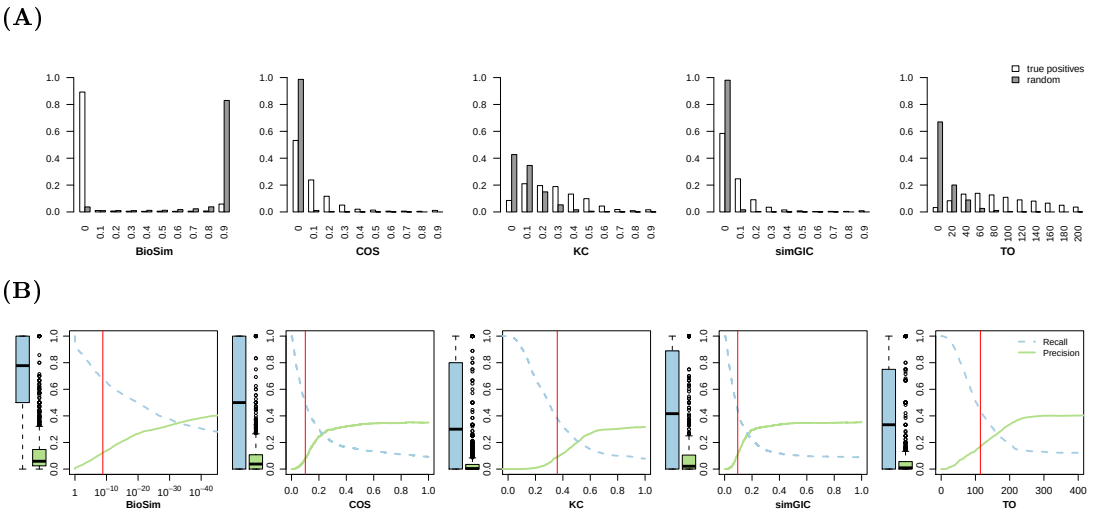


Fig. 5.5: Comparison of functional similarity methods. (**A**) Histograms of the functional similarity scores that were obtained for 6,907 pairs of gold standard positives and for 10,000 random pairs. (**B**) Precision (straight lines) and recalls (dashed lines) are averaged at different cut-offs. The vertical red lines highlight the SC_{50} score cut-offs that yield, on average, 50 false positives. The box plots to the left of the y-axis shows the distribution of recalls (light-blue) and precision (light-green) at this cut-off. *BioSim* scores are in logarithmic scale for better visualization.

at the SC_{50} score cut-off had the highest median (0.82) and the corresponding distribution concentrated around high values.

The limited consistency of the scores of *COS*, *KC*, *simGIC*, and *TO* is probably caused by annotation bias towards better studied molecules (Rhee et al., 2008) as these methods appear to be best suited for unbiased data (Wang et al., 2010). In our data warehouse, a handful of proteins have over thousand annotations, while the majority has less than ten annotations. A similar pattern can be observed when considering only GO annotations (Figs. 5.6 and 5.7). About 16 % of all proteins are annotated only with less specific terms such as the UniProtKB keyword "Receptor" or the GO term "protein binding". The functional similarity of any two proteins sharing such terms is overestimated by the *COS*, *KC* and *simGIC* methods, which yield the highest score of 1. This misleading result is undistinguishable from a genuine functional similarity based on several shared annotation terms.

Furthermore, the same methods tend to underestimate the genuine similarity of any two proteins that are annotated with numerous terms and do not share a large proportion of their annotation terms. For example, the cellular tumor antigen TP53 (with 1,642 annotation terms including 332 literature-curated protein interactions) shares approximately 19 % of its annotation terms with the closely

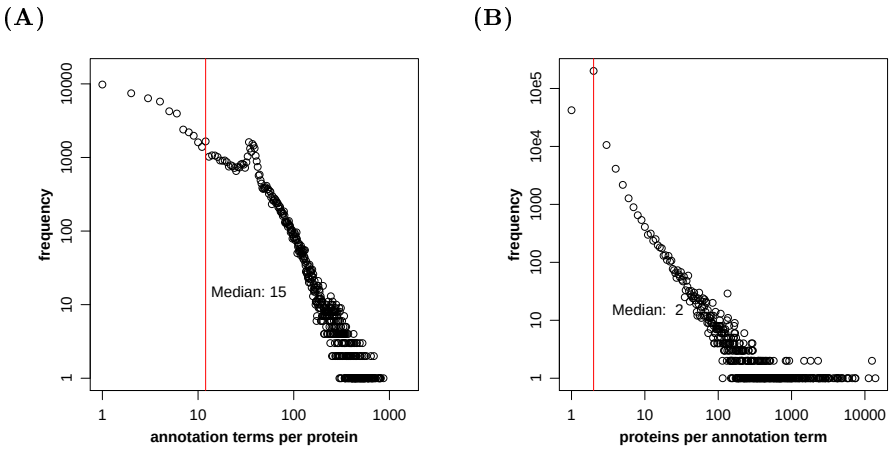


Fig. 5.6: Distribution bias in the annotation of human proteins. (**A**) Frequency plot of the number of annotation terms per protein. Few proteins have many annotations, while a large number of proteins are barely annotated. For instance, the protein with the most annotations is p53 with 1,642 annotation terms, which include 332 literature-curated protein-protein interactions. Generally, proteins with a large number of annotations are very well studied due to strong biomedical interests. On average, each protein is annotated with 29.7 terms. (**B**) Frequency plot of the number of proteins per annotation term. Most annotation terms are specific for few proteins, while other terms are broad and cover many proteins. The increased number of terms annotated to only two proteins is due to pairwise protein-protein interactions.

related E3 ubiquitin-protein ligase MDM2, which is known to bind and inhibit TP53 (Vassilev et al., 2004). Relevant terms indicate common metabolic and signaling pathways, disease associations and protein interactions. However, the remaining 81 % of TP53 annotation terms that are not shared with MDM2 lead to the following low functional similarity scores:

	functional similarity score	
method	**multiple sources**	**only GO**
COS	0.097	0.206
KC	0.120	0.379
simGIC	0.056	0.142

These functional similarity scores are even below the SC_{50} cut-offs for the respective methods. This means that low functional similarity scores are often obtained for truly functionally related proteins. Moreover, such low similarity scores are also obtained when only GO annotations are considered.

The *TO* method, which is simply the count of annotation terms shared by two proteins, avoids some of the described shortcomings by focusing only on

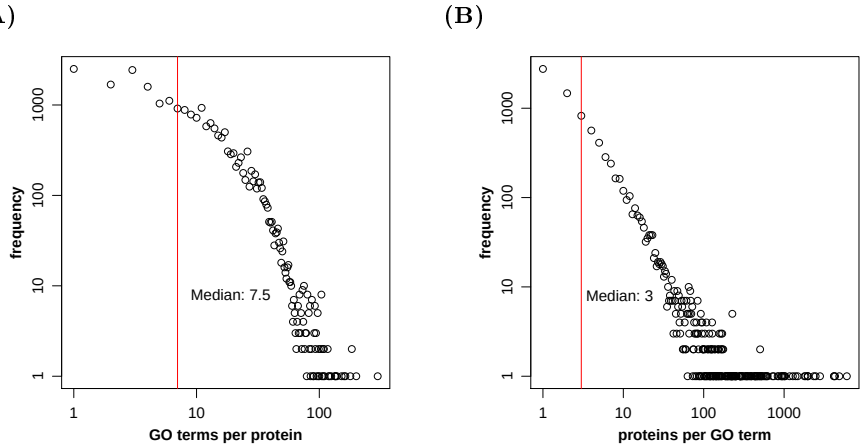


Fig. 5.7: Distribution bias in the GO annotation of human proteins. (**A**) Frequency plot of the number of GO terms per protein. (**B**) Frequency plot of the number of proteins with the same GO term. Both distributions are similar to those in Fig. 5.6.

the shared annotations. However, it cannot distinguish those annotations that occur by accident because it judges an event of two proteins sharing a rather unspecific, frequent annotation term (e.g. "protein binding") as likely as an event of two proteins sharing a very specific, rare annotation term (e.g. "actin filament binding").

5.9 Functional similarity vs. sequence similarity

The correlation between the functional similarity of two proteins and their sequence similarity is often used to evaluate functional similarity methods (Pesquita et al., 2008; Lord et al., 2003). In our results, rank correlations for all methods were close to 0.1 when comparing BLAST bit scores and functional similarity scores for 100.000 random pairs of proteins. This low correlation is likely due to many protein pairs with almost no sequence similarity, but some functional similarity (Fig. 5.8). To filter out protein pairs with low sequence similarity, we discarded all pairs having a ln(bit score) below 3.3. This threshold was chosen after observing that, for all methods, the averaged functional similarity scores increases above this value. In total, 631 (0.63 %) of the random pairs had a ln(bit score) of at least 3.3. The rank correlations for these pairs were COS: 0.77, KC: 0.67, $BioSim$: 0.69, $simGIC$: 0.73, TO: 0.48.

Since $BioSim$ showed a slightly lower correlation than COS and $simGIC$, we additionally analyzed some interesting cases manually. Table 5.2 summarizes the

gene symbol	bit score	BioSim	COS	KC	simGIC	TO	shared specific annotations	shared general annotations
CDH10 – CDH9	1135.5	1.7×10^{-129}	0.81	0.92	0.77	130	protein-protein interaction, Ensembl Family, Pfam architecture	GO BP, InterPro
HSPH1 – HSPA4L	1011.0	2.6×10^{-13}	0.11	0.34	0.11	48	UniProtKB Keyword Biological process, Pfam architecture, Pfam family	GO BP, UniProtKB Keyword Ligand
RHBDF1 – RHBDF2	947.5	2.4×10^{-43}	0.51	0.45	0.26	31	Ensembl Family, Pfam family	GO CC, UniProtKB Keyword Cellular component
ZNF229 – ZNF208	614.5	5.1×10^{-5}	0.40	0.60	0.36	49		InterPro, Pfam family
PCDHB6 – PCDHGB1	606.0	3.8×10^{-9}	0.21	0.31	0.10	34	Pfam architecture, Pfam family	GO BP, Pfam family
VLDLR – LRP4	541.0	2.9×10^{-20}	0.23	0.49	0.16	94	InterPro, Pfam family	InterPro, Pfam family
ZNF284 – ZNF45	539.0	2.8×10^{-4}	0.49	0.52	0.34	49		InterPro, Pfam family
MTSS1 – MTSS1L	503.0	1.2×10^{-32}	0.33	0.38	0.18	47	Ensembl Family, InterPro, Pfam family	GO BP, GO MF, UniProtKB Keyword Ligand
ZNF7 – ZNF845	500.5	1.7×10^{-2}	0.31	0.61	0.32	43		GO BP, InterPro, Pfam family
ZNF732 – ZNF790	495.0	0.12	0.50	0.66	0.40	388		GO BP, UniProtKB Keyword Domain
ZNF93 – ZNF442	482.0	2.0×10^{-3}	0.42	0.76	0.43	74		Ensembl Family, InterPro, Pfam family
ZNF300 – ZNF181	460.0	4.5×10^{-5}	0.42	0.72	0.47	49		Ensembl Family, InterPro, Pfam family
ZNF623 – ZNF84	457.0	1.2×10^{-2}	0.32	0.55	0.30	43		Ensembl Family, InterPro, Pfam family
ZNF429 – ZNF320	449.0	5.2×10^{-10}	0.59	0.74	0.56	50	Pfam architecture	InterPro, Pfam family
TPO – MPO	444.5	2.1×10^{-37}	0.15	0.38	0.12	97	Ensembl Family, UniProtKB Keyword Biological process, UniProtKB Keyword Molecular function	GO BP, GO MF, UniProtKB Keyword Ligand

Table 5.2: Manual inspection of annotations terms shared by proteins with very similar sequences. The table lists fifteen protein pairs with the highest BLAST sequence similarity bit score out of 100,000 analyzed random protein pairs. The columns 'shared specific annotations' and 'shared general annotations' summarize the very specific and rather unspecific annotations common to the respective protein pair. Here, specific terms are annotated to less than 100 proteins, while general terms are annotated to at least 100 proteins. GO BP, CC, and MF are abbreviations for the three Gene Ontology domains Biological Process, Cellular Component, and Molecular Function, respectively.

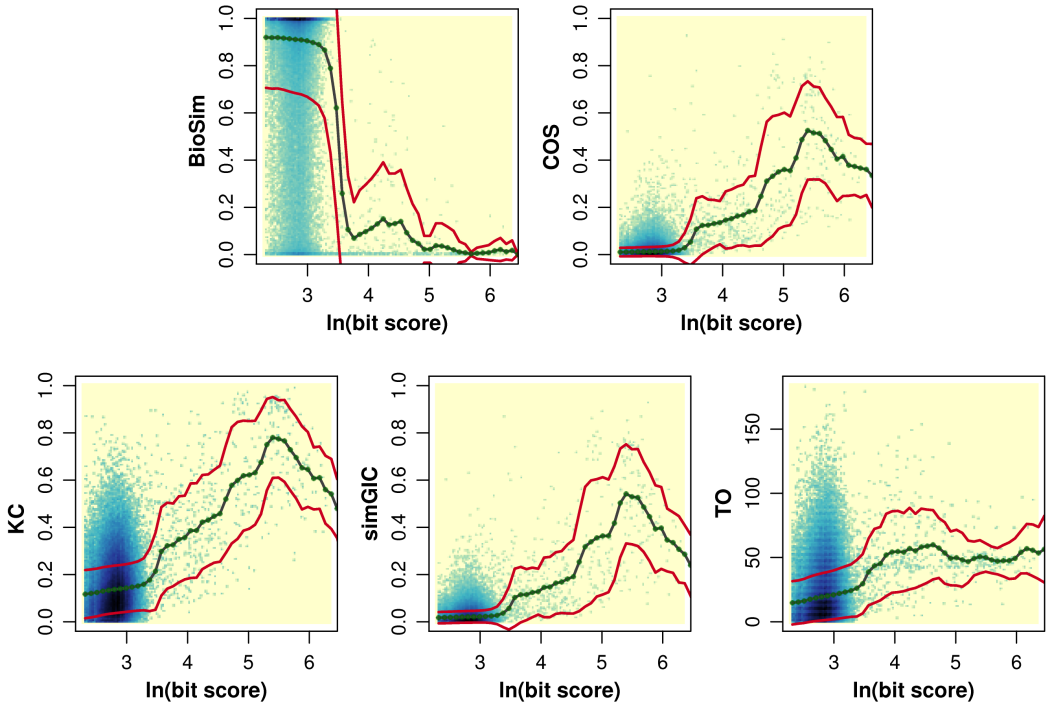


Fig. 5.8: Comparison of functional similarity and sequence similarity scores. 100,000 random pairs of proteins were analyzed. Sequence similarity is measured as ln(bit score). Green lines depict the average functional similarity. Red lines illustrate the standard deviation. In each plot, the background contains a scatter plot where darker colors indicate higher density of dots. A number of protein pairs with low sequence similarity have some functional similarity in all cases.

manual inspection of annotations shared by the fifteen pairs of proteins with the highest sequence similarity bit score. Seven protein pairs do not share specific annotation terms suitable for inferring a clear functional relationship. Accordingly, the low *BioSim* scores of those pairs are above the previously determined SC_{50} score cut-off of 1.18×10^{-9}, which indicates a weak functional similarity. In contrast, a true functional relationship between the remaining eight protein pairs is more evident due to several shared specific annotations terms. This agrees well with *BioSim* scores below or very close to the SC_{50} cut-off, which suggests a considerable certainty of a real functional similarity. However, in contrast to *BioSim*, the scores from the other methods do not allow a clear-cut distinction in those cases as explained in the preceding Section 5.8. For example, the second and fifteenth rows in Table 5.2 are cases of low functional similarity scores for *COS*, *KC*, and *simGIC* in contrast to *BioSim* although the respective proteins share numerous annotations. This suggests that a meaningful comparison of scoring methods based on the correlation of functional similarity and sequence similarity is limited by the available annotation datasets and their overall characteristics

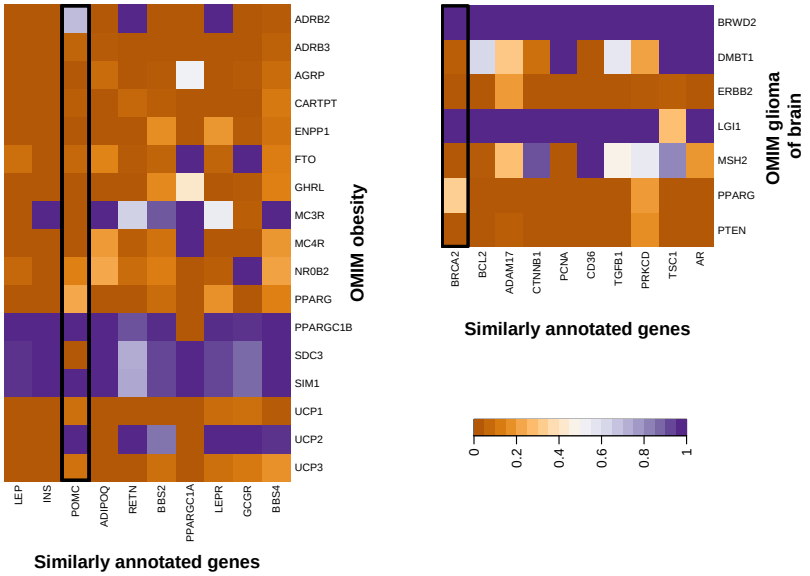


Fig. 5.9: OMIM disease-associated genes and its top 10 most functionally similar genes. The *BioSim* method was used to identify related genes for obesity (left) and the familial glioma of brain (right). The black frames highlight the new genes POMC and BRCA2 found by using *BioSim*. The vertical axis alphabetically lists the previously known disease genes. The horizontal axis ranks the most similar genes from left (most similar) to right. The colors indicate the strength of the functional similarity scores between the respective genes as computed by *BioSim*; lower scores indicate stronger similarity, see depicted color bar.

and quality, which can also be affected by annotation bias and incompleteness. Since *BioSim* is particularly designed to be more sensitive to the number and specificity of annotation terms in contrast to the other methods, its overall performance depends more on the annotation datasets and the individual annotation terms.

5.10 Discovery of disease-associated genes

Genes associated with the same disease phenotype tend to be functionally related (Schlicker et al., 2010; Vidal et al., 2011). Using *BioSim*, we ranked genes based on their functional similarity to genes known to be associated with a particular OMIM disease phenotype (Amberger et al., 2009). To this end, for each gene not associated with a disease phenotype, we averaged the computed scores of its functional similarity to the previously known disease genes. The functional similarity scores were computed using a snapshot of the data warehouse that con-

tained only gene annotations from before January 1, 2009. We then compared our results with an updated version of OMIM from October 31, 2009. This update contained 54 new gene associations for 46 diseases. In the results, eleven of the new genes were found at the top four ranks and twelve others between ranks 6 and 54 (Table 5.3 and Appendix Tables B.1–B.23). The median rank of the new genes was 9.5. This is a drastic improvement due to the use of multiple annotation sources in contrast to the ranks obtained when using only GO annotations with a resultant median of 133.5.

Figure 5.9 highlights two disease phenotypes: obesity, which had 17 associated genes known before January 2009, and familial glioma of brain, which had seven associated genes. The new gene POMC, which was added to the obesity phenotype in the updated version of OMIM, was found on the third rank. Annotations shared by POMC and the other known disease genes included protein-protein interactions (with AGRP, ENPP1, GHRL, MC3R and MC4R) and the annotation term "obesity" from UniProtKB keywords, which covers POMC and 10 other obesity genes (Appendix Table B.6). The genes ranked first and second, LEP (leptin) and INS (insulin), are also related to obesity (Spiegelman & Flier, 2001) even if they are not among the genes of the specific obesity phenotype in OMIM.

BRCA2, the new gene included into the updated version of OMIM for the glioma of brain phenotype, achieved the first rank of genes functionally related to the disease. BRCA2 showed strong *BioSim* functional similarity to five of the seven previously known genes for glioma of brain. Some of the annotations shared by BRCA2 and the five disease genes are protein-protein interactions (with ERBB2, MSH2 and PTEN), the joint disease association of BRCA2 and DMBT1 to medulloblastoma as well as of BRCA2 and PTEN to prostate cancer in OMIM, and a number of GO and pathway annotations (Appendix Table B.1).

5.11 Summary

This chapter presented the novel method *BioSim* to compute and search for functional similarities of genes and proteins based on diverse annotations such as protein interactions, domain architectures, biological pathways, and disease associations. *BioSim* was evaluated together with four other published methods. All methods are fast to compute and just depend on the number of available annotation terms; thus they can scale well to larger datasets.

phenotype	# genes	new gene	gene description	rank	GO rank	shared annotations
Familial glioma of brain	7	BRCA2	breast cancer 2, early onset	1	102	direct and indirect protein-protein interactions (PPI); same disease, GO, and pathway annotation
Epidermolytic palmoplantar keratoderma	2	KRT1	keratin 1	2	26	direct and indirect PPI; same disease, domain, and GO annotation
Antley-Bixler syndrome	1	FGFR1	fibroblast growth factor receptor 1	2	1	indirect PPI; same disease, domain, GO, and pathway annotation
Cardiofaciocutaneous syndrome	3	MAP2K1	mitogen-activated protein kinase kinase 1	2	16	direct and indirect PPI; same pathway annotation
Folate-sensitive neural tube defects	3	MTHFR	5,10-methylenetetrahydrofolate reductase	2	3	indirect PPI; same GO, and pathway annotation.
Obesity	17	POMC	proopiomelanocortin	3	83	direct and indirect PPI; same GO, pathway and UniProtKB keyword annotation
Autosomal recessive deafness-1A	1	GJB6	gap junction protein, beta 6, 30kDa	3	6	same disease, domain, and GO annotation
Autosomal idiopathic short stature	3	GHR	growth hormone receptor	3	182	direct PPI; same GO annotation
Hypogonadotropic hypogonadism	3	FGFR1	fibroblast growth factor receptor 1	3	1183	direct PPI; same GO, and UniProtKB keyword annotation
Noninsulin-dependent diabetes mellitus	25	PPARG	peroxisome proliferator-activated receptor gamma	4	31	direct and indirect PPI; same disease, domain, and GO annotation
Susceptibility to atypical hemolytic uremic syndrome-1	2	CFI	complement factor I	4	14	indirect PPI; same GO, pathway, and UniProtKB keyword annotation
Noninsulin-dependent diabetes mellitus	25	SLC2A4	solute carrier family 2 (facilitated glucose transporter), member 4	6	424	indirect PPI; same GO, pathway, and UniProtKB keyword annotation
Autosomal recessive deafness	1	GJB3	gap junction protein, beta 3, 31kDa	9	5	same domain, GO, and UniProtKB keyword annotation.
Autosomal recessive dyskeratosis congenita	1	NHP2	NHP2 ribonucleoprotein homolog (yeast)	10	1	direct PPI; same GO annotation
Orofacial cleft 1	1	MTHFR	5,10-methylenetetrahydrofolate reductase	11	7	direct PPI; same GO pathway, and GO annotation; orthologs in same species
Alzheimer disease	9	APP	amyloid beta (A4) precursor protein	11	868	direct and indirect PPI; same GO, and pathway annotation
Susceptibility to atypical hemolytic uremic syndrome-1	2	CFHR1	complement factor H-related 1	19	35	direct and indirect PPI; same domain, and GO annotation
Endometrial cancer	1	MLH3	mutL homolog 3 (E. coli)	28	17	PPI, same UniProtKB keyword, GO and KEGG annotation
Susceptibility to atypical hemolytic uremic syndrome-1	2	CFHR3	complement factor H-related 3	35	2702	direct and indirect PPI, same domain architecture
Mitochondrial neurogastrointestinal encephalopathy syndrome	1	POLG	polymerase (DNA directed), gamma	39	90	PPI, same GO annotation
Colorectal cancer	18	CCND1	cyclin D1	43	378	PPI, indirect PPI, same GO annotations
Osteogenic sarcoma	3	TP53	tumor protein p53	48	806	direct and indirect PPI, same pathway and GO annotation
Mitochondrial complex I deficiency	9	NDUFA11	NADH dehydrogenase (ubiquinone)	54	57	same protein complexes and pathways

Table 5.3: Disease genes recently added to OMIM and identified by the *BioSim* method. The table lists 23 new disease gene associations found between ranks 1 and 54. The table column '# genes' gives the number of known genes associated with the disease phenotype before January 1, 2009. The column 'new gene' contains the symbol of the gene that was added to the phenotype between January and October 2009 and correctly identified by *BioSim*. The columns 'rank' and 'GO rank' give the position of the new gene in the ranking list if all annotations were used or only GO, respectively. The column 'shared annotations' contains a summary of the most specific annotation terms shared by the known genes and the new gene. The detailed list of shared annotations can be found in Apendix Tables B.1–B.23.

In the benchmarks, the use of multiple annotation sources resulted in improved performance of most methods than the use of solely GO annotations. *BioSim* achieved the best performance by consistently ranking functionally related proteins among the top two out of over 18,000 human gene products. *BioSim* in contrast to other scoring methods might be particular useful for applications based on functional similarity when consistent scores are especially desirable, for example, for the quality assessment of protein-protein interactions (Ramírez et al., 2007) and for the clustering of genes or proteins by function (Huang et al., 2007). We also showed how *BioSim* can be applied to discover potential disease genes.

6

Web Portal to Analyze Genes and Proteins

This chapter describes *BioMyn*, a web portal that allows access to the data warehouse presented in Chapter 2, and to the *BioSim* method introduced in Chapter 5. The web portal facilitates the analysis of large sets of genes of proteins that are submitted by the portal users in the context of all the information integrated in the data warehouse. *BioMyn* offers enrichment analysis that are complemented with novel visualization methods to explore the results.

6.1 Introduction

One of the main challenges faced by bioinformaticians today is the development of efficient methods and techniques to interpret the results from high-throughput experiments such as microarrays, yeast two-hybrid, proteomic methods based on mass spectrometry, next-generation sequencing, and RNAi. Such technologies usually generate lists containing hundreds of potentially relevant genes, which need to be further investigated in order to understand their biological significance and their cellular roles. A common approach for the interpretation of results is through data mining of the accumulated biological knowledge, typically by identifying statistically over-represented annotations associated to the set of interesting genes. This method, known as gene set enrichment analysis (Rivals et al., 2007), is based on the assumption that sets of genes responsible for a biological activity are likely to be selected together in an experiment.

Currently, there are dozens of tools that can perform enrichment analysis (see reviews by Khatri & Drăghici (2005) and Huang et al. (2009)); however, most of them are limited to functional categories based on Gene Ontology (GO) (Ashburner et al., 2000), such as BiNGO (Maere et al., 2005), topGO (Alexa et al., 2006), FatiGO (Al-Shahrour et al., 2004), and GoMiner (Zeeberg et al., 2003) just

to name some of the popular tools. Few other enrichment analysis tools available additionally include metabolic and signaling pathways such as GeneMAPP (Salomonis et al., 2007), ArrayXPath (Chung et al., 2005), WebGestalt (Zhang et al., 2005), KEGG-spider (Antonov et al., 2008), SubpathwayMiner (Li et al., 2009), PathExpress (Goffard et al., 2009), GFINDer (Masseroli et al., 2004) and KOBAS (Masseroli et al., 2004). Some of these tools also offer enrichments analysis for disease associations and orthologous groups in other species. Yet few of the available tools allow comprehensive analysis using many other sources of information that may be relevant for the interpretation of results like protein-protein interactions (Vidal et al., 2011; Ideker & Sharan, 2008) and functional similarities (Pesquita et al., 2009). Only DAVID (Huang et al., 2009) implements gene set enrichment analysis using several annotation databases available together with other analysis tools. However, the update cycle of DAVID takes between one and two years while most of the gene and protein annotations are updated several times a year.

Here, we present *BioMyn*, a web portal for the analysis of human gene and protein sets that is based on a large collection of annotations from different biological sources that are integrated in our data warehouse (Table 2.2). Although *BioMyn* focuses on enrichment analysis, it also presents predicted and experimental protein-protein interactions and computes functional similarities to complement the analyses. A wide variety of gene and protein identifier systems are supported to easily submit gene or protein sets. *BioMyn* also offers a powerful search engine based on CompleteSearch (Bast & Weber, 2007) for quickly searching the integrated annotations.

The web portal seeks to improve how researchers analyze, manipulate, share, find, interact, and compare important experimental results more effectively and efficiently. The development of these features was guided from the feedback given by biologists and on the experiences gained from several studies (Ramírez et al., 2007; Ramírez & Albrecht, 2010; Reiss et al., 2011).

6.2 Concept

BioMyn aims to facilitate the bioinformatic analysis of gene or protein sets by offering access to integrated human gene and protein annotations and to powerful data mining tools. The web front-end of *BioMyn* acts as an interface that translates tasks and results to the end-user from the integrative data warehouse and from the data analysis tools, which are written in R (R Development Core Team, 2011).

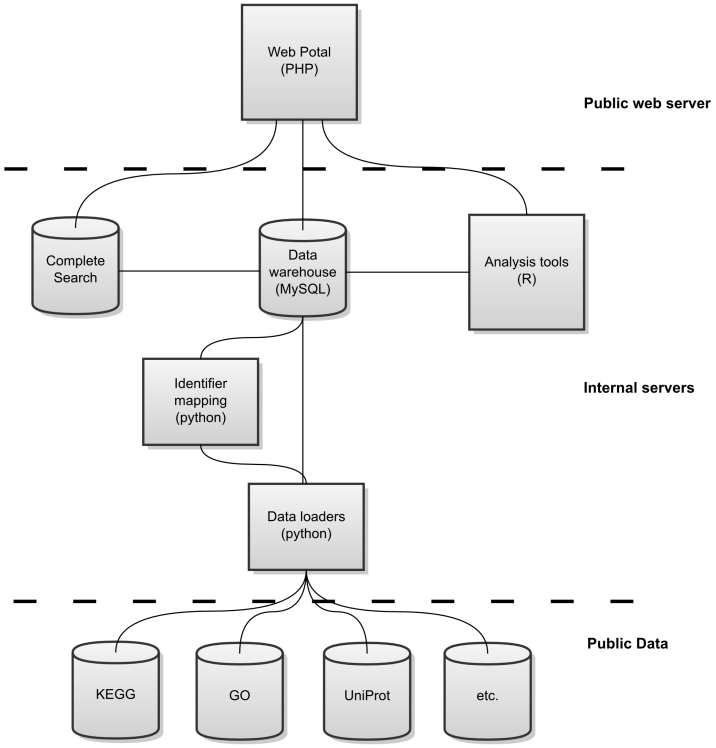


Fig. 6.1: Portal structure overview. The *BioMyn* web portal is divided into several modules that are located in different servers and developed using various programming languages and tools. Central to *BioMyn* is the database warehouse containing the integrated information from different biological data sources. The data warehouse uses the relational database MySQL to efficiently store and search the data. The analysis tools, written in the R language, compute the *BioSim* similarity and the enrichments of genes and proteins based on the integrated data. The CompleteSearch (Bast & Weber, 2007) server allows the creation of complex queries over the integrated annotations. The data warehouse is fed by several programs written using the Python programming language to collect, process and unify the data from the different biological databases. Finally, the web portal, which is developed using PHP and Javascript, allows public access to the different components.

BioMyn is composed of three modules: (*i*) the enrichment analysis module, (*ii*) the protein-protein interaction module, and (*iii*) the functional similarity module. Since *BioMyn* allows researchers to submit their own set of genes or proteins for analysis, an important aspect for the functioning of the site is the mapping of the submitted identifiers to a unified scheme. This is achieved by using the mapping algorithm developed to integrate the plethora of identifiers found in data sources, which was discussed in Chapter 2.

BioMyn also has a powerful search engine based on CompleteSearch (Bast & Weber, 2007) that allows quick searches over the integrated annotations. The results from the queries are protein sets that can be further analyzed with our data mining tools. For instance, using CompleteSearch all human kinases (as

annotated in the Gene Ontology) that are known to be involved in a disease (as reported by OMIM) can be easily retrieved without the need to write SQL commands.

The web portal also presents the integrated annotations stored in the data warehouse on a gene-by-gene or protein-by-protein basis, similar to other websites like GeneCards (Safran et al., 2010) and NextProt (`http://nextprot.org`). This, however, is not the main goal of *BioMyn* and these views exist to browse the results from the analysis tools.

An overview of the structure of web portal is depicted in Fig. 6.1.

6.3 Uploading gene and protein sets

BioMyn analysis starts by uploading a set of human genes or proteins identifiers of any size. The user is not confined to any particular identifier system (see Section 2.2, page 8) but, for unification purposes, all set members are mapped to either the Entrez Gene (Maglott et al., 2007) or UniProtKB (The UniProt Consortium, 2010) identifier systems, as chosen by the user. These are the most widely used identifiers for genes and proteins. Alternatively, the *BioMyn* search engine can be used to obtain a set of genes to analyze all human transcription factors (Section 6.8).

Once a set has been uploaded, the first view is that of a treemap containing the different biological databases integrated into the data warehouse (Table 2.2). Treemaps (Shneiderman, 1992) are visual representations of quantities that use available space efficiently using rectangles. In *BioMyn*, the rectangles have an area proportional to the number of genes from the submitted set that are associated with an annotation in the respective database (Fig. 6.2). The number of genes can be seen by positioning the mouse over any of the rectangles.

By clicking in any of the rectangles, representing the distinct biological databases, the user is taken to a new treemap containing the enrichment analysis results for the selected database, for instance GO biological process or KEGG.

6.4 Enrichment analysis

Enrichment of an annotation term means that the user set contains more genes or proteins annotated with that term than what would be expected from a random

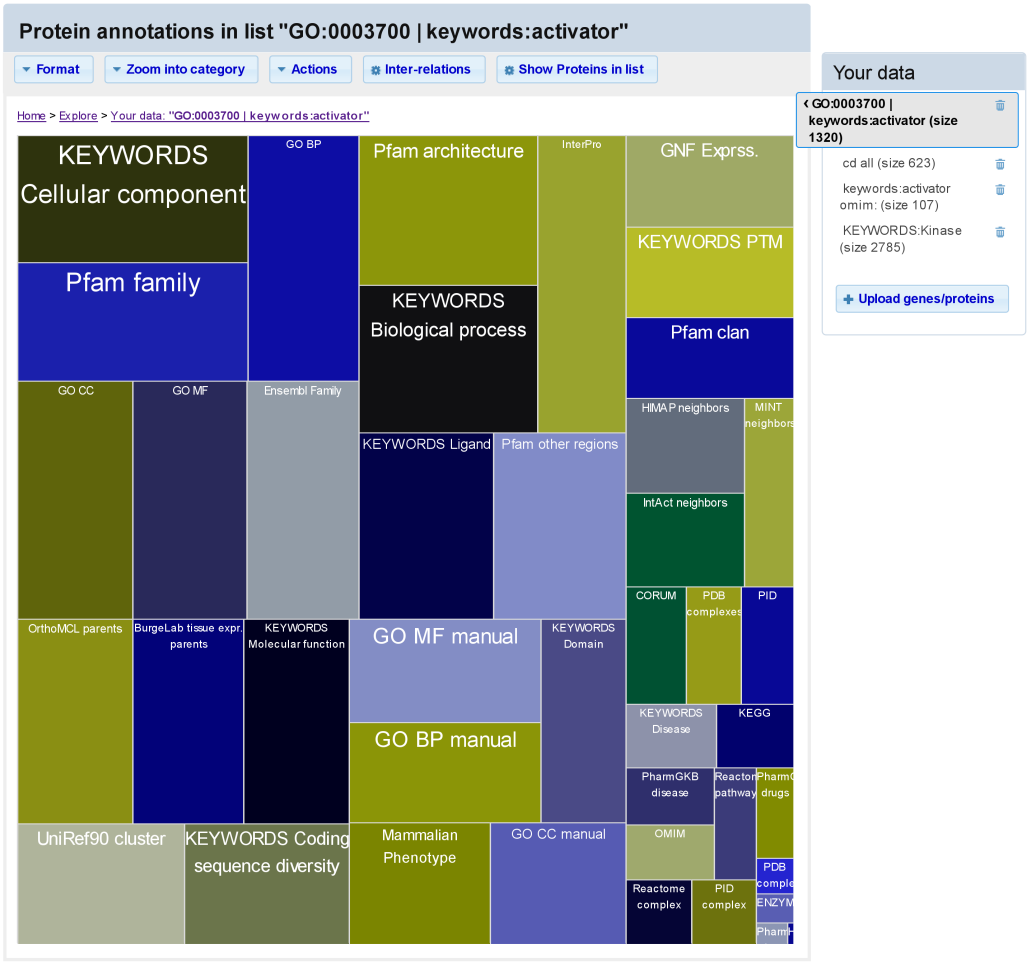


Fig. 6.2: Overview of data sources. After loading a set of genes or protein identifiers, *BioMyn* shows a treemap in which the squares represent each of the biological databases containing an annotation for any of the genes or proteins submitted by the user. The area of the rectangles is proportional to the number of genes or proteins having an annotation in the respective biological database.

selection containing the same number of genes as those in the set. Similarly, depletion means that fewer genes or proteins are associated with that annotation term than randomly expected. The results from the enrichment and depletion test are p-values that indicate the probability of obtaining the observed counts compared to random expectations. *BioMyn* uses the hypergeometric distribution to compute p-values as recommended by Rivals et al. (2007).

The p-value computed for each annotation term as well as the number of genes or proteins expected to be annotated with the term can be seen by directing the mouse pointer over any of the treemap rectangles. The rectangle colors vary from red and orange for those annotation terms that are enriched (overrepresented) in the submitted gene or protein set, to grey for cases that match the random expec-

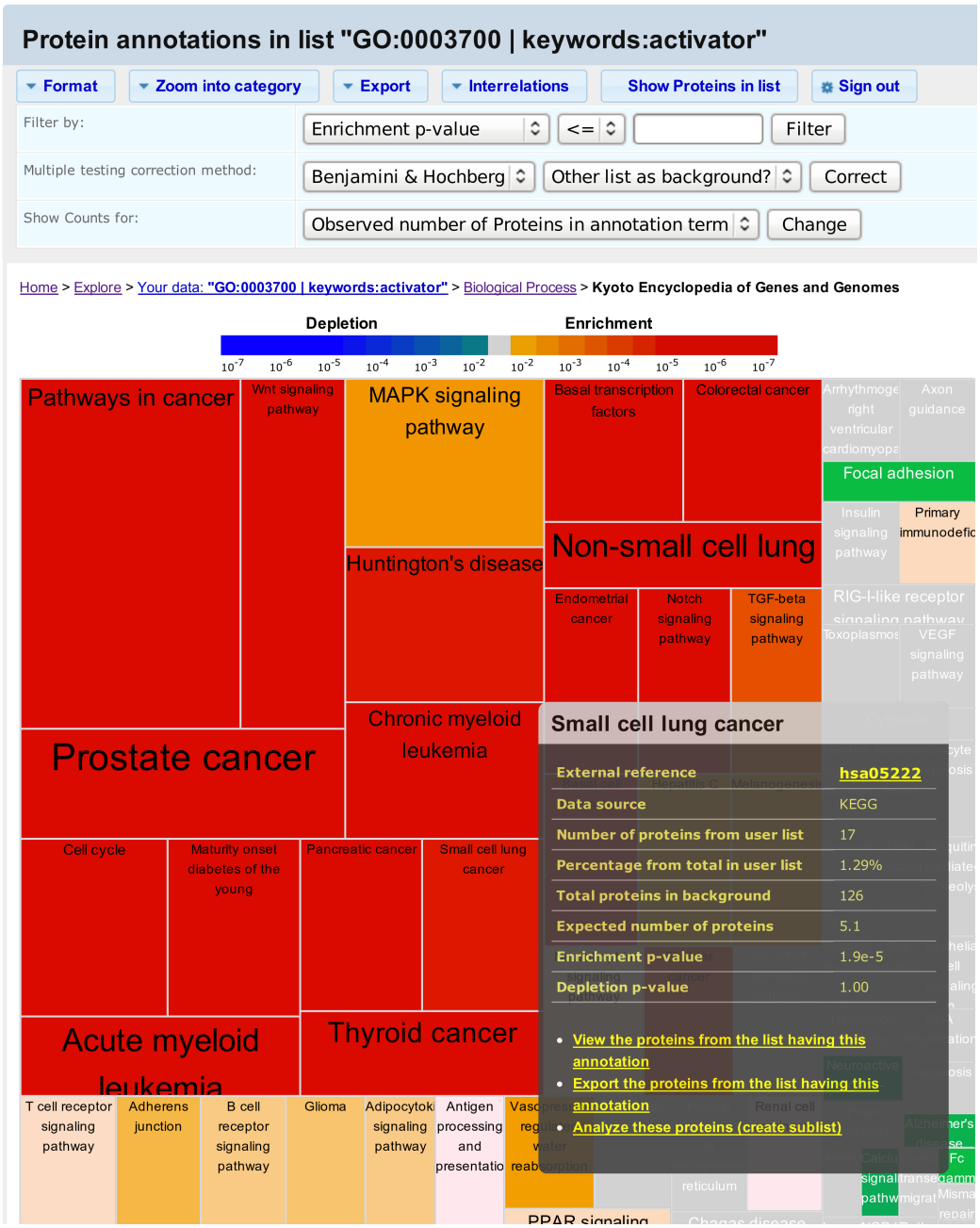


Fig. 6.3: Enrichment analysis. Enriched pathways found in KEGG (Kanehisa et al., 2008) for a selection of 1,320 human transcription factors. The red and orange colors are assigned to enriched pathways, the green and blue colors are assigned to depleted pathways, and the grey color is for cases that are not enriched or depleted. The enrichment and depletion p-values and other information are shown in a pop-up window that appears by hovering the mouse over the treemap rectangles. The link for analyzing a subset of proteins appears at the botton of this pop-up window. The visualization and enrichment options on top of the treemap allow filtering the results by p-value, and by the number of proteins in either the user set or in the background. There is also the option to use another submitted set as a background for the computation of enrichments, and the option to change the rectangle area to be proportional to the expected number of proteins from the set instead of the default number of observed proteins annotated to the respective annotation term.

tation, and finally to green and blue for depleted (underrepresented) annotation terms.

Apart from the treemap, results can also be visualized as a histogram, as a table, or can be exported for further analysis in an electronic spreadsheet file format.

Enrichment analysis options

Adjusting *p*-values

The computed p-values help to decide if the hypothesis that an annotation term is enriched or depleted should be rejected or accepted. However, p-values should be corrected in order to avoid erroneous conclusions when numerous p-values are evaluated as in the case of the enrichment analysis in which p-values are computed for hundreds of annotation terms; this adjustment of p-values is called *multiple testing correction*.

The enrichment analysis of *BioMyn* is accompanied by several multiple testing correction options to adjust p-values. While a given p-value may be appropriate for each individual test, it is not for the set of all tests. The methods offered are: Bonferroni, Holm (Holm, 1979), Benjamini & Hochberg (Benjamini & Hochberg, 1995), and Benjamini & Yekutieli (Benjamini & Yekutieli, 2001). The Bonferroni and Holm procedures lower the p-values to ensure that the overall probability of declaring an enrichment as significant is maintained at some significant level, for example 0.05.

The Bonferroni and Holm methods, however, are considered too conservative for many practical cases in which researchers prefer a small proportion of false positives as a trade-off for getting a larger set of significant hypothesis (Kerr, 2009). As an alternative, the Benjamini & Hochberg (BH) and the Benjamini & Yekutieli (BY) methods control the false discovery rate (FDR), which is the expected proportion of false positive findings among the enrichments declared significant (i.e. rejected null hypotheses). For the BH and BY methods the reported values, analog to the p-values, are called q-values (Storey, 2002). As an example, the expected number of false positives is 1 % for the set of enrichments having a q-value $<= 0.01$.

Selecting a background set

The computation of p-values and q-values is based on the distribution of annotation terms in a background. By default, the computation of p-values uses a background distribution containing all human genes or proteins that have an annotation in the selected database; however, any other background set, chosen by the user, can be used as well. *BioMyn* includes the option to select one of the submitted sets as background instead of the default. Improper use of a background set can invalidate the results. For example, the enrichment analysis of a group of interesting genes from a study that is based on all human kinases should use, as background, this same set of all human kinases; otherwise, the enrichment analysis will identify annotations terms already enriched in the background set. In the case of kinases, these annotation terms are for instance, phosphotransferase function, intra-cellular location, and the participation in signaling transduction pathways.

Generating subsets of genes

The selection of a subset of genes or proteins from a submitted set is useful to further study smaller groups of genes that share an annotation term, for instance, that participate in a particular signaling pathway. Subsets can also be used to filter a submitted set, for example, from a set of proteins, a subset containing only proteins expressed in brain tissues can be created to focus the analysis. In *BioMyn*, subsets are created from annotation terms (Fig. 6.3). Subsequently, the set of genes or proteins having the selected annotation term is automatically loaded into *BioMyn* as a new set that can be further analyzed and explored.

Filtering the results

Often, the enrichment analysis results contain hundreds of annotation terms that render the visualization of the results difficult. Using the option to filter the results, the user can concentrate on certain parts of the visualization. The filtering options control which annotation terms are shown based on enrichment or depletion p-value, number of genes or proteins from the submitted set annotated with the term, and the number of genes or proteins in the background annotated with the term (Fig. 6.3).

(A)

Protein A	Protein B	IntAct	DIP	HPRD	MINT	BioGRID
CTNB1_HUMAN (CTNNB1) Catenin beta-1 -Entrez id: 1499 -UniProtKB acc: P35222	TF7L2_HUMAN (TCF7L2) Transcription factor 7-like 2 -Entrez id: 6934 -UniProtKB acc: Q9NQB0					
HDAC1_HUMAN (HDAC1) Histone deacetylase 1 -Entrez id: 3065 -UniProtKB acc: Q13547	SIN3A_HUMAN (SIN3A) Paired amphipathic helix protein Sin3a -Entrez id: 25942 -UniProtKB acc: Q96ST3					
RUVB2_HUMAN (RUVBL2) RuvB-like 2 -Entrez id: 10856 -UniProtKB acc: Q9Y230	RUVB1_HUMAN (RUVBL1) RuvB-like 1 -Entrez id: 8607 -UniProtKB acc: Q9Y265					

(B)

Protein A	Protein B	HiMAP	I2D	PIPs	STRING	HomoMINT
RUVB2_HUMAN (RUVBL2) RuvB-like 2 -Entrez id: 10856 -UniProtKB acc: Q9Y230	RUVB1_HUMAN (RUVBL1) RuvB-like 1 -Entrez id: 8607 -UniProtKB acc: Q9Y265					
SMAD4_HUMAN (SMAD4) Mothers against decapentaplegic homolog 4 -Entrez id: 4089 -UniProtKB acc: Q13485	SMAD1_HUMAN (SMAD1) Mothers against decapentaplegic homolog 1 -Entrez id: 4086 -UniProtKB acc: Q15797					

Fig. 6.4: Protein-protein interactions. *BioMyn* reports all experimentally verified and predicted protein-protein interactions known between the set of genes or proteins submitted by the user. (**A**) View of experimentally derived protein-protein interactions. (**B**) View of predicted protein-protein interactions.

Visualizing ancestor terms from ontologies

The data warehouse contains different types of annotations including ontological annotations characterized by a hierarchical structure (See Section 2.4.1, page 18). In cases such as GO, the visualization of the enrichment analysis is hindered by the sheer amount of relations and terms found in such hierarchical structure. Thus, to avoid cluttering the view, all ancestor terms are hidden by default. This feature can be changed easily if needed.

Also, in the case of the GO, only the so called 'slim' terms are shown by default. The GO slim (`http://www.geneontology.org/GO.slims.shtml`) is a subset of GO containing only few broad categories, manually chosen for summarizing the GO annotations.

6.5 Protein-protein interactions

The protein-protein interactions section of *BioMyn* lists all experimental and predicted interactions that occur between the genes or proteins in the submitted set. Figure 6.4 shows an example of the results where each interaction appears in a separate row. The columns list the databases providing the interaction informa-

tion. Each dot represents an experiment supporting the interaction or, in the case of a prediction, each dot indicates a different evidence. The dots are linked to the source database where further details of the interaction can be found. Those interactions having the highest number of experimental support or, in the case of predictions, with the highest number of methods predicting the interaction are placed on top. The list of interactions can be easily exported for further analysis to Cytoscape (Smoot et al., 2011) or other network visualization software.

Currently, *BioMyn* presents experimental interactions from BioGrid (Stark et al., 2011), DIP (Salwinski et al., 2004), HPRD (Prasad et al., 2009), IntAct (Kerrien et al., 2007) and MINT (Ceol et al., 2010); and predicted interactions from HiMap (Rhodes et al., 2005), HomoMINT (Persico et al., 2005), I2D (Brown & Jurisica, 2005), PIPs (McDowall et al., 2009), and STRING (Jensen et al., 2009).

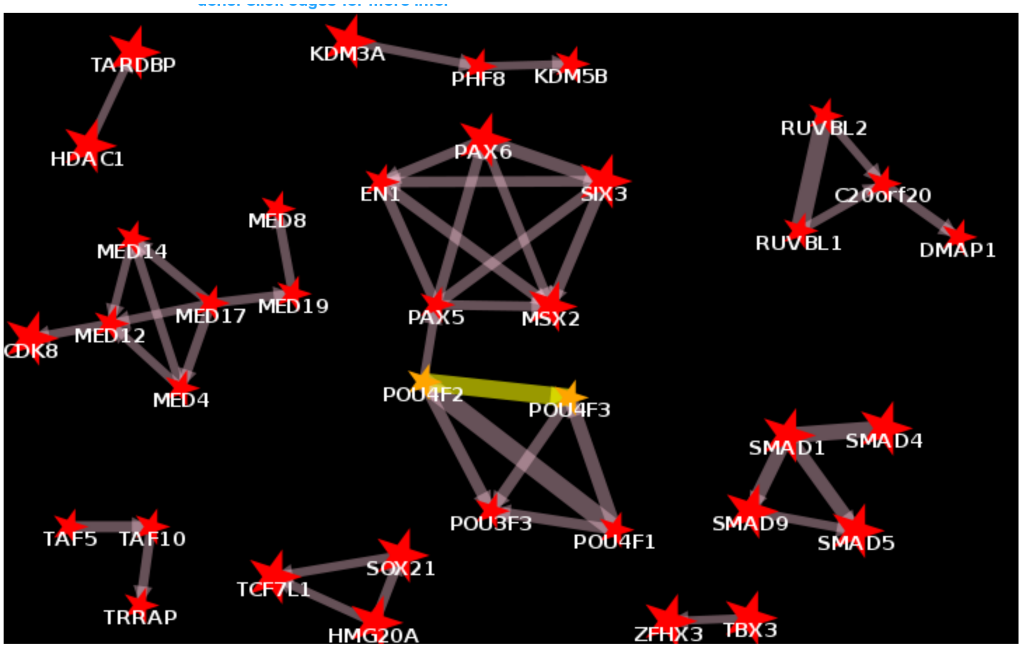
6.6 Functional similarity

The functional similarity module of the portal uses the *BioSim* method introduced in the previous chapter to identify protein pairs having similar annotation terms from the user submitted set. When using this module, a network is shown containing the top twenty proteins having the highest functional similarity (Fig. 6.5). Besides reporting the functional similarities, *BioMyn* also facilitates the navigation through the annotation terms shared between genes or proteins (Fig. 6.6). This is of importance to let the researchers understand and assess the similarities found.

The functional similarity scores can be downloaded for further analysis. For example, the results can be clustered using the Cytoscape plugin clusterMaker (`http://www.cgl.ucsf.edu/cytoscape/cluster/clusterMaker.html`) and GLay (Su et al., 2010).

6.7 Search engine

BioMyn uses CompleteSearch (Bast & Weber, 2007) to index its large catalog of annotations. The search engine accepts full-text query strings related to annotations terms, and gene and protein identifiers, symbols and synonyms. Searches can be narrowed by adding contextual search terms that are separated from the search query by a colon. For example, to search for the pro-insulin gene using its



POU domain, class 4, transcription factor 3 (POU4F3)
Entrez id: 5459 -UniProtKb acc: Q15319 - View all annotations
POU domain, class 4, transcription factor 2 (POU4F2)
Entrez id: 5458 -UniProtKb acc: Q12837 - View all annotations
BioSim similarity value: $10^{-244.019}$

Annotation category	Number of shared annotations terms	Coverage of most specific annotation
component	7	5448
domain	13	3
neighbor	37	7
phenotype	4	696
function	5	664
orthology	236	5782

Fig. 6.5: Functional similarity network. The genes or proteins from the submitted set having the highest *BioSim* functional similarity score are shown as a network. In this network, the edges represent genes or proteins from the user submitted set and the edges indicate a functional similarity. Thicker edges indicate a stronger functional similarity. By clicking over any of the edges, a summary of the shared annotation terms is shown. The figure shows the similarities found for a set of human transcription factors. A summary of the shared annotation terms can be seen in the lower part of the image for the proteins: POU domain, class 4 transcription factor 2 (POU4F2), and the POU domain, class 4 transcription factor 3 (POU4F3). Further information about the shared annotation terms is found by following the links found in this summary as shown in Fig. 6.6.

symbol, the query is: 'symbol:INS'. The built-in auto-complete function helps to find the right term and context names while typing.

Search terms can be combined using the `OR` operator represented by the pipe '|' character, the `AND` operator which is assumed by default, and the `NOT` operator represented by the minus '-' character.

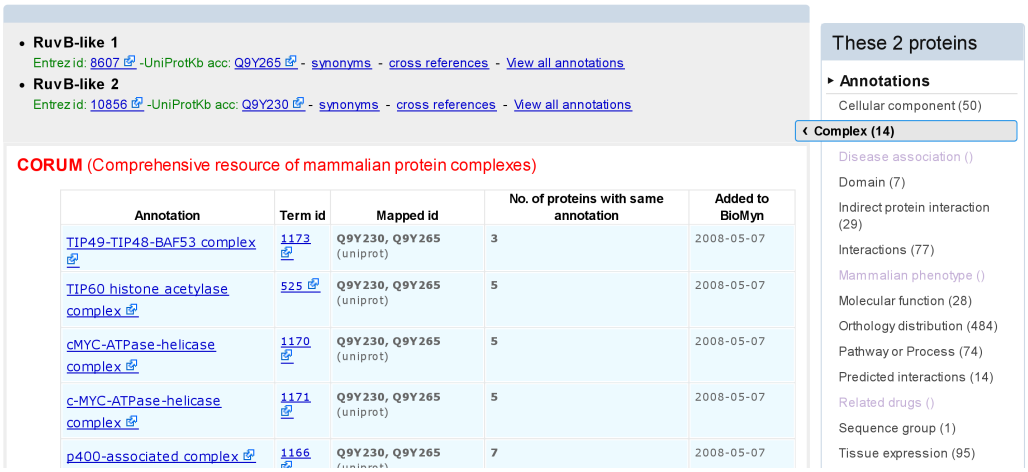


Fig. 6.6: Shared annotations between two proteins. This figure shows all protein complexes in which the transcription factors RuvB-like1 and RuvB-like2 participate, according to the CORUM database (Ruepp et al., 2008). The hypertext links in the 'Annotation' column point to the original biological databases. Besides complexes, other shared annotations can be seen by following the links in the right panel.

The search results are complemented with additional information classified into: 'Pathway/Process', 'Function', 'Component', and 'Disease'. These information contain frequent annotation terms found within the search results.

The set of proteins matching the search query can be automatically loaded as a user set for which all the analysis tools presented before can be used. In this way, the search engine serves as a second entry point to *BioMyn* analysis tools and can facilitate the bioinformatic analysis of available annotations.

6.8 Case study

We used the search engine to obtain a set of human transcription factors from the data warehouse using the annotations found in two database sources: GO and UniProtKB Keywords. The GO term for transcription factor is: *sequence-specific DNA binding transcription factor activity* that corresponds to the identifier: GO:0003700 (`http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0003700`). In UniProtKB Keywords, the term for transcription factor is *Activator* (`http://www.uniprot.org/keywords/KW-0010`). Using the search query 'GO:0003700 | keywords:activator' we obtained a set of 1,320 proteins. This set was subsequently loaded into *BioMyn* for further analysis.

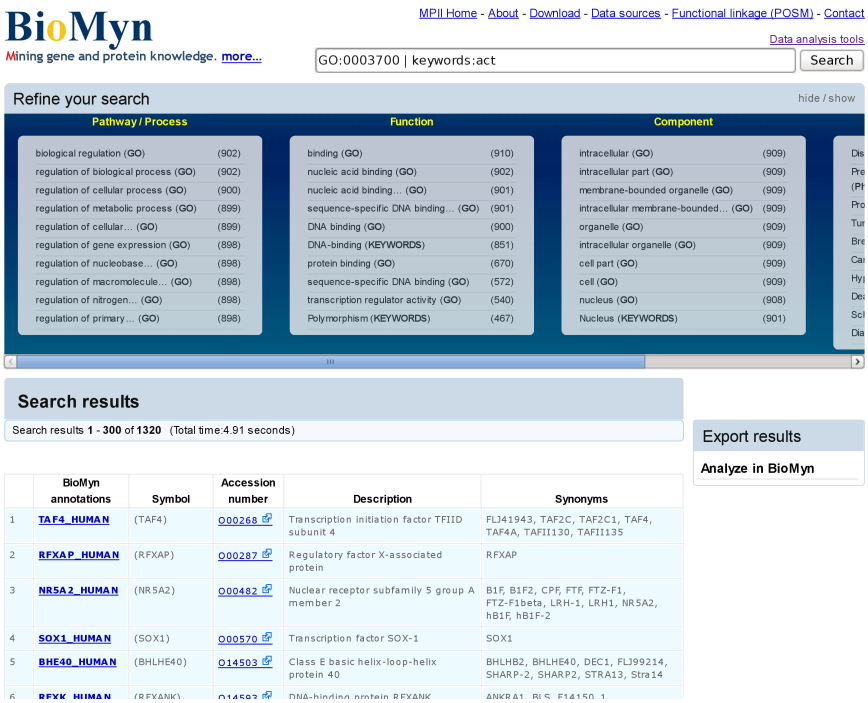


Fig. 6.7: *BioMyn* search Engine Search results for the query 'GO:0003700 | keywords:activator' used for retrieving transcription factors. 'GO:0003700' is the GO identifier for this type of proteins that are named 'activators' in UniProtKB Keywords. The pipe character is the `OR` operator. The result list shows the first 300 proteins that match the search query. The annotations for each of these proteins can be seen by following the link in the column 'BioMyn annotations'. Alternatively, a new set can be submitted following the link on the right called 'Analyze in *BioMyn*'. The boxes above the results table contain additional information useful to refine the search. Each box lists up to ten annotations terms, sorted in decreasing order by the number of proteins having such annotation term from the search results. This number is shown in parentheses after the annotation term name and source. Clicking over any of the annotation terms found in the rectangles adds the respective term to the search query.

We analyzed the GO biological processes associated with the set of transcription factors. The default background was used since the set was obtained from all human proteins having GO or UniProtKB Keyword annotations. The Benjamini & Hochberg method was used to correct for multiple testing instead of the more stringent Bonferroni or Holm methods. Electronic annotations were excluded by selecting the annotations from the *GO Biological Process manual*. Using the default GO slim visualization of *BioMyn* we could identify the following enriched GO terms: *transcription*, *chromosome organization* and *embryo development*. The GO term *transcription* should be obviously enriched because all proteins in the set are transcription factors. Most other GO slim terms appear depleted, for example: *signal transduction*, *response to stress* and *immune system process*.

Besides GO biological processes we also studied the protein domain composition of the set of transcription factors. Transcription factors are usually charac-

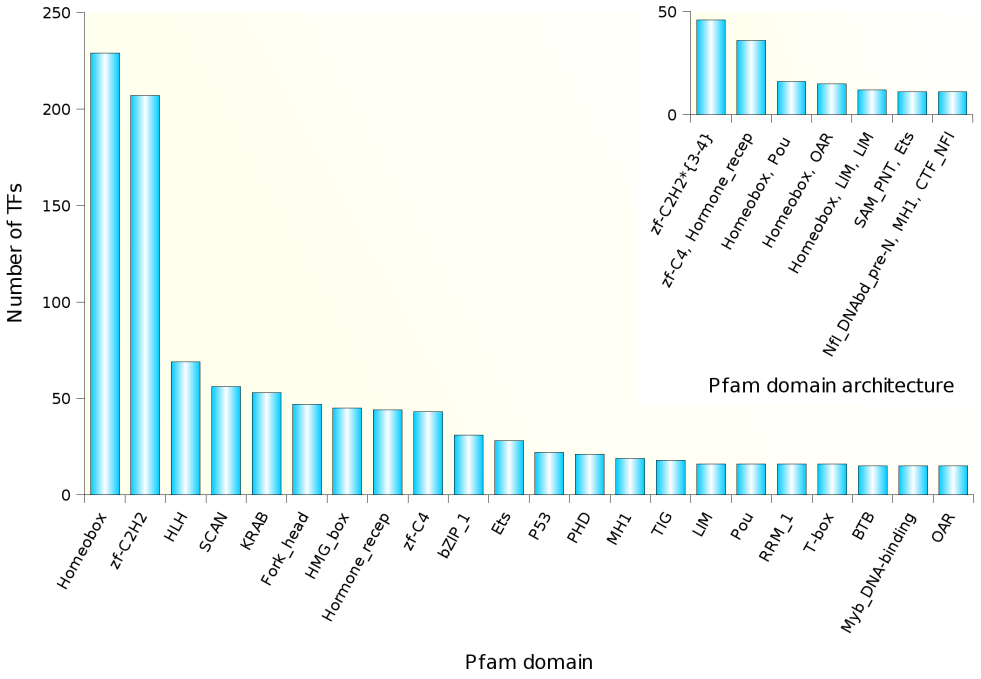


Fig. 6.8: Transcription factor domains. Enriched Pfam domains and Pfam domain architectures for human transcription factors. Here, we show the enriched Pfam domains found in a set of 1,320 human transcription factors. We compared the results with common Pfam domain architectures (domain combinations) found for the same set.

terized by several DNA binding domains including the Helix-Loop-Helix (HLH) domain (Massari & Murre, 2000), the Homeobox domain (Levine & Hoey, 1988), and several zinc finger domains (Wolfe et al., 2000; Laity et al., 2001). We used *BioMyn* to identify enrichments for Pfam protein domain annotations and for Pfam architectures (Finn et al., 2008). Pfam domain architectures correspond combinations of domains found in multi-domain proteins. In many cases there are different domains within a protein, but is also not rare to find repetitions of the same domain within a protein.

Our analysis of transcription factor domains found the previously mentioned domains and others that had been reported elsewhere (Vaquerizas et al., 2009) (Fig. 6.8). Interestingly, the analysis of domain architectures identified several groups of domains that often appear together like the zF-C4 and Hormone receptors which are only found individually in eight cases out of 44. Similarly, the POU and OAR domains are always found in conjunction with the Homeobox domain and never independently. The zinc finger C2H2 often appears in groups of 3 or 4 repetitions in 22 % of the proteins having that domain (Fig. 6.8).

We generated a subset containing 74 transcription factors belonging to the enriched GO term *chromosome organization* to investigate the protein domain composition characterizing the subset. For the enrichment analysis of this subset we also selected Pfam domains, the Benjamini & Hochberg method for multiple testing correction, and as background we chose the set containing the original 1,320 transcription factors. The results show that transcription factors annotated as participating in *chromosome organization* are depleted in the Homeobox domain, one of the most commonly found domain of transcription factors in our analysis (Fig. 6.8). However, Levine & Hoey (1988) argues that the Homeobox domain is mostly involved in developmental processes which would explain the scarcity of transcription factors in *chromosome organization*.

The subset of transcription factors was found enriched with the SET and PHD finger-domains. Proteins having the SET domain methylate lysine amino acids from histones while the PHD finger domain apparently is a protein-protein interaction domain that binds tri-methylated lysines on target histones. Ten of the eleven transcription factors having the PHD domain also contain other domains like MOZ_SAS whose function is unclear and zf-CXXC that binds to non-methylated CpG dinucleotides. We also found that eight of the eleven proteins having the PHD domain often interact with each other.

These results may point to the formation of protein complexes composed of PHD-containing proteins responsible for the activation of genes that are regulated by the lysine methylation of histones.

6.9 Summary

The *BioMyn* web portal allows a multiple-perspective analysis of human gene and protein sets based on the integrated annotations from dozens of biological databases. The goal of the web portal is to aid in the interpretation of such sets that are often the result of high-throughput experiments. Using *BioMyn*, the visualization of enrichment analysis for user-submitted sets is straightforward, allowing the researcher to concentrate on the analysis of the results rather than acquiring the integrated information.

BioMyn is useful for exploratory analysis of the submitted sets and for the generation and testing of hypotheses. We showed how the portal can be used to study different aspects of a set of proteins, including enrichment analysis of annotations, protein-protein interactions (experimental and predicted) and functional similarity relations based on the method presented in the previous chapter. The

offered analyses implement novel uses of visualization techniques and incorporate fast and easy-to-use searching capabilities over most of the annotations known for human genes and proteins.

BioMyn results can be easily saved for further analysis using spreadsheets, R (R Development Core Team, 2011), and Cytoscape (Smoot et al., 2011) to complement the analysis.

Although *BioMyn* has not been officially released yet, it receives about twenty daily visitors that are mostly browsing the portal. A manuscript describing *BioMyn* is in preparation.

Currently, the web portal most similar to *BioMyn* is the popular site DAVID (Huang et al., 2007). An advantage of DAVID is the various model organisms for which integrated annotations are available and some additional analysis modules offered. However, in my opinion our web portal is more user-friendly, provides better visualization and navigation tools, offers unification for genes and proteins (DAVID only unifies gene annotations), and allows the creation of subsets for further analysis. *BioMyn* also has more up-to-date annotations as DAVID is only updated annually.

7

Conclusions

This closing chapter summarizes the work presented in this thesis and presents the perspectives and future directions. Additionally, this chapter shows several ongoing projects within our research group that are supported by the data warehouse and related data mining and visualization tools. Finally, this dissertation concludes with a perspective about current data integration endeavors and future outcomes.

7.1 Summarizing remarks

Current experimental and computational high-throughput methods are generating biological data at unprecedented speeds. The possibility to interrogate whole cellular systems at the level of genes, proteins and metabolites has spawned cell-wide studies focusing on the dynamics and interactions of cellular elements. Such comprehensive studies are part of what is known as *systems biology*, a relatively new branch of biology aimed at understanding complex cellular processes. Scientists have used high-throughput technologies to accumulate a tremendous amount of data that, in order to be useful, needs to be analyzed and interpreted in the context of all other information available. To ease such analysis, integrative repositories aggregating and merging biological knowledge are required.

The construction of such integrative repositories, however, has proven to be a challenging task due to the inherent complexities of the data found in biological databases and to the spread of knowledge in multiple places. Moreover, such repositories need to offer appropriate data mining and analysis tools that are easily accessible through simple interfaces to promote their use and adoption by the biological and medical researchers.

The work presented in this dissertation had addressed these challenges by (*i*) creating an integrative repository containing millions of annotations from over 30 major molecular biology databases related to human genes and proteins, (*ii*) proposing and validating a data mining method to identify similarly annotated gene products, and (*iii*) developing a web portal to analyze large sets of genes submitted by users using the integrated annotations in the repository.

The integrative repository uses a data warehousing model in which information is periodically downloaded from more than 30 biological databases and locally stored in a relational database for efficient searches. The integrated biological databases contain information about co-expressed genes, disease associations, drug associations, metabolic and signaling pathways, molecular functions, orthologous species groups, phenotype associations, protein complexes, protein domain classifications, protein-protein interactions, predicted protein-protein interactions, sequence clusters, sub-cellular locations and gene expression in tissues.

The integrated data contained in the data warehouse was used to support two important studies presented in this thesis. In the first study, we assessed the reliability of human protein-protein interactions either predicted by computational methods or produced by the yeast two-hybrid high-throughput method (Ramírez et al., 2007). For this study we used the information from several protein-protein interaction sources stored in the data warehouse together with protein domain information and Gene Ontology associations that were used to assess the interaction data. The results revealed less reliable interactions from high-throughput experiments than from computational predictions and manually curated interactions. We also found that highly reliable interactions were those reported in at least three different experiments. This comprehensive study was highlighted in the editorial section of the *Proteomics* journal where it was first published and is frequently cited in studies discussing the reliability of yeast two-hybrid results. The predicted and experimental interactions collected into the data warehouse for this study have been instrumental in subsequent studies from our research group (Schlicker et al., 2007; Tress et al., 2007; Schlicker et al., 2010; Blankenburg et al., 2009b; Kacprowski, 2010) and the new topological measure introduced in Chapter 3 *shared neighbors* was integrated as part of the Cytoscape plug-in called *NetworkAnalyzer* (Assenov et al., 2008).

In a second study, we combined protein-protein interactions with protein functions and signaling pathways contained in the data warehouse to predict an important type of proteins involved in signaling cascades called scaffold proteins (Ramírez & Albrecht, 2010). Before this study, the few scaffold proteins known had been found by chance through direct methods. Our analysis is considered

the first attempt to estimate the number and abundance of these proteins in the human proteome (Alexa et al., 2010).

Having a solid infrastructure for storing integrated information allowed us to devise a computational method for the discovery of biological relationships based on the similarity of gene and protein annotations. This method, called *BioSim*, overcomes the disadvantages associated to other similar methods that are only available for genes and proteins annotated with GO terms (**?**). Our results demonstrated that the use of annotations from multiple sources drastically improved the performance of our method. Using this method we were able to accurately identify known disease-gene associations, thus opening the possibility for future prediction of disease-causing genes.

Finally, we set up a web portal to offer public access to the data warehouse and to the *BioSim* method. This web portal is called *BioMyn* and can be reached via the following URL: `http://biomyn.de`. The *BioMyn* web portal is optimized for the analysis of large sets of genes or proteins submitted by the portal users. *BioMyn* allows enrichment analysis based on the integrated data and is complemented by novel visualization methods to explore the results. The web portal offers tremendous advantages to biological and medical researchers that are now able to concentrate directly on the analysis of their data rather than acquiring the integrated information. Sets of genes or proteins can be easily imported to *BioMyn* and the different analysis can be also easily exported and saved. The web portal has been successfully used to analyze a list of candidate genes obtained in a RNA interference screen targeting kinases of cultured cells infected with hepatitis C virus (Reiss et al., 2011).

In summary, the data warehouse presented in this dissertation is a powerful resource that facilitates large-scale analyses of data. To support this claim, we described two important studies including the assessment of predicted protein-protein interactions (Ramírez et al., 2007) and the computational discovery of scaffold proteins (Ramírez & Albrecht, 2010). The integrated information of the data warehouse allowed us to develop a new method to search similarly annotated proteins likely to be functionally related (**?**). Lastly, we developed the web portal *BioMyn* to facilitate the analysis of large sets of genes. Using *BioMyn*, the interpretation and analysis of system-wide results that take into account the accumulated biological knowledge is greatly simplified by the access to search and mining methods based on the integration of molecular biology databases.

7.2 Perspectives

Ongoing projects

Several ongoing research projects are currently exploiting the practical utility of the data warehouse.

- The *BioSim* method is being used to assess a large protein-protein interaction network obtained to investigate the exocytosis of presynaptic vesicles in neurons. This is a cooperation with the Max Planck Institute for Biophysical Chemistry in Göttingen, Department of Neurobiology.
- The enrichment analysis tools from *BioMyn* are used to investigate candidate proteins obtained from an interference RNA screen involving cells infected with the Dengue virus. This project is a cooperation with the Department for Molecular Virology at the University of Heidelberg.
- *BioMyn* is used to run exploratory analysis on genes associated to Crohn's disease by identifying enriched and depleted cellular pathways. This is a cooperation with the University Hospital of Kiel supported by the National German Research Network (NGFN) for environmental diseases.

Enhancements to the data warehouse and methods

An avalanche of biological data is expected to arrive in the following years, mostly from inexpensive next-generation sequencing technologies (Africa, 2010; Metzker, 2010). Also, the completion of the human interactome using improved yeast two-hybrid methods is foreseen to reach near completion in the next few years (Vidal et al., 2011). Meanwhile, new molecular knowledge, specially regarding gene and protein regulation, has resulted in the creation of novel biological databases devoted to epigenomics (Fingerman et al., 2011; Turinsky et al., 2011), microRNA targets (Yang et al., 2011; Kaya et al., 2011), and transcription factor binding sites (Portales-Casamar et al., 2010; Yamashita et al., 2010).

These new sources of information will need to be revised and integrated into the data warehouse in order to keep it up-to-date. Additionally, the abundant information available for model organisms such as mouse, yeast and fly, need to be added to the data warehouse. Besides facilitating the analysis and interpretation of the high-throughput data from these organisms, this new information will also permit us to transfer annotations between organisms. Such future steps are

simplified by the database structure, which allows to add new information easily. Currently, the inclusion of mouse data is underway.

The methods introduced in Chapter 3 to assess protein-protein interactions can now be improved by the *BioSim* method. Using *BioSim*, protein-protein interactions can be evaluated using data sources integrated in the data warehouse. This will increase the scope of the methods and improve the reliability of the assessments.

The interaction within available services developed in our group at the Max Planck for Informatics, such as EpiExplorer (`http://cosgen.bioinf.mpi-inf.mpg.de`) and FunSimMat (Schlicker et al., 2010), is another interesting area for further development. EpiExplorer allows the quick analysis of a large number of genetic features from genomic regions submitted by the users. To complement EpiExplorer results, the sets of genes located in regions interesting to the user can be automatically imported into *BioMyn* to explore the associated annotations for those genes. FunSimMat allows the user to obtain a list of proteins probably related to a given disease. These results could be directly analyzed using the different tools from the *BioMyn* web portal. At the moment, this integration is being prepared in our research department.

In order to extend the usage of the data warehouse, the integration with popular stand-alone software such as Cytoscape (Smoot et al., 2011) can be provided. Mike Wininger, a summer student at the MPII, already developed a plug-in for Cytoscape that connects to our servers, collect annotations from the data warehouse, and shows this information in the Cytoscape window. This plug-in can be further developed to retrieve and show enrichments for groups of genes as well as *BioSim* similarities from our servers. More generally, programmatic access to the data warehouse can be added using common protocols like SOAP, REST and PSICQUIC (Aranda et al., 2011).

Finally, *BioMyn* still needs to be exposed to more researchers who will provide further feedback for improvement. For this, a manuscript presenting the web portal is being prepared.

Wider perspective of the data integration field

Data integration will continue to be an issue in the upcoming years, mostly motivated by the growing demand of studies that require access to multiple sources of information. Although diverse methods to integrate biological information exist, prevalent difficulties discussed in this dissertation related to data format, location,

availability, and content hamper integrative solutions. An important limitation of biological data integration is that it is virtually impossible to construct a definitive integrative repository able to capture all the specific aspects and details of the data maintained by biological databases. Therefore, integrative repositories necessarily need to make choices and ignore some of the available information to merge disparate data sources. Nevertheless, it is feasible to integrate a large fraction of the available information into solutions that address specific problems such as the one presented here.

Researchers working in data integration actively encourage the transition towards a Semantic Web (http://www.w3.org/2001/sw/) (Berners-Lee et al., 2001) for the life sciences (Ruttenberg et al., 2009; Neumann, 2005; Goble & Stevens, 2008). Under the Semantic Web, information needs to be formated using standard languages that encourage the clear definition of relations that are structured using ontologies. Such standardization of the information already facilitates integration by homogenizing the diverse data models available. Furthermore, current standards used in the life sciences such as the Biological Pathway Exchange (BioPAX) Demir et al. (2010) request that external identifiers for different reference systems should be provided along with the semantic data. This will greatly facilitate the identification of common biological entities such as genes or proteins.

The Semantic Web will allow computers to interpret and make assertions of the semantic data and, in turn, improve current search methods based on the matching of keywords. For this, a number of software is already available such as the query language SPARQL (Prud'hommeaux & Seaborne, 2008) and semantic reasoners (software able to make logical inferences) such as Pellet (`http://pellet.owldl.com/`), and Jena (`http://openjena.org`).

However, for the life sciences Semantic Web to work, databases should provide the information using a semantic format based on the Resource Description Framework (RDF) or derived languages such as the Web Ontology Language (OWL) and the aforementioned BioPAX (Ruttenberg et al., 2009). This content should be delivered through standardized web services. In practice, this means that many biological databases will have to program new software and transform their data in order to participate, something that is unlikely to happen in the near future. Furthermore, the use of the semantic formats to share the data needs to be carefully controlled. As discussed in Section 2.2, page 13, data currently available in such formats is hard to use because of deviations from the standard that are specific to each data provider.

To circumvent these hurdles, workflow and view integration software as well as data warehouses like *BioMyn* have developed specific modules to access the available information using the current format. Naturally, the availability of semantic data will simplify much of the integration processes but, until such data becomes widely accessible, the use of the information in its present form is the only option.

However, data warehouses such as *BioMyn* are well prepared to make a quick transition towards semantic data analysis. In fact, the organization of the database schema of our data warehouse resembles the structure of one of the cornerstones of the Semantic Web, the Resource Description Framework (RDF) used to store semantic data. The RDF format uses expressions composed of a *subject*, a *predicate* and an *object* for the conceptual description of the information. The database schema of *BioMyn* contains the following matching tables to elaborate RDF expressions: `Molecule`, `Participant` and `MolecularConcept`. Because of the performed unification steps, the data already present into our data warehouse is probably more amenable for semantic searches than some semantic data found somewhere else.

In conclusion, I believe that a transition towards a Semantic Web in the life sciences is slowly happening. Data warehouses, as the one presented in this thesis, are in an advantageous situation because they can easily adapt to semantic queries and analyses by converting the information they already store into a semantic format. Thus, such data warehouses don't need to wait for other biological databases to provide their contents in the appropriate semantic format, instead they can quickly start using the available software to run semantic searches and analysis.

Bibliography

Aerts, S., Lambrechts, D., Maity, S., Loo, P. V., Coessens, B., Smet, F. D., Tranchevent, L.-C., Moor, B. D., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnol*, 24(5), 537–544.

Africa, W. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 1061–73.

Al-Shahrour, F., Díaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4), 578–80.

Albrecht, M., Huthmacher, C., Tosatto, S. C., & Lengauer, T. (2005). Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21 Suppl 2, ii220–ii221.

Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13), 1600–7.

Alexa, A., Varga, J., & Reményi, A. (2010). Scaffolds are 'active' regulators of signaling modules. *The FEBS journal*, 277, 4376–4382.

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M. J., Dumontier, M. R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J. P., Parker, B., Pintilie, G., Pirone, R., Salama, J. J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B. F., & Hogue, C. W. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*, 33, D418–D424.

Alibés, A., Yankilevich, P., Cañada, A., & Díaz-Uriarte, R. (2007). IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC bioinformatics*, 8, 9.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17), 3389–3402.

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

Antonov, A. V., Dietmann, S., & Mewes, H. W. (2008). KEGG spider: interpretation of genomics data in the context of the global gene metabolic network. *Genome biology*, 9(12), R179.

Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S. L., Ceol, A., Chautard, E., Dana, J. M., De Las Rivas, J., Dumousseau, M., Galeota, E., Gaulton, A., Goll, J., Hancock, R. E. W., Isserlin, R., Jimenez, R. C., Kerssemakers, J., Khadake, J., Lynn, D. J., Michaut, M., O'Kelly, G., Ono, K., Orchard, S., Prieto, C., Razick, S., Rigina, O., Salwinski, L., Simonovic, M., Velankar, S., Winter, A., Wu, G., Bader, G. D., Cesareni, G., Donaldson, I. M., Eisenberg, D., Kleywegt, G. J., Overington, J., Ricard-Blum, S., Tyers, M., Albrecht, M., & Hermjakob, H. (2011). PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nature Methods*, 8(7), 528–529.

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., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1), 25–9.

Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T., & Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2), 282–4.

Bader, G. D. & Hogue, C. W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnol*, 20, 991–997.

Bader, J. S., Chaudhuri, A., Rothberg, J. M., & Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnol*, 22, 78–85.

Bahcall, O. (2007). Nature Milestones in DNA technologies, Milestone 15: BLAST-off for genomes. *Nature Rev Genet*, 8, S14–S15.

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Res*, 28(1), 304–305.

Barabasi, A. L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.

Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Rev Genet*, 5(2), 101–13.

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

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. a., Phillippy, K. H., Sherman, P. M., Muertter, R. N., Holko, M., Ayanbule, O., Yefanov, A., & Soboleva, A. (2010). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res*, 39(November 2010), 1005–1010.

Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I., & Wrana, J. L. (2005). High-throughput mapping of a dynamic signaling network in mammalian cells. *Science*, 307(5715), 1621–1625.

Bast, H. & Weber, I. (2007). The CompleteSearch Engine: Interactive, Efficient, and Towards IR & DB integration. In G. Weikum (Ed.), *CIDR 2007 : 3rd Biennial Conference on Innovative Data Systems Research* (pp. 88–95).: VLDB Endowment.

Ben-Hur, A. & Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. *Bioinformatics*, 21 Suppl 1, i38–46.

Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., & Devignes, M.-D. (2010). IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11, 588.

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

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4), 1165–1188.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Res*, 34(Database issue), D16–D20.

Berman, H., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nature Struct Biol*, 10(12), 980.

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

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5), 34–43.

Bhattacharyya, R. P., Reményi, A., Yeh, B. J., & Lim, W. A. (2006). Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. *Annu Rev Biochem*, 75, 655–680.

Birkland, A. & Yona, G. (2006). BIOZON: a system for unification, management and analysis of heterogeneous biological data. *BMC Bioinformatics*, 7, 70.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Graf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios, D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., & Hubbard, T. J. (2006). Ensembl 2006. *Nucleic Acids Res*, 34(Database issue), D556–D561.

Biswas, M., O'Rourke, J. F., Camon, E., Fraser, G., Kanapin, A., Karavidopoulou, Y., Kersey, P., Kriventseva, E., Mittard, V., Mulder, N., Phan, I., Servant, F., & Apweiler, R. (2002). Applications of InterPro in protein annotation and genome analysis. *Brief Bioinform*, 3(3), 285–295.

Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Eppig, J. T., & the Mouse Genome Database Group (2010). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*, 39(Database issue): D842–D848

Blankenburg, H., Finn, R. D., Prlić, A., Jenkinson, A. M., Ramírez, F., Emig, D., Schelhorn, S.-E., Büch, J., Lengauer, T., & Albrecht, M. (2009a). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10), 1321–8.

Blankenburg, H., Ramírez, F., Büch, J., & Albrecht, M. (2009b). DASMIweb: online integration, analysis and assessment of distributed protein interaction data. *Nucleic Acids Res*, 37(Web Server issue), W122–W128.

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

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P. O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., Hopf, C., Huhse, B., Mangano, R., Michon, A. M., Schirle, M., Schlegl, J., Schwab, M., Stein, M. A., Bauer, A., Casari, G., Drewes, G., Gavin, A. C., Jackson, D. B., Joberty, G., Neubauer, G., Rick, J., Kuster, B., & Superti-Furga, G. (2004). A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. *Nature Cell Biol*, 6, 97–105.

Brazma, a., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. a., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., & Vingron, M. (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, 29(4), 365–71.

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

Buckley, C. & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM New York, NY, USA.

Buday, L. & Tompa, P. (2010). Functional classification of scaffold proteins and related molecules. *The FEBS journal*, 277, 4348–4355.

Buffa, L., Fuchs, E., Pietropaolo, M., Barr, F., & Solimena, M. (2008). ICA69 is a novel Rab2 effector regulating ER-Golgi trafficking in insulinoma cells. *Eur J Cell Biol*, 87(4), 197–209.

Camargo, L. M., Collura, V., Rain, J. C., Mizuguchi, K., Hermjakob, H., Kerrien, S., Bonnert, T. P., Whiting, P. J., & Brandon, N. J. (2007). Disrupted in Schizophrenia 1 Interactome: evidence for the close connectivity of risk genes and a potential synaptic basis for schizophrenia. *Mol Psychiatry*, 12(1), 74–86.

Caviston, J. P. & Holzbaur, E. L. F. (2009). Huntingtin as an essential integrator of intracellular vesicular trafficking. *Trends Cell Biol*, 19(4), 147–155.

Ceol, A., Aryamontri, A. C., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., & Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res*, 38(Database issue), D532–D539.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., & Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res*, 39(Database issue), D685–D690.

Chabalier, J., Mosser, J., & Burgun, A. (2007). A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8, 235.

Chan, J. N. Y., Nislow, C., & Emili, A. (2010). Recent advances and method development for drug target identification. *Trends Pharmacol Sci*, 31(2), 82–88.

Chen, F., Mackey, A. J., Stoeckert, C. J., & Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, 34(Database issue), D363–D368.

Choi, K. Y., Satterberg, B., Lyons, D. M., & Elion, E. A. (1994). Ste5 tethers multiple protein kinases in the MAP kinase cascade required for mating in S. cerevisiae. *Cell*, 78(3), 499–512.

Chung, H.-J., Park, C. H., Han, M. R., Lee, S., Ohn, J. H., Kim, J., Kim, J., & Kim, J. H. (2005). ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res*, 33(Web Server issue), W621–W626.

Cochrane, G., Karsch-Mizrachi, I., & Nakamura, Y. (2011). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res*, 39(Database issue), D15–D18.

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P., & Gauthier, J. M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome Res*, 14, 1324–1332.

Côté, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R., & Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC bioinformatics*, 8, 401.

Cusick, M. E., Klitgord, N., Vidal, M., & Hill, D. E. (2005). Interactome: gateway into systems biology. *Hum Mol Genet*, 14 Spec No, R171–R181.

Deane, C. M., Salwinski, L., Xenarios, I., & Eisenberg, D. (2002). Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol Cell Proteomics*, 1(5), 349–356.

del Pozo, A., Pazos, F., & Valencia, A. (2008). Defining functional distances over Gene Ontology. *BMC Bioinformatics*, 9, 50.

Demir, E., Cary, M. P., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., Schacherer, F., Martinez-Flores, I., Hu, Z., Jimenez-Jacinto, V., Joshi-Tope, G., Kandasamy, K., Lopez-Fuentes, A. C., Mi, H., Pichler, E., Rodchenkov, I., Splendiani, A., Tkachev, S., Zucker, J., Gopinath, G., Rajasimha, H., Ramakrishnan, R., Shah, I., Syed, M., Anwar, N., Babur, O., Blinov, M., Brauner, E., Corwin, D., Donaldson, S., Gibbons, F., Goldberg, R., Hornbeck, P., Luna, A., Murray-Rust, P., Neumann, E., Reubenacker, O., Samwald, M., van Iersel, M., Wimalaratne, S., Allen, K., Braun, B., Whirl-Carrillo, M., Cheung, K.-H., Dahlquist, K., Finney, A., Gillespie, M., Glass, E., Gong, L., Haw, R., Honig, M., Hubaut, O., Kane, D., Krupa, S., Kutmon, M., Leonard, J., Marks, D., Merberg, D., Petri, V., Pico, A., Ravenscroft, D., Ren, L., Shah, N., Sunshine, M., Tang, R., Whaley, R., Letovksy, S., Buetow, K. H., Rzhetsky, A., Schachter, V., Sobral, B. S., Dogrusoz, U., McWeeney, S., Aladjem, M., Birney, E., Collado-Vides, J., Goto, S., Hucka, M., Novère, N. L., Maltsev, N., Pandey, A., Thomas, P., Wingender, E., Karp, P. D., Sander, C., & Bader, G. D. (2010). The BioPAX community standard for pathway data sharing. *Nature Biotechnology*, 28(9), 935–942.

Dowell, R., Jokerst, R., Day, A., Eddy, S., & Stein, L. (2001). The distributed annotation system. *BMC bioinformatics*, 2(1), 7.

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., & Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet*, 18(10), 529–536.

Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R., Oughtred, R., Park, J., Skrzypek, M. S., Weng, S., Wong, E. D., Dolinski, K., Botstein, D., & Cherry, J. M. (2010). Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res*, 38(Database issue), D433–D436.

Etzold, T., Ulyanov, A., & Argos, P. (1996). SRS: information retrieval system for molecular biology data banks. *Methods in enzymology*, 266, 114–28.

Fingerman, I. M., McDaniel, L., Zhang, X., Ratzat, W., Hassan, T., Jiang, Z., Cohen, R. F., & Schuler, G. D. (2011). NCBI Epigenomics: a new public resource for exploring epigenomic data sets. *Nucleic Acids Res*, 39(Database issue), D908–D912.

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

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., & Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res*, 36(Database issue), D281–D288.

Flicek, P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Gräf, S., Haider, S., Hammond, M., Holland, R., Howe, K. L., Howe, K., Johnson, N., Jenkinson, A., Kähäri, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl,

P., Overduin, B., Parker, A., Pritchard, B., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A. J., Vogel, J., White, S., Wood, M., Birney, E., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Kasprzyk, A., Proctor, G., Smith, J., Ureta-Vidal, A., & Searle, S. (2008). Ensembl 2008. *Nucleic Acids Res*, 36(Database issue), D707–D714.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gordon, L., Hendrix, M., Hourlier, T., Johnson, N., Kähäri, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Larsson, P., Longden, I., McLaren, W., Overduin, B., Pritchard, B., Riat, H. S., Rios, D., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sobral, D., Spudich, G., Tang, Y. A., Trevanion, S., Vandrovcova, J., Vilella, A. J., White, S., Wilder, S. P., Zadissa, A., Zamora, J., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Vogel, J., & Searle, S. M. J. (2011). Ensembl 2011. *Nucleic Acids Res*, 39(Database issue), D800–D806.

Friedberg, I. (2006). Automated protein function prediction–the genomic challenge. *Briefings in bioinformatics*, 7(3), 225–42.

Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., & Kanehisa, M. (1998). DBGET/LinkDB: an integrated database retrieval system. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, (pp. 683–94).

Fujita, P. a., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. a., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., & Kent, W. J. (2011). The UCSC Genome Browser database: update 2011. *Nucleic Acids Res*, 39(Database issue), D876–D882.

Galperin, M. Y. & Cochrane, G. R. (2011). The 2011 Nucleic Acids Res Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Res*, 39(Database), D1–D6.

Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., & Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, 38(3), 285–293.

Gavin, A. C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M. A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., & Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415, 141–147.

Ge, H., Liu, Z., Church, G. M., & Vidal, M. (2001). Correlation between transcriptome and interactome mapping data from {S}accharomyces cerevisiae. *Nature Genetics*, 29, 482–486.

Ge, H., Walhout, A. J. M., & Vidal, M. (2003). Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*, 19(10), 551–560.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, W., KP, B., M, J., T, G., S, L., M, K., J, S., RA, M., MP, C., J, R., & JM (2003). A protein interaction map of Drosophila melanogaster. *Science*, 302, 1727–1736.

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

Goecks, J., Nekrutenko, A., & Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8), R86.

Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K. S., Knoblich, M., Haenig, C., Herbst, M., Suopanki, J., Scherzinger, E., Abraham, C., Bauer, B., Hasenbank, R., Fritzsche, A., Ludewig, A. H., Buessow, K., Coleman, S. H., Gutekunst, C. A., Landwehrmeyer, B. G., Lehrach, H., & Wanker, E. E. (2004). A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Mol Cell*, 15, 853–865.

Goffard, N., Frickey, T., & Weiller, G. (2009). PathExpress update: the enzyme neighbourhood method of associating gene-expression data with metabolic pathways. *Nucleic Acids Res*, 37(Web Server issue), W335–W339.

Goll, J. & Uetz, P. (2006). The elusive yeast interactome. *Genome Biol*, 7(6), 223.

Gribskov, M. & Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem*, 20(1), 25–33.

Guo, D., Han, J., Adam, B. L., Colburn, N. H., Wang, M. H., Dong, Z., Eizirik, D. L., She, J. X., & Wang, C. Y. (2005). Proteomic analysis of SUMO4 substrates in HEK293 cells under serum starvation-induced stress. *Biochem Biophys Res Commun*, 337(4), 1308–1318.

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

Han, J., Kamber, M., & Pei, J. (2005). *Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., & Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue), D258–D261.

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, a. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761 Suppl), C47–52.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sø rensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W., Figeys, D., & Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415, 180–183.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, (December 1977), 65–70.

Homma, K., Suzuki, K., & Sugawara, H. (2011). The Autophagy Database: an all-inclusive information resource on autophagy that provides nourishment for research. *Nucleic Acids Res*, 39(Database

issue), D986–D990.

Huang, D. W., Sherman, B. T., & Lempicki, R. a. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1), 1–13.

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

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

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue), W729–W732.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., & Yeats, C. (2008). InterPro: the integrative protein signature database. *Nucleic Acids Res*, 37(Database issue): D211–D215.

Huttenhower, C., Haley, E. M., Hibbs, M. A., Dumeaux, V., Barrett, D. R., Coller, H. A., & Troyanskaya, O. G. (2009). Exploring the human genome with functional maps. *Genome research*, 19(6), 1093–106.

Ideker, T. & Sharan, R. (2008). Protein networks in disease. *Genome research*, 18(4), 644–52.

Imarisio, S., Carmichael, J., Korolchuk, V., Chen, C.-W., Saiki, S., Rose, C., Krishna, G., Davies, J. E., Ttofi, E., Underwood, B. R., & Rubinsztein, D. C. (2008). Huntington's disease: from pathology and genetics to potential therapies. *Biochem J*, 412(2), 191–209.

Isserlin, R., El-Badrawi, R. A., & Bader, G. D. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. *Database : the journal of biological databases and curation*, 2011, baq037.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., & Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98, 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., & Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302, 449–453.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., & von Mering, C. (2009). STRING 8 – a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, 37(Database issue), D412–D416.

Jin, J., Smith, F. D., Stark, C., Wells, C. D., Fawcett, J. P., Kulkarni, S., Metalnikov, P., O'Donnell, P., Taylor, P., Taylor, L., Zougman, A., Woodgett, J. R., Langeberg, L. K., Scott, J. D., & Pawson, T. (2004). Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Curr Biol*, 14(16), 1436–1450.

Kacprowski, T. M. P. I. f. I. (2010). *Analysis of protein complexes and tissue-spedific expression*. PhD thesis, Universität des Saarlandes.

Kamburov, A., Wierling, C., Lehrach, H., & Herwig, R. (2009). ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res*, 37(Database issue), D623–D628.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., & Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue), D480–D484.

Kaya, K. D., Karakülah, G., Yakicier, C. M., Acar, A. C., & Konu, O. (2011). mESAdb: microRNA expression and sequence analysis database. *Nucleic Acids Res*, 39(Database issue), D170–D180.

Kerr, K. F. (2009). Comments on the analysis of unbalanced microarray data. *Bioinformatics*, 25(16), 2035–41.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., & Hermjakob, H. (2007). IntAct – open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database issue), D561–D565.

Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y., Birney, E., & Apweiler, R. (2004). The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, 4(7), 1985–1988.

Khatri, P. & Drăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18), 3587–95.

Kitano, H. (2002). Systems biology: a brief overview. *Science*, 295(5560), 1662–1664.

Klein, T. E., Chang, J. T., Cho, M. K., Easton, K. L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D. E., Rubin, D. L., Shafa, F., Stuart, J. M., & Altman, R. B. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. *The pharmacogenomics journal*, 1(3), 167–70.

Kolesnick, R. & Xing, H. R. (2004). Inflammatory bowel disease reveals the kinase activity of KSR1. *J Clin Invest*, 114(9), 1233–1237.

Küntzer, J., Blum, T., Gerasch, A., Backes, C., Hildebrandt, A., Kaufmann, M., Kohlbacher, O., & Lenhof, H. (2006). BN++-a biological information system. *J Integr Bioinformatics*, 3(2), 34.

Laity, J. H., Lee, B. M., & Wright, P. E. (2001). Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*, 11(1), 39–46.

Lee, T. J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D. W. J., Tenenbaum, J. D., & Karp, P. D. (2006). BioWarehouse: a bioinformatics database warehouse toolkit. *BMC bioinformatics*, 7, 170.

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

Lehner, B., Semple, J. I., Brown, S. E., Counsell, D., Campbell, R. D., & Sanderson, C. M. (2004). Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics*, 83(1), 153–167.

Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., & Apweiler, R. (2004). UniProt archive. *Bioinformatics*, 20(17), 3236–7.

Lerman, G. & Shakhnovich, B. E. (2007). Defining functional distance using manifold embeddings of Gene Ontology annotations. *Proc Natl Acad Sci U S A*, 104(27), 11334–11339.

Levine, M. & Hoey, T. (1988). Homeobox proteins as sequence-specific transcription factors. *Cell*, 55(4), 537–40.

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., Li, J., Han, J., Zhang, F., Gong, B., & Xu, L. (2009). SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res*, 37(19), e131.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., & Vidal, M. (2004). A map of the interactome network of the metazoan C. elegans. *Science*, 303, 540–543.

Liebel, U., Kindler, B., & Pepperkok, R. (2004). 'Harvester': a fast meta search engine of human protein resources. *Bioinformatics*, 20(12), 1962–3.

Lim, J., Hao, T., Shaw, C., Patel, A. J., Szabo, G., Rual, J. F., Fisk, C. J., Li, N., Smolyar, A., Hill, D. E., Barabasi, A. L., Vidal, M., & Zoghbi, H. Y. (2006). A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. *Cell*, 125(4), 801–814.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *Proc. of the 15th International Conference on Machine Learning*.

Locasale, J. W., Shaw, A. S., & Chakraborty, A. K. (2007). Scaffold proteins confer diverse regulatory properties to protein kinase cascades. *Proc Natl Acad Sci U S A*, 104(33), 13307–13312.

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

Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16), 3448–9.

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

Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(Database issue), D26–D31.

Marcora, E., Gowan, K., & Lee, J. E. (2003). Stimulation of NeuroD activity by huntingtin and huntingtin-associated proteins HAP1 and MLK2. *Proc Natl Acad Sci U S A*, 100(16), 9578–9583.

Massari, M. E. & Murre, C. (2000). Helix-Loop-Helix Proteins: Regulators of Transcription in Eucaryotic Organisms. *Molecular and Cellular Biology*, 20(2), 429–440.

Masseroli, M., Martucci, D., & Pinciroli, F. (2004). GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res*, 32(Web Server issue), W293–W300.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., & D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res*, 37(Database issue), D619–D622

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

McDowall, M. D., Scott, M. S., & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res*, 37(Database issue), D651–D656.

McMahon, H. T., Missler, M., Li, C., & Südhof, T. C. (1995). Complexins: cytosolic proteins that regulate SNAP receptor function. *Cell*, 83(1), 111–119.

McWilliam, H., Valentin, F., Goujon, M., Li, W., Narayanasamy, M., Martin, J., Miyar, T., & Lopez, R. (2009). Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res*, 37(Web Server issue), W6–W10.

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Rev Genet*, 11(1), 31–46.

Mika, S. & Rost, B. (2006). Protein-protein interactions more conserved within species than across species. *PLoS Comput Biol*, 2(7), e79.

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

Mrowka, R., Patzak, A., & Herzel, H. (2001). Is there a bias in proteome research? *Genome Res*, 11, 1971–1973.

Nakayama, M., Kikuno, R., & Ohara, O. (2002). Protein-protein interactions between large proteins: two-hybrid screening using a functionally classified library composed of long cDNAs. *Genome Res*, 12(11), 1773–1784.

Neumann, E. (2005). A life science Semantic Web: are we there yet? *Sci STKE*, 2005(283), pe22.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, 2(3), 117–120.

Obayashi, T. & Kinoshita, K. (2011). COXPRESdb: a database to compare gene coexpression in seven model animals. *Nucleic Acids Res*, 39(Database issue), D1016–D1022.

O'Brien, K. P., Remm, M., & Sonnhammer, E. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33, D476–D480.

Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., & Brazma, A. (2010). ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res*, 38(Database issue), D690–D698.

Pawson, T. & Nash, P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science*, 300(5618), 445–452.

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

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

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

Popescu, M., Keller, J. M., & Mitchell, J. A. (2006). Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 3(3), 263–274.

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W., & Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 38(Database

issue), D105–D110.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Kishore, C. J. H., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., & Pandey, A. (2009). Human Protein Reference Database – 2009 update. *Nucleic Acids Res*, 37(Database issue), D767–D772.

Prud'hommeaux, E. & Seaborne, A. (2008). SPARQL Query Language for RDF.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue), D501–D504.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing.

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A. Y., Walter, J. C., & LaBaer, J. (2004). Self-assembling protein microarrays. *Science*, 305(5680), 86–90.

Ramani, A. K., Bunescu, R. C., Mooney, R. J., & Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*, 6, R40.

Ramírez, F. & Albrecht, M. (2010). Finding scaffold proteins in interactomes. *Trends in cell biology*, 20(1), 2–4.

Ramirez, F., Lawyer, G., & Albrecht, M. (2011). Novel search method for the discovery of functional relationships. *Bioinformatics*, (in press).

Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., & Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics*, 7(15), 2541–52.

Ramsköld, D., Wang, E. T., Burge, C. B., & Sandberg, R. (2009). An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS computational biology*, 5(12), e1000598.

Razick, S., Magklaras, G., & Donaldson, I. M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, 9(1), 405.

Reeves, G. A., Eilbeck, K., Magrane, M., O'Donovan, C., Montecchi-Palazzi, L., Harris, M. A., Orchard, S., Jimenez, R. C., Prlic, A., Hubbard, T. J. P., Hermjakob, H., & Thornton, J. M. (2008). The Protein Feature Ontology: a tool for the unification of protein feature annotations. *Bioinformatics*, 24(23), 2767–2772.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B. J., Hon, G. C., Myers, C. L., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O. G., Ideker, T., Dolinski, K., Batada, N. N., & Tyers, M. (2006). Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J Biol*, 5(4), 11.

Reiss, S., Rebhan, I., Backes, P., Romero-Brey, I., Erfle, H., Matula, P., Kaderali, L., Poenisch, M., Blankenburg, H., Hiet, M.-S., Longerich, T., Diehl, S., Ramírez, F., Balla, T., Rohr, K., Kaul, A., Bühler, S., Pepperkok, R., Lengauer, T., Albrecht, M., Eils, R., Schirmacher, P., Lohmann, V., & Bartenschlager, R. (2011). Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell Host Microbe*, 9(1), 32–45.

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

Resnik, P. (1999). Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11, 95–130.

Rhee, S. Y., Wood, V., Dolinski, K., & Draghici, S. (2008). Use and misuse of the gene ontology annotations. *Nature Rev Genet*, 9(7), 509–15.

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

Rivals, I., Personnaz, L., Taing, L., & Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, 23(4), 401–407.

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., & Karp, P. D. (2005). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol*, 6(1), R2.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., & Bourne, P. E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*, 39(Database issue), D392–D401.

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., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., & Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 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., & Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 36(Database issue), D646–D650.

Ruttenberg, A., Rees, J. A., Samwald, M., & Marshall, M. S. (2009). Life sciences on the Semantic Web: the Neurocommons and beyond. *Briefings in bioinformatics*, 10(2), 193–204.

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., & Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database : the journal of biological databases and curation*, 2010, baq020.

Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., Ben-Dor, U., Esterman, N., Rosen, N., Peter, I., & Others (2002). GeneCards^{TM} 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, 18(11), 1542–3.

Salomonis, N., Hanspers, K., Zambon, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., Doniger, S. W., Stuart, J., Conklin, B. R., & Pico, A. R. (2007). GenMAPP 2: new features and resources for pathway analysis. *BMC bioinformatics*, 8(1), 217.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11), 613–620.

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

Sauer, U., Heinemann, M., & Zamboni, N. (2007). Genetics. Getting closer to the whole picture. *Science (New York, N.Y.)*, 316(5824), 550–1.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., Feolo, M., Fingerman, I. M., Geer, L. Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D. J., Lu, Z., Madden, T. L., Madej, T., Maglott, D. R., Marchler-Bauer, A., Miller, V., Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T. A., Wagner, L., Wang, Y., Wilbur, W. J., Yaschenko, E., & Ye, J. (2011). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 39(Database issue), D38–D51.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res*, 37(Database issue), D674–D679.

Schelhorn, S.-E., Lengauer, T., & Albrecht, M. (2008). An integrative approach for predicting interactions of protein regions. *Bioinformatics*, 24(16), i35–i41.

Schlicker, A., Domingues, F. S., Rahnenführer, J., & Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC bioinformatics*, 7, 302.

Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., & Albrecht, M. (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7), 859–65.

Schlicker, A., Lengauer, T., & Albrecht, M. (2010). Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18), i561–i567.

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5(12), e1000605.

Seal, R. L., Gordon, S. M., Lush, M. J., Wright, M. W., & Bruford, E. A. (2011). genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*, 39(Database issue), D514–D519.

Sevilla, J. L., Segura, V., Podhorski, A., Guruceaga, E., Mato, J. M., Martínez-Cruz, L. A., Corrales, F. J., & Rubio, A. (2005). Correlation between gene expression and GO semantic similarity. *IEEE/ACM Trans Comput Biol Bioinform*, 2(4), 330–338.

Shah, S. P., Huang, Y., Xu, T., Yuen, M. M. S., Ling, J., & Ouellette, B. F. F. (2005). Atlas - a data warehouse for integrative bioinformatics. *BMC bioinformatics*, 6, 34.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498–2504.

Shaw, A. S. & Filbert, E. L. (2009). Scaffold proteins and immune-cell signalling. *Nature Rev Immunol*, 9(1), 47–56.

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*, 11(1), 92–99.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20), 3940–3941.

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

Speer, N., Spieth, C., & Zell, A. (2004). A memetic clustering algorithm for the functional partition of genes based on the Gene Ontology. In *Computational Intelligence in Bioinformatics and Computational Biology, 2004. CIBCB'04. Proceedings of the 2004 IEEE Symposium on* (pp. 252–259).

Spiegelman, B. M. & Flier, J. S. (2001). Obesity and the regulation of energy balance. *Cell*, 104(4), 531–543.

Sprinzak, E., Sattath, S., & Margalit, H. (2003). How reliable are experimental protein-protein interaction data? *J Mol Biol*, 327, 919–923.

Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., & Tyers, M. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res*, 39(Database issue), D698–D704.

Stein, L. D. (2003). Integrating biological databases. *Nature Rev Genet*, 4(5), 337–45.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., & Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122, 957–968.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3), 479–498.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., & Hogenesch, J. B. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, 99(7), 4465–4470.

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., & Meng, F. (2010). GLay: community structure analysis of biological networks. *Bioinformatics*, 26(24), 3135–7.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10), 1282–1288.

Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., Oda, H., Endo, T., Kondo, S., Bono, H., Okazaki, Y., & Hayashizaki, Y. (2001). Protein-protein interaction panel using mouse full-length cDNAs. *Genome Res*, 11, 1758–1765.

The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, 38(Database issue), D142–D148.

The UniProt Consortium (2011). Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, 39(Database issue), D214–D219.

Tomlins, S. A., Mehra, R., Rhodes, D. R., Cao, X., Wang, L., Dhanasekaran, S. M., Kalyana-Sundaram, S., Wei, J. T., Rubin, M. A., Pienta, K. J., Shah, R. B., & Chinnaiyan, A. M. (2007). Integrative molecular concept modeling of prostate cancer progression. *Nature Genetics*, 39(1), 41–51.

Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., Andrews, B., Tyers, M., & Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, 294, 2364–2368.

Tress, M. L., Martelli, P. L., Frankish, A., Reeves, G. a., Wesselink, J. J., Yeats, C., Olason, P. I., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J., Laskowski, R. a., López, G., Sadowski, M. I., Watson, J. D., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Stø rling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S. E., Reymond, A., Birney, E., Brunak, S. r., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D. T., Lengauer, T., Orengo, C. a., Patthy, L., Thornton, J. M., Tramontano, A., & Valencia, A. (2007). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13), 5495–500.

Tsang, H. T., Connell, J. W., Brown, S. E., Thompson, A., Reid, E., & Sanderson, C. M. (2006). A systematic analysis of human CHMP protein interactions: additional MIT domain-containing proteins bind to multiple components of the human ESCRT III complex. *Genomics*, 88(3), 333–346.

Turinsky, A. L., Razick, S., Turner, B., Donaldson, I. M., & Wodak, S. J. (2011). Interaction databases on the same page. *Nature Biotechnology*, 29(5), 391–3.

Uetz, P., Dong, Y.-A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., & Haas, J. (2006). Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758), 239–242.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403, 623–627.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Rev Genet*, 10(4), 252–263.

Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., & Liu, E. A. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science*, 303(5659), 844–848.

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

Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, 144(6), 986–98.

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

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., & Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417, 399–403.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18, 1283–1292.

Walhout, A. J., Sordella, R., Lu, X., Hartley, J. L., Temple, G. F., Brasch, M. A., Thierry-Mieg, N., & Vidal, M. (2000). Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science*, 287, 116–122.

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

Wang, P. I. & Marcotte, E. M. (2010). It's the machine that matters: Predicting gene function and phenotype from protein networks. *Journal of Proteomics*, 73(11), 2277–2289.

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res*, 38(Web Server), W214–W220.

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences*, 38(6), 983–996.

Wolfe, S. A., Nekludova, L., & Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure*, 29, 183–212.

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Mazumder, R., O'Donovan, C., Redaschi, N., & Suzek, B. (2006a). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*, 34(Database issue), D187–D191.

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

Yamashita, R., Wakaguri, H., Sugano, S., Suzuki, Y., & Nakai, K. (2010). DBTSS provides a tissue specific dynamic view of Transcription Start Sites. *Nucleic Acids Res*, 38(Database issue), D98–D104.

Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., & Qu, L.-H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res*, 39(Database issue), D202–D209.

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

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., & Cesareni, G. (2002). MINT: a Molecular INTeraction database. *FEBS Lett*, 513(1), 135–140.

Zeeberg, B., Feng, W., Wang, G., Wang, M., Fojo, A., Sunshine, M., Narasimhan, S., Kane, D., Reinhold, W., Lababidi, S., Bussey, K., Riss, J., Barrett, J., & Weinstein, J. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, 4(4), R28.

Zeke, A., Lukács, M., Lim, W. a., & Reményi, A. (2009). Scaffolds: interaction platforms for cellular signalling circuits. *Trends in cell biology*, 19(8), 364–74.

Zhang, B., Kirov, S., & Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, 33(Web Server issue), W741–W748.

Appendix

A

Manually validated scaffold proteins

symbol	type	pubmed id	text snippet
AKAP12	scaffold	17686059	A-kinase anchor protein 12 (**AKAP12**) is a **scaffold** protein that participates in mitotic regulation and others signalling processes and probably exerts tumour suppressor function.
	scaffold	16442664	A-kinase anchoring proteins (**AKAPs**) define an expanding group of **scaffold** proteins that display a signature binding site for the RI/RII subunit of protein kinase A.
	scaffold	17626016	SSeCKS (Src-suppressed C kinase substrate), also called gravin/**AKAP12**, is a large **scaffolding** protein with metastasis suppressor activity.
		and others	
AKAP9 (AKAP350)	scaffold	14569596	Recently, we demonstrated that several CLIC proteins, including CLIC4, interact with **AKAP350**. **AKAP350** is concentrated at the Golgi apparatus, centrosome, and midbody and acts as a **scaffolding** protein for several protein kinases and phosphatases.
	scaffold	12163481	The protein kinase A-anchoring proteins (**AKAP**s) are defined by their ability to **scaffold** protein kinase A to specific subcellular compartments.
	scaffold	12163479	**AKAP350** can **scaffold** a number of protein kinases and phosphatases at the centrosome and the Golgi apparatus.
		and others	
APC	scaffold	19434056	The PDZ-proteins also promote functional nicotinic innervation of the neurons, as does the **scaffold** protein **APC** and transmembrane proteins such as neuroligin and the EphB2 receptor.
BCAR1 (p130Cas)	scaffold	18256281	For over a decade, p130CAS/**BCAR1**, HEF1/NEDD9/CAS-L, and Efs/Sin have defined the CAS (Crk-associated substrate) **scaffolding** protein family.
	adaptor	19357231	Phosphorylation of the **adaptor** protein **p130Cas** was inversely related to phagocytosis
	adaptor	19064994	Exposure to L. major resulted in degradation of the phosphorylated **adaptor** protein **p130Cas** and the protein-tyrosine phosphatase-PEST
		and others	
BIRC2	adaptor	18997792	The **adaptor** and signaling proteins TRAF2, TRAF3, **cIAP1** and cIAP2 may inhibit alternative nuclear factor-kappaB (NF-kappaB) signaling in resting cells by targeting NF-kappaB-inducing kinase (NIK) for ubiquitin-dependent degradation

Continued on next page

symbol	type	pubmed id	text snippet
CD81	scaffold	15004227	These data reveal a potential role for tetraspanins CD9 and **CD81** as GPCR **scaffolding** proteins
CRK (CRKII)	adaptor	18599349	One family of **adaptor** molecules includes the **CRKII/CRKL** proteins that are also involved in the regulation of lymphocyte function.
	adaptor	19350053	**CRK adaptor** protein-induced phosphorylation of Gab1 on tyrosine 307 via Src is important for organization of focal adhesions and enhanced cell migration.
	adaptor	18981215	Experiments with cultured neurons have shown that when Dab1 is phosphorylated on tyrosine, it activates Akt and provides a scaffold for assembling signaling complexes, including the paralogous **CRK** and **CRKL adaptors**
		and others	
DAB1	adaptor/ scaffold	18981215	These results show that Reelin-induced Akt stimulation and **DAB1** turnover are not sufficient for normal development and suggest that **DAB1** acts both as a kinase switch and as a **scaffold** for assembling signaling complexes in vivo. They also call DAB1 'adapto
	adaptor	10827173	Signaling through these receptors requires the interaction of their cytoplasmic tails with the intracellular **adaptor** protein Disabled-1 (**DAB1**).
DBNL (ABP1, HIP-55, SH3P7)	adaptor	15798181	**ABP1** acts as an **adaptor** protein in the localization or concentration of Sjl2 during late stages of endocytic vesicle formation.
	adaptor	14729663	We found that the cytoplasmic **adaptor HIP-55**, a Src/Syk-kinases substrate and member of the drebrin/**ABP1** family of actin-binding proteins
	adaptor	14718626	The CD2v protein of African swine fever virus interacts with the actin-binding **adaptor** protein **SH3P7**.
	adaptor	19725075	Further analysis of **HIP-55** revealed that this **adaptor** protein becomes increasingly associated with both Syk and integrin beta3 upon platelet activation.
		and others	
DOK1/2 (P62DOK)	adaptor	16338067	**DOK1** is an **adaptor** tyrosine kinase substrate with tumor-suppressive activity.
	adaptor	16823827	The data in this report show that both the **DOK1** and the **DOK2 adaptor** proteins are constitutively expressed in the myelomonoblastic leukemia cell line
	adaptor	11254695	**p62(dok)** belongs to a newly identified family of **adaptor** proteins.
		and others	

Continued on next page

symbol	type	pubmed id	text snippet
EFS (SYN)	scaffold	11867627	Instead, the pathway involved relies on increased tyrosine phosphorylation of, and recruitment of Crk to, the SRC substrate **SIN**/**EFS**. The latter is a **scaffolding** protein structurally similar to the SRC substrate Cas
	scaffold	18256281	For over a decade, p130Cas/BCAR1, HEF1/NEDD9/Cas-L, and **EFS**/**SIN** have defined the Cas (Crk-associated substrate) **scaffolding** protein family.
FADD	adaptor	19583773	Fas-associated protein with death domain (**FADD**) is an essential **adaptor** protein in death receptor-mediated signal transduction
	adaptor	18661484	To investigate the role of the apoptosis **adaptor** molecules TRADD and **FADD** in the development of hematological diseases, patient samples were screened for mutations in these genes.
	adaptor	9582077	When activated, membrane-bound receptors for Fas and tumour-necrosis factor initiate programmed cell death by recruiting the death domain of the **adaptor** protein **FADD** to the membrane.
		and others	
FRS2/3	adaptor	16887332	**FRS2/3** are homologs that function as **adaptor** proteins to mediate signaling of multiple receptor tyrosine kinases.
	adaptor	16702953	Unique role of **FRS3** docking/**adaptor** protein for negative regulation in EGF receptor tyrosine kinase signaling pathways.
GIPC1 (SYNECTIN)	adaptor	15459234	CD93 interacts with the PDZ domain-containing **adaptor** protein **GIPC**: implications in the modulation of phagocytosis
	adaptor	16467373	The PDZ **adaptor** protein **SYNECTIN** bound the longer splice variant, Syx1, which was targeted to the plasma membrane in a **SYNECTIN**-dependent manner
	scaffold	16940428	**SYNECTIN** (**GIPC1**), a receptor **scaffold** protein, has been isolated by our laboratory as a syndecan-4 cytoplasmic domain binding partner that regulates important aspects of cell motility
	scaffold	15356268	We show that recruitment of GAIP (RGS19) by the dopamine D2 receptor (D2R), a GPCR, required the **scaffold** protein **GIPC** (GAIP-interacting protein, C terminus) and that all three were coexpressed in neurons and neuroendocrine cells
		and others	
HTT	scaffold	12881483	We propose that **HTT**, together with HAP1, may function as a **scaffold** for the activation of ND by MLK2.
	scaffold	19269181	Overall, the predicted structure of **huntingtin** is consistent with a cellular role as a **scaffold** protein.
	scaffold	19429504	Unexpectedly, the faulty gene product, mutant **huntingtin** (mtHtt), is an extremely large protein of 350 kDa and might act as a **scaffold** protein regulating vesicle and organelle trafficking and signaling pathways.

Continued on next page

symbol	type	pubmed id	text snippet
INADL (PATJ)	scaffold	17234746	Here we report, using a two-hybrid assay, a direct molecular interaction between **TSC2** C-terminal part and PDZ 2 and 3 of **PATJ**, a **scaffold** member of the Crumbs 3 (CRB 3) complex in human intestinal epithelial cells, Caco2.
	scaffold	16697075	One evolutionarily conserved protein complex, which can be found both in Drosophila and mammalian epithelial cells, is composed of the transmembrane protein Crumbs/Crb3 and the **scaffolding** proteins Stardust/Pals1 and D**PATJ**/**PATJ**, respectively, and localise
	scaffold	15863617	A unified assembly mode revealed by the structures of tetrameric L27 domain complexes formed by mLin-2/mLin-7 and **PATJ**/Pals1 **scaffold** proteins. AXIN1 almost all references mention it as a Scaffold
IRS1/2	scaffold	19564410	Here we demonstrate the physical association of these signaling pathways using a proteomic approach that identified insulin-regulated complexes of JIPs together with **IRS scaffold** proteins
	adaptor	11024460	Interleukin-9 (IL-9) stimulation results in JAK, STAT and **IRS1**/**2** phosphorylation. The role of IRS **adaptor** proteins in IL-9 signaling is not clear
	adaptor	19671761	In addition, we find that levels of other components of the signaling pathway such as the **adaptor** proteins **IRS1** and **IRS2**
	adaptor	11024460 and others	The role of **IRS adaptor** proteins in IL-9 signaling is not clear. We show that IL-9 induces **IRS2** ...
MAGI1	scaffold	18971469	**MAGI-1**, a candidate stereociliary **scaffolding** protein, associates with the tip-link component cadherin 23.
	scaffold	19017743	TRIP6, a novel molecular partner of the **MAGI-1 scaffolding** molecule, promotes invasiveness.
	scaffold	19403801 and others	**MAGI-1** is a **scaffolding** protein that allows formation of complexes between certain transmembrane proteins, actin-binding proteins, and others regulatory proteins.
MLLT4 (AF-6, AFADIN)	scaffold	16819513	The AF-6/**MLLT4** gene, telomeric of PARK2, encodes the **AFADIN scaffold** protein, which is essential for epithelial integrity.
	scaffold	17473018	The human **AF-6**, a **scaffold** protein between cell membrane-associated proteins and the actin cytoskeleton, plays an important role in special cell-cell junctions and signal transduction.
	adaptor	18593353 and others	**AFADIN** additionally serves as an **adaptor** protein by further binding many **scaffolding** proteins
MAPKSP1 (MP1)	adaptor/ scaffold	15263099	Taken together, the presented work provides insight into the spatial regulation of MAPK signaling, illustrating how p14 and **MP1** collaborate as an endosomal **adaptor**/**scaffold** complex

Continued on next page

symbol	type	pubmed id	text snippet
	scaffold	19289794	On these LEs, we also identified the p14-**MP1 scaffolding** complex and activated extracellular signal-regulated kinase 1/2.
	scaffold	19177150 and others	p18 specifically binds to the p14-**MP1** complex, a **scaffold** for MEK1.
NF2 (Merlin)	adaptor	16341207	**Merlin** is an **adaptor** protein with a FERM domain and it is thought to transduce a growth-regulatory signal.
NPHS1 (Nephrin)	scaffold	18480178	**Nephrin**, an essential adhesion and **scaffolding** molecule expressed in podocytes, emerged in this screen
	scaffold	19443634	Within the glomerulus, the **scaffolding** protein **nephrin** bridges the actin-rich foot processes that extend from adjacent podocytes to form the slit diaphragm
PIK3R1 (p85)	scaffold	17024187	We show that **p85** acts as a **scaffold** to bind Cdc42 and septin 2 simultaneously. **p85** is thus involved in the spatial control of cytosolic division through regulation of Cdc42 and septin 2, in a PI3K-activity independent manner.
RASSF1 (RASSF1A)	scaffold	17878233	Current evidence supports the hypothesis that **RASSF1** serves as a **scaffold.**
	scaffold	18641684	The failure of recombinant **RASSF1A** to activate recombinant Aurora-A indicates that **RASSF1A** may not activate Aurora-A directly and suggests that **RASSF1A** may function as a **scaffold** to bring together Aurora-A and its activator(s).
RPTOR (Raptor)	scaffold	16824195	4E-BP1 has been shown to associate with the **scaffold** protein **raptor** through its TOS and RAIP motifs to be recognized by mTOR.
	adaptor	19439614	We demonstrate that mTOR exerts its effects on oligodendrocyte differentiation through two distinct signaling complexes, mTORC1 and mTORC2, defined by the presence of the **adaptor** proteins **raptor** and rictor, respectively
	scaffold	18722121	The **scaffolding** protein **Raptor** binds to mTOR and recruits substrates to the rapamycin-sensitive mTOR complex 1 (mTORC1).
	scaffold	16354680 and others	**Raptor** directly binds to and serves as a **scaffold** for mTOR-mediated phosphorylation of IRS-1 on Ser636/639
SHC2 (SHCB, SLI)	adaptor	12006576	The signaling adapters **SHCB** and ShcC, but not ShcA, are thought to be the primary Shc **adaptor** proteins in neurons as both are highly expressed in both the developing and adult nervous system.

Continued on next page

symbol	type	pubmed id	text snippet
	adaptor	15893635	Shc family of **adaptor** (including **SHCB** and ShcC) molecules has been demonstrated to play an important role during the transition from proliferating neural stem cells to postmitotic neurons.
	adaptor	17409413	The Src homology and collagen (Src) family of **adaptor** proteins comprises six Shc-like proteins encoded by three loci in mammals (Shc, Rai, and **SLI**)
SLC9A3R1 (NHERF1, EBP50)	adaptor	19073137	The **adaptor** protein **NHERF1** has previously been implicated in MRP4 internalization in non-polarized cells.
	adaptor	19857202	The **adaptor** protein **EBP50** is important for localization of the protein kinase A-Ezrin complex in T cells and the immunomodulating effect of cAMP.
	scaffold	19591839 and others	Ezrin induces long-range interdomain allostery in the **scaffolding** protein **NHERF1**.
SOCS3	adaptor	14707129	Suppressor of cytokine signaling (**SOCS**) proteins are a family of Src homology 2-containing **adaptor** proteins.
	adaptor	15541651	Regulation of the immune system by **SOCS** family **adaptor** proteins.
	adaptor	18948053	The SOCS box can also add unique features to individual SOCS proteins: it can function as an **adaptor** domain as was demonstrated for **SOCS3**, or as a modulator of substrate binding in case of CIS.
SPTBN1 (ELF)	adaptor	16650383	We have shown that loss of **ELF**, a stem cell **adaptor** protein, disrupts TGF-beta signaling through Smad3 and Smad4 localization
	adaptor/ scaffold	12543979	Disruption of the **adaptor** protein **ELF**, a beta-spectrin, leads to disruption of transforming growth factor-beta (TGF-beta) signaling by Smad proteins in mice.
	adaptor	16359909 and others	TGF-beta signaling mediator Smads are tightly dependent on modulation by **adaptor** proteins, such as **ELF**, SARA, filamin, and crkl as well as ubiquitinators, such as PRAJA and SMURFs.
SQSTM1 (p62)	scaffold	17229006	Mutations in **SQSTM1**, which encodes an important **scaffold** protein in this pathway, have been found to be a common cause of classical Paget's disease of bone (PDB)
	scaffold	19850933	Using a combined Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)-proteomics methodology, we identified the cytoplasmic **scaffold** protein **p62** as the molecular target of IDR-1
	scaffold	18174161	Sequestosome 1 (**SQSTM1**)/**p62** is an interacting partner of the atypical protein kinase C zeta/iota and serves as a **scaffold** for cell signaling and ubiquitin binding,

Continued on next page

symbol	type	pubmed id	text snippet
	adaptor	19765191	In cells lacking **p62**, the existence of mutant SOD1 in acidic autolysosomes decreased, suggesting that **p62** can function as an **adaptor** between mutant SOD1 and the autophagy machinery.
		and others	
TANK	scaffold	18353649	Recent data provide insight into the requirement for **scaffold** proteins in complex assembly; NF-kappaB essential modulator coordinates some IKK complexes, whereas **TANK**, NF-kappaB-activating kinase-associated protein 1 (NAP1) or similar to NAP1 TBK1 adaptor
	scaffold	17823124	we have identified **TANK** as a **scaffold** protein that assembles some but not all IRF3/7-phosphorylating TBK1-IKKepsilon complexes
TICAM1 (TRIF)	adaptor	19648648	LPS and lauric acid enhanced the association of TLR4 with MD-2 and downstream **adaptor** molecules, **TRIF** and MyD88
	adaptor	15032644	TLR intracellular domains could then specifically recruit several **adaptor** proteins including MyD88, TIRAP/MAL, **TRIF**, and TOLLIP.
	adaptor	19825364	TIR domain-containing **adaptor** protein (**TRIF**) is an **adaptor** protein in Toll-like
		and-more	
TOLLIP	adaptor	17113392	Our findings suggest that **TOLLIP** functions as an endosomal **adaptor** linking IL-1RI, via Tom1, to the endosomal degradation machinery.
	adaptor	15032644	TLR intracellular domains could then specifically recruit several **adaptor** proteins including MyD88, TIRAP/MAL, TRIF, and **TOLLIP**.
	adaptor	11751856	The **adaptor** protein **TOLLIP** was identified initially as an intermediate in interleukin (IL)-1 signaling
TRAF2/3/5	scaffold	16299380	Central role of the **scaffold** protein tumor necrosis factor receptor-associated factor 2 (**TRAF2**) in regulating endoplasmic reticulum stress-induced apoptosis.
	adaptor	18997792	The **adaptor** and signaling proteins **TRAF2**, **TRAF3**, cIAP1 and cIAP2 may inhibit alternative nuclear factor-kappaB (NF-kappaB) signaling in resting cells by targeting NF-kappaB-inducing kinase (NIK) for ubiquitin-dependent degradation ...
	adaptor	19198591	Depletion of membrane cholesterol inhibited the assembly of an IL-12-inducing CD40 signalosome containing the **adaptors TRAF2**, **TRAF3** and **TRAF5**
	adaptor	17991829	DUBA bound tumor necrosis factor receptor-associated factor 3 (**TRAF3**), an **adaptor** protein essential for the IFN-I response

Continued on next page

symbol	type	pubmed id	text snippet
TRIP10 (CIP4)	scaffold	12456510	Altogether, these data suggest that **CIP4**/Felic constitute a novel family of cytoskeletal **scaffolding** proteins, integrating Src and Cdc42 pathways.
	adaptor	17785506	Thus, **CIP4** is an important cytoskeletal **adaptor** that functions after filamentous actin accumulation and Cdc42
	adaptor	19632321	Cdc42-Interacting Protein-4 (**CIP4**) family **adaptors**
TXN (Thioredoxin, TRX)	scaffold	19177362	OptGraft: A computational procedure for transferring a binding site onto an existing protein **scaffold TXN**.
	scaffold	17875722	Expression of a **scaffold** protein (**Thioredoxin**) ....
	scaffold	16827663	Upon immunization, the V3 peptide-inserted **TRX scaffold** was able to generate anti-V3 antibodies ...
		and others	
WASF2 (WAVE2)	adaptor	15899863	We find that Tiam1 contributes to both of these processes by binding to IRSp53, an **adaptor** protein that is an effector for both Rac and Cdc42. Tiam1 directs IRSp53 to Rac signaling by enhancing IRSp53 binding to both active Rac and the **WAVE2 scaffold**.
YWHAZ (14-3-3 zeta)	scaffold	19218246	these results show that **14-3-3 (zeta)**:Shc **scaffold**s can act as multivalent signaling nodes for the integration of both phosphoserine/threonine and phosphotyrosine pathways to regulate specific cellular responses.

B

Disease associated genes identified by *BioSim*

Table B.1: Familial glioma of brain. Annotation terms shared by BRCA2 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	MSH2
STRING	Protein interactions	2	ERBB2
STRING	Protein interactions	2	PTEN
OMIM	MEDULLOBLASTOMA; MDB	8	DMBT1
GO BP	response to X-ray	12	MSH2
HIMAP	Indirect interaction	12	MSH2
OMIM	PROSTATE CANCER	14	PTEN
GO BP manual	regulation of epithelial cell proliferation	19	ERBB2
HIMAP	Indirect interaction	27	MSH2
GO BP	mammary gland development	29	ERBB2
GO BP	negative regulation of DNA metabolic process	30	MSH2
GO BP	response to ionizing radiation	34	MSH2
GO BP	double-strand break repair	40	MSH2
GO BP	aging	44	MSH2
HIMAP	Indirect interaction	45	MSH2
GO BP	regulation of epithelial cell proliferation	46	ERBB2
GO BP	response to UV	47	MSH2
GO MF	single-stranded DNA binding	49	MSH2
GO BP manual	response to organic substance	50	PPARG
GO BP	cell maturation	52	PPARG
GO BP	DNA damage response, signal transduction	54	MSH2
GO BP	response to nutrient	60	PPARG
GO BP	germ cell development	62	MSH2
GO BP	developmental maturation	67	PPARG
KEGG	Pancreatic cancer	72	ERBB2
GO BP	gonad development	72	MSH2
GO BP	reproductive structure development	74	MSH2
GO BP manual	response to hormone stimulus	77	PPARG
GO BP manual	response to endogenous stimulus	79	PPARG
GO BP	gland development	80	ERBB2
GO BP	regulation of DNA metabolic process	81	MSH2
GO BP	development of primary sexual characteristics	81	MSH2
GO BP	response to nutrient levels	92	PPARG
GO BP	DNA recombination	92	MSH2
GO BP	in utero embryonic development	96	MSH2
GO BP	response to extracellular stimulus	98	PPARG
GO BP	sex differentiation	102	MSH2
GO BP	response to light stimulus	105	MSH2
GO BP	response to organic substance	105	PPARG
GO MF	structure-specific DNA binding	105	MSH2

Table B.2: Epidermolytic palmoplantar keratoderma. Annotation terms shared by KRT1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
OMIM	Palmoplantar keratoderma, nonepidermolytic	2	KRT16
IntAct	Protein interaction	4	KRT9
IntAct	Protein interaction	16	KRT9
Keyword Disease	Palmoplantar keratoderma	16	KRT9, KRT16
IntAct	Protein interaction	17	KRT9
IntAct	Protein interaction	24	KRT9
IntAct	Indirect interaction	24	KRT9, KRT16
MINT	Indirect interaction	24	KRT9, KRT16
MINT	Indirect interaction	27	KRT9
InterPro	Prion protein	31	KRT9, KRT16
IntAct	Protein interaction	33	KRT9
IntAct	Indirect interaction	34	KRT9
IntAct	Protein interaction	35	KRT9
IntAct	Protein interaction	37	KRT9, KRT16
IntAct	Protein interaction	50	KRT9, KRT16
IntAct	Protein interaction	62	KRT9, KRT16
InterPro	Keratin, type I	63	KRT9, KRT16
GO MF manual	structural constituent of cytoskeleton	66	KRT9, KRT16
Pfam architecture	Filament	66	KRT9, KRT16
InterPro	Filament	68	KRT9, KRT16
HPRD	Indirect interaction	68	KRT9
GO MF	structural constituent of cytoskeleton	71	KRT9, KRT16
Pfam family	Intermediate filament protein	74	KRT9, KRT16
Keyword Cellular component	Intermediate filament	77	KRT9, KRT16
GO BP manual	epidermis development	80	KRT9, KRT16
GO BP manual	ectoderm development	88	KRT9, KRT16
GO BP manual	tissue development	136	KRT9, KRT16
Keyword Cellular component	Keratin	153	KRT9, KRT16
GO BP	epidermis development	154	KRT9, KRT16
InterPro	Prefoldin	159	KRT16
GO BP	ectoderm development	168	KRT9, KRT16
GO CC	intermediate filament	177	KRT9, KRT16
GO CC	intermediate filament cytoskeleton	178	KRT9, KRT16
GO BP	tissue development	331	KRT9, KRT16
GO MF manual	structural molecule activity	351	KRT9, KRT16
GO CC manual	cytoskeleton	477	KRT9, KRT16
GO BP manual	organ development	502	KRT9, KRT16
GO MF	structural molecule activity	596	KRT9, KRT16
GO CC	cytoskeletal part	772	KRT9, KRT16
GO BP manual	system development	869	KRT9, KRT16

Table B.3: Antley-Bixler syndrome. Annotation terms shared by FGFR1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
OMIM	Pfeiffer syndrome	2	FGFR2
OMIM	Jackson-Weiss syndromeS	2	FGFR2
PDB complexes	Indirect interaction	2	FGFR2
PDB complexes	Indirect interaction	3	FGFR2
IntAct	Indirect interaction	4	FGFR2
InterPro	Fibroblast growth factor receptor	4	FGFR2
HPRD	Indirect interaction	4	FGFR2
GO MF	fibroblast growth factor receptor activity	5	FGFR2
GO MF manual	fibroblast growth factor receptor activity	5	FGFR2
Keyword Disease	Craniosynostosis	8	FGFR2
HPRD	Indirect interaction	11	FGFR2
Pfam architecture	I-set	12	FGFR2
HPRD	Indirect interaction	13	FGFR2
HPRD	Indirect interaction	14	FGFR2
Reactome pathway	Signaling by FGFR	22	FGFR2
Reactome pathway	FGFR ligand binding and activation	22	FGFR2
GO BP manual	cell growth	22	FGFR2
GO BP manual	fibroblast growth factor receptor signaling pathway	25	FGFR2
GO BP	fibroblast growth factor receptor signaling pathway	26	FGFR2
GO BP manual	growth	29	FGFR2
GO BP	cell growth	41	FGFR2
GO MF manual	transmembrane receptor protein tyrosine kinase activity	47	FGFR2
ENZYME	Receptor protein-tyrosine kinase	52	FGFR2
GO MF manual	transmembrane receptor protein kinase activity	59	FGFR2
GO MF	transmembrane receptor protein tyrosine kinase activity	64	FGFR2
Keyword Ligand	Heparin-binding	64	FGFR2
GO MF manual	protein tyrosine kinase activity	69	FGFR2
GO MF	transmembrane receptor protein kinase activity	81	FGFR2
GO MF	heparin binding	83	FGFR2
KEGG	Prostate cancer	89	FGFR2
ENZYME	Protein-tyrosine kinases	89	FGFR2
GO BP manual	regulation of cell size	91	FGFR2

Table B.4: Cardiofaciocutaneous syndrome. Annotation terms shared by MAP2K1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	KRAS
STRING	Protein interactions	2	MAP2K2
STRING	Protein interactions	2	BRAF
IntAct	Protein interaction	3	BRAF

Table B.4: *(continued)*

annotation source	description	size	genes
HIMAP	Indirect interaction	3	MAP2K2
Reactome pathway	ERK activation	4	MAP2K2
IntAct	Protein interaction	4	MAP2K2, BRAF
HIMAP	Indirect interaction	5	MAP2K2
HIMAP	Indirect interaction	6	MAP2K2
IntAct	Protein interaction	7	MAP2K2
Reactome pathway	MEK activation	7	MAP2K2, KRAS
Reactome pathway	RAF phosphorylates MEK	7	MAP2K2, KRAS
MINT	Indirect interaction	7	MAP2K2
HIMAP	Indirect interaction	7	MAP2K2
ENZYME	Mitogen-activated protein kinase kinase	8	MAP2K2
HPRD	Indirect interaction	8	MAP2K2
Reactome pathway	MAP kinase cascade	9	MAP2K2, KRAS
IntAct	Proteins interacting with MEK1	10	BRAF
HIMAP	Indirect interaction	10	MAP2K2
Reactome pathway	SHC-mediated signaling	12	MAP2K2, KRAS
HIMAP	Indirect interaction	12	MAP2K2
Reactome pathway	SOS-mediated signaling	13	MAP2K2, KRAS
Reactome pathway	SHC-related events	14	MAP2K2, KRAS
Reactome pathway	Signaling to p38 via RIT and RIN	14	MAP2K2, KRAS, BRAF
Reactome pathway	Frs2-mediated activation	16	MAP2K2, KRAS, BRAF
Reactome pathway	Prolonged ERK activation events	17	MAP2K2, KRAS, BRAF
Reactome pathway	ARMS-mediated activation	17	MAP2K2, KRAS, BRAF
IntAct	Protein interaction	17	MAP2K2, BRAF
HPRD	Indirect interaction	17	MAP2K2
ENZYME	Dual-specificity kinases (those acting on Ser/Thr and Tyr	20	MAP2K2
HPRD	Indirect interaction	21	MAP2K2
IntAct	Indirect interaction	24	MAP2K2
IntAct	Protein interaction	26	MAP2K2
Reactome pathway	Signaling to RAS	26	MAP2K2, KRAS
KEGG	Thyroid cancer	28	MAP2K2, KRAS, BRAF
Reactome pathway	Signaling to ERKs	34	MAP2K2, KRAS, BRAF
Reactome pathway	IRS-mediated signaling	36	MAP2K2, KRAS
Reactome pathway	IRS-related events	38	MAP2K2, KRAS
Reactome pathway	Signaling by Insulin receptor	39	MAP2K2, KRAS
Reactome pathway	Insulin receptor signaling cascade	39	MAP2K2, KRAS
KEGG	Bladder cancer	42	MAP2K2, KRAS, BRAF

Table B.5: Folate-sensitive neural tube defects. Annotation terms shared by MTHFR and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
HIMAP	Indirect interaction	6	MTR, MTHFD1, MTRR

Table B.5: *(continued)*

annotation source	description	size	genes
HumanCyc	reductive acetyl coenzyme A pathway	8	MTHFD1
Reactome pathway	Metabolism of folate and pterins	8	MTHFD1
HIMAP	Indirect interaction	10	MTR, MTHFD1
HIMAP	Indirect interaction	10	MTR, MTHFD1, MTRR
HumanCyc	folate transformations	11	MTR, MTHFD1
GO BP	methionine metabolic process	12	MTR, MTHFD1, MTRR
HIMAP	Indirect interactions	14	MTR, MTHFD1, MTRR
ENZYME	Oxidoreductases acting on the CH-NH group of donors with NAD(+) or NADP(+) as acceptor	15	MTHFD1
KEGG	One carbon pool by folate	16	MTR, MTHFD1
GO MF	oxidoreductase activity, acting on the CH-NH group of donors, NAD or NADP as acceptor	17	MTHFD1
HIMAP	Indirect interaction	17	MTHFD1, MTRR
ENZYME	Oxidoreductases acting on the CH-NH group of donors.	21	MTHFD1
GO BP	sulfur amino acid metabolic process	22	MTR, MTHFD1, MTRR
GO BP	aspartate family amino acid metabolic process	23	MTR, MTHFD1, MTRR
HIMAP	Indirect interaction	25	MTR, MTRR
GO MF	oxidoreductase activity, acting on the CH-NH group of donors	27	MTHFD1
Reactome pathway	Metabolism of vitamins and cofactors	44	MTHFD1
Reactome pathway	Metabolism of water-soluble vitamins and cofactors	44	MTHFD1
GO BP	sulfur metabolic process	89	MTR, MTHFD1, MTRR
Keyword Ligand	Flavoprotein	101	MTRR
Keyword Ligand	FAD	110	MTRR
Keyword Ligand	NADP	153	MTHFD1, MTRR
GO BP	cellular amino acid metabolic process	218	MTR, MTHFD1, MTRR
GO BP	cellular amino acid and derivative metabolic process	292	MTR, MTHFD1, MTRR
GO MF manual	oxidoreductase activity	333	MTRR
GO BP	amine metabolic process	338	MTR, MTHFD1, MTRR
GO BP	nitrogen compound metabolic process	387	MTR, MTHFD1, MTRR
OrthoMCL	Campylobacter jejuni subsp. jejuni NCTC 11168	435	MTHFD1
OrthoMCL	Thermotoga maritima MSB8	439	MTR, MTHFD1
OrthoMCL	Wolinella succinogenes DSM 1740	447	MTR, MTHFD1
GO BP	carboxylic acid metabolic process	483	MTR, MTHFD1, MTRR
OrthoMCL	Streptococcus pneumoniae TIGR4	486	MTHFD1
GO BP	organic acid metabolic process	487	MTR, MTHFD1, MTRR
OrthoMCL	Aquifex aeolicus VF5	491	MTHFD1
OrthoMCL	Chlorobium tepidum TLS	516	MTR
Keyword Molecular function	Oxidoreductase	527	MTHFD1, MTRR
OrthoMCL	Synechococcus sp. WH 8102	543	MTR, MTHFD1
OrthoMCL	Coxiella burnetii RSA 493	558	MTHFD1

Table B.6: Obesity. Annotation terms shared by POMC and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	AGRP
STRING	Protein interactions	2	MC4R
STRING	Protein interactions	2	GHRL
STRING	Protein interactions	2	SDC3
STRING	Protein interactions	2	ENPP1
STRING	Protein interactions	2	MC3R
HPRD	Indirect interaction	3	AGRP
HPRD	Indirect interaction	3	MC4R
HPRD	Indirect interaction	8	AGRP
Keyword Disease	Obesity	29	ENPP1, CARTPT, ADRB3, UCP1, FTO, PPARG, MC4R, AGRP, UCP3, NR0B2
Keyword PTM	Amidation	44	GHRL
GO MF manual	hormone activity	55	AGRP
KEGG	Adipocytokine signaling pathway	67	AGRP
GO BP	regulation of blood pressure	77	CARTPT, ADRB2, ADRB3, PPARG
Keyword Molecular function	Hormone	83	GHRL
GO BP	neuropeptide signaling pathway	87	CARTPT, AGRP
GO MF	hormone activity	107	GHRL, AGRP
GO BP manual	generation of precursor metabolites and energy	142	ENPP1, ADRB3
GO BP	circulatory system process	151	CARTPT, ADRB2, PPARG
GO BP	blood circulation	154	CARTPT, ADRB2, ADRB3, PPARG
Keyword PTM	Cleavage on pair of basic residues	270	CARTPT
GO BP	generation of precursor metabolites and energy	306	ENPP1, ADRB3
GO BP manual	cell-cell signaling	389	GHRL

Table B.7: Autosomal recessive deafness-1A. Annotation terms shared by GJB6 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
OMIM	Deafness, autosomal dominant 3A; DFNA3A	2	GJB2
Pfam architecture	Connexin	3	GJB2
Keyword Disease	Palmoplantar keratoderma	16	GJB2
Pfam architecture	Connexin, Connexin_CCC	18	GJB2
InterPro	Connexin, N-terminal	20	GJB2
GO CC	connexon complex	21	GJB2
Pfam family	Connexin	21	GJB2
Pfam family	Gap junction channel protein cysteine-rich domain	21	GJB2
Keyword Disease	Ectodermal dysplasia	23	GJB2
Keyword Cellular component	Gap junction	26	GJB2
GO CC	gap junction	27	GJB2
GO BP manual	sensory perception of sound	47	GJB2

Table B.7: *(continued)*

annotation source	description	size	genes
GO BP manual	sensory perception of mechanical stimulus	47	GJB2
GO BP	sensory perception of sound	83	GJB2
GO BP	sensory perception of mechanical stimulus	86	GJB2
Keyword Disease	Deafness	98	GJB2
GO CC	cell-cell junction	156	GJB2
GO BP manual	sensory perception	229	GJB2
GO BP manual	cognition	249	GJB2
Keyword Cellular component	Cell junction	371	GJB2
GO BP manual	neurological system process	423	GJB2
GO CC	cell junction	461	GJB2
Mammalian Phenotype	hearing/vestibular/ear phenotype	480	GJB2
GO BP manual	system process	625	GJB2
GO BP	sensory perception	787	GJB2

Table B.8: Autosomal idiopathic short stature. Annotation terms shared by GHR and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
IntAct	Protein interaction	2	GH1
IntAct	Protein interaction	2	GH1
IntAct	Protein interaction	2	GH1
PDB complexes	human growth hormone bound to single receptor	2	GH1
PDB complexes	structural plasticity at the hgh:hghbp interface	2	GH1
PDB complexes	human growth hormone with its soluble binding protein	2	GH1
PDB complexes	human growth hormone mutant g120r with its soluble binding protein	2	GH1
PDB complexes	human growth hormone and extracellular domain of its receptor	2	GH1
DIP	Protein interaction	2	GH1
STRING	Protein interactions	2	GHSR
STRING	Protein interactions	2	GH1
GO BP manual	regulation of multicellular organism growth	4	GH1, GHSR
GO BP manual	positive regulation of multicellular organism growth	4	GH1, GHSR
GO BP	positive regulation of tyrosine phosphorylation of Stat5 protein	10	GH1
GO BP	response to estradiol stimulus	10	GH1
GO BP manual	response to estradiol stimulus	10	GH1
GO BP	positive regulation of tyrosine phosphorylation of Stat3 protein	11	GH1
GO BP	regulation of tyrosine phosphorylation of Stat5 protein	12	GH1
GO BP	cellular response to hormone stimulus	13	GHSR
GO MF	peptide hormone binding	15	GHSR
GO BP	regulation of tyrosine phosphorylation of Stat3 protein	17	GH1
GO BP manual	positive regulation of growth	19	GH1, GHSR
GO BP	positive regulation of multicellular organism growth	20	GH1, GHSR

Table B.8: *(continued)*

annotation source	description	size	genes
GO BP manual	response to estrogen stimulus	20	GH1
GO BP	positive regulation of tyrosine phosphorylation of STAT protein	21	GH1
GO BP	positive regulation of JAK-STAT cascade	23	GH1
GO BP	regulation of tyrosine phosphorylation of STAT protein	28	GH1
GO BP manual	response to steroid hormone stimulus	28	GH1
GO MF	hormone binding	29	GHSR
GO BP	regulation of JAK-STAT cascade	34	GH1
Keyword Disease	Dwarfism	34	GH1, SHOX
GO BP	response to estrogen stimulus	36	GH1
GO BP	positive regulation of peptidyl-tyrosine phosphorylation	37	GH1
GO BP	positive regulation of growth	43	GH1, GHSR
GO BP	regulation of multicellular organism growth	46	GH1, GHSR
GO BP	regulation of peptidyl-tyrosine phosphorylation	52	GH1
GO BP	positive regulation of protein amino acid phosphorylation	54	GH1
GO BP	response to steroid hormone stimulus	60	GH1
GO BP	positive regulation of phosphorylation	62	GH1
GO BP	positive regulation of phosphorus metabolic process	63	GH1

Table B.9: Hypogonadotropic hypogonadism. Annotation terms shared by FGFR1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	KISS1R
STRING	Protein interactions	2	PROK2
STRING	Protein interactions	2	NELF
Keyword Disease	Kallmann syndrome	4	PROK2
GO BP manual	MAPKKK cascade	95	PROK2
GO BP	MAPKKK cascade	149	PROK2
GO BP manual	protein kinase cascade	211	PROK2
GO BP manual	protein amino acid phosphorylation	286	PROK2
GO BP	protein kinase cascade	309	PROK2
GO BP manual	phosphorylation	359	PROK2
GO BP manual	phosphate metabolic process	444	PROK2
GO BP manual	phosphorus metabolic process	444	PROK2
GO BP manual	post-translational protein modification	494	PROK2
GO BP manual	regulation of biological quality	563	PROK2
GO MF manual	transmembrane receptor activity	589	KISS1R
GO BP	regulation of cell proliferation	598	KISS1R
Mammalian Phenotype	renal/urinary system phenotype	605	KISS1R
GO BP	protein amino acid phosphorylation	609	PROK2
GO BP manual	intracellular signaling cascade	631	PROK2
GO BP manual	cell surface receptor linked signal transduction	674	KISS1R

Table B.9: *(continued)*

annotation source	description	size	genes
GO BP manual	protein modification process	679	PROK2
GO BP manual	biopolymer modification	706	PROK2
GO BP	phosphorylation	740	PROK2
GO MF manual	receptor activity	750	KISS1R
Mammalian Phenotype	digestive/alimentary phenotype	790	KISS1R
GO BP manual	system development	869	PROK2
GO BP	phosphate metabolic process	899	PROK2
GO BP	phosphorus metabolic process	899	PROK2
GO BP manual	cellular protein metabolic process	996	PROK2
GO MF manual	molecular transducer activity	996	KISS1R
GO MF manual	signal transducer activity	996	KISS1R
GO BP manual	anatomical structure development	1002	PROK2
GO BP	post-translational protein modification	1031	PROK2
GO BP	regulation of biological quality	1058	PROK2
GO BP manual	multicellular organismal development	1082	PROK2
GO BP	intracellular signaling cascade	1134	PROK2
GO BP manual	protein metabolic process	1158	PROK2
GO MF	transmembrane receptor activity	1247	KISS1R
GO BP manual	developmental process	1268	PROK2
GO BP	protein modification process	1299	PROK2

Table B.10: Noninsulin-dependent diabetes mellitus. Annotation terms shared by PPARG and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	RETN
STRING	Protein interactions	2	IL6
STRING	Protein interactions	2	SLC30A8
STRING	Protein interactions	2	IRS1
STRING	Protein interactions	2	GPD2
STRING	Protein interactions	2	KCNJ11
STRING	Protein interactions	2	IRS2
GO BP manual	cellular response to insulin stimulus	4	ENPP1
GO BP manual	fat cell differentiation	6	TCF7L2
GO BP manual	response to insulin stimulus	7	ENPP1
GO BP manual	response to peptide hormone stimulus	7	ENPP1
GO BP	cellular response to insulin stimulus	8	ENPP1
GO BP manual	cellular response to hormone stimulus	8	ENPP1
GO BP	regulation of fat cell differentiation	9	ENPP1
HPRD	Indirect interaction	12	HNF4A
GO BP	cellular response to hormone stimulus	13	ENPP1
GO BP manual	glucose homeostasis	15	TCF7L2, WFS1, HNF1A

Table B.10: *(continued)*

annotation source	description	size	genes
GO BP manual	carbohydrate homeostasis	15	TCF7L2, WFS1, HNF1A
Pfam architecture	zf-C4	15	HNF4A
OMIM	OBESITY	17	ENPP1
GO BP	response to insulin stimulus	17	ENPP1
GO BP manual	lipid homeostasis	19	HNF4A
HPRD	Indirect interaction	19	HNF4A
HPRD	Indirect interaction	23	HNF4A
GO BP	fat cell differentiation	26	TCF7L2
GO MF	fatty acid binding	26	HNF4A
HPRD	Indirect interaction	26	HNF4A
KEGG	Thyroid cancer	28	TCF7L2
GO BP manual	response to nutrient	29	GCGR
Keyword Disease	Obesity	29	RETN, ENPP1
GO BP	response to peptide hormone stimulus	34	IRS2, ENPP1, IRS1
GO BP	glucose homeostasis	36	TCF7L2, WFS1, HNF1A, GCK, PDX1, NEUROD1
GO BP	lipid homeostasis	36	LIPC, HNF4A
GO BP	carbohydrate homeostasis	36	TCF7L2, WFS1, HNF1A, GCK, PDX1, NEUROD1
GO BP manual	regulation of blood pressure	38	GCGR
Pfam architecture	zf-C4, Hormone_recep	38	HNF4A
GO BP manual	response to nutrient levels	40	GCGR
HIMAP	Indirect interaction	42	HNF4A
Keyword Disease	Diabetes mellitus	42	WFS1, HNF1A, ABCC8, KCNJ11, GCK, GCGR, HNF4A, RETN, ENPP1, MAPK8IP1, HNF1B, PDX1, IRS1, NEUROD1, SLC30A8
HPRD	Indirect interaction	43	HNF4A
Pfam family	Zinc finger, C4 type (two domains)	44	HNF4A
InterPro	Vitamin D receptor	45	HNF4A
GO BP manual	response to extracellular stimulus	45	GCGR
InterPro	Zinc finger, nuclear hormone receptor-type	46	HNF4A
InterPro	Steroid hormone receptor	46	HNF4A
Pfam family	Ligand-binding domain of nuclear hormone receptor	46	HNF4A
InterPro	Nuclear hormone receptor, ligand-binding, core	47	HNF4A
InterPro	Nuclear hormone receptor, ligand-binding	47	HNF4A
GO MF	steroid hormone receptor activity	49	HNF4A

Table B.11: Susceptibility to atypical hemolytic uremic syndrome-1. Annotation terms shared by CFI and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
HIMAP	Indirect interaction	9	CFH
HIMAP	Indirect interaction	9	CFH
HIMAP	Indirect interaction	15	CFH
HIMAP	Indirect interaction	26	CFH
GO BP	complement activation, classical pathway	28	CD46
Keyword Biological process	Complement pathway	28	CD46
GO BP	humoral immune response mediated by circulating immunoglobulin	29	CD46
HIMAP	Indirect interaction	30	CFH, CD46
HIMAP	Indirect interaction	32	CFH, CD46
HIMAP	Indirect interaction	35	CFH, CD46
GO BP	activation of plasma proteins involved in acute inflammatory response	36	CFH, CD46
GO BP	complement activation	36	CFH, CD46
HIMAP	Indirect interaction	37	CFH, CD46
HIMAP	Indirect interaction	41	CFH, CD46
GO BP	immunoglobulin mediated immune response	47	CD46
GO BP	B cell mediated immunity	48	CD46
HIMAP	Indirect interaction	49	CFH, CD46
HIMAP	Indirect interaction	49	CFH, CD46
Keyword Biological process	Innate immunity	56	CFH, CD46
GO BP	lymphocyte mediated immunity	59	CD46
HIMAP neighbors	Indirect interactioncomplement component 1, r subcomponent	64	CFH, CD46
GO BP	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	65	CD46
GO BP	adaptive immune response	65	CD46
GO BP	activation of immune response	66	CFH, CD46
KEGG	Complement and coagulation cascades	68	CFH, CD46
GO BP	leukocyte mediated immunity	68	CD46
GO BP	humoral immune response	68	CFH, CD46
GO BP	acute inflammatory response	74	CFH, CD46
GO BP	positive regulation of immune response	96	CFH, CD46
GO BP	immune effector process	100	CFH, CD46
GO BP	innate immune response	122	CFH, CD46
GO BP	positive regulation of response to stimulus	137	CFH, CD46
GO BP	regulation of immune response	152	CFH, CD46
GO BP	positive regulation of immune system process	166	CFH, CD46
Keyword Biological process	Immune response	199	CFH, CD46
GO BP	regulation of response to stimulus	245	CFH, CD46
GO BP	regulation of immune system process	270	CFH, CD46
GO BP	inflammatory response	274	CFH, CD46
GO BP	response to wounding	408	CFH, CD46
GO BP	defense response	539	CFH, CD46

Table B.12: Noninsulin-dependent diabetes mellitus. Annotation terms shared by SLC2A4 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
Reactome pathway	Glucose uptake	9	GCK
GO BP manual	glucose homeostasis	15	TCF7L2, WFS1, HNF1A
GO BP manual	carbohydrate homeostasis	15	TCF7L2, WFS1, HNF1A
IntAct neighbors	Indirect interactionDeath domain-associated protein 6	30	TCF7L2
GO BP	glucose homeostasis	36	TCF7L2, WFS1, HNF1A, GCK, PDX1, NEUROD1
GO BP	carbohydrate homeostasis	36	TCF7L2, WFS1, HNF1A, GCK, PDX1, NEUROD1
KEGG	Type II diabetes mellitus	42	ABCC8, KCNJ11, IRS2, PDX1, IRS1
Keyword Disease	Diabetes mellitus	42	WFS1, HNF1A, ABCC8, KCNJ11, GCK, GCGR, HNF4A, RETN, ENPP1, MAPK8IP1, HNF1B, PDX1, IRS1, NEUROD1, SLC30A8
Reactome pathway	Metabolism of vitamins and cofactors	44	ENPP1
Reactome pathway	Metabolism of water-soluble vitamins and cofactors	44	ENPP1
KEGG	Adipocytokine signaling pathway	67	AKT2, IRS2, IRS1
Reactome pathway	Glucose metabolism	70	GCK
HPRD	Indirect interaction	73	AKT2
Reactome pathway	Metabolism of carbohydrates	87	GCK
GO CC manual	cell surface	123	VEGFA, ENPP1
KEGG	Insulin signaling pathway	136	AKT2, IRS2, IRS1
GO CC manual	membrane-bounded vesicle	174	VEGFA, SLC30A8
GO CC	microsome	176	KCNJ11, IRS1
GO CC	vesicular fraction	182	KCNJ11, IRS1
GO CC manual	vesicle	184	VEGFA, SLC30A8
GO BP manual	chemical homeostasis	199	TCF7L2, WFS1, HNF1A, KCNJ11, HNF4A, ENPP1, SLC30A8
GO CC	cell surface	234	VEGFA, ENPP1
GO BP manual	carbohydrate metabolic process	247	ABCC8, KCNJ11
GO BP manual	homeostatic process	251	TCF7L2, WFS1, HNF1A, KCNJ11, HNF4A, ENPP1, IL6, SLC30A8
GO CC manual	endomembrane system	296	WFS1
Mammalian Phenotype	adipose tissue phenotype	299	HNF1A, GPD2, GCGR, RETN, AKT2, IRS2, IRS1
GO BP	chemical homeostasis	361	LIPC, TCF7L2, WFS1, HNF1A, KCNJ11, GCK, HNF4A, ENPP1, PDX1, NEUROD1, SLC30A8
GO CC	cytoplasmic membrane-bounded vesicle	400	VEGFA, SLC30A8
GO MF manual	substrate-specific transmembrane transporter activity	404	ABCC8, SLC30A8
GO CC	membrane-bounded vesicle	407	VEGFA, SLC30A8
GO CC manual	organelle membrane	424	WFS1
Keyword PTM	Ubl conjugation	428	TCF7L2, MAPK8IP1
GO MF manual	transmembrane transporter activity	437	ABCC8, SLC30A8

Table B.12: *(continued)*

annotation source	description	size	genes
GO MF manual	substrate-specific transporter activity	471	ABCC8, SLC30A8
GO BP	carbohydrate metabolic process	472	ABCC8, KCNJ11, GCK, GPD2, PDX1
GO CC	cytoplasmic vesicle	522	VEGFA, SLC30A8
GO BP	homeostatic process	529	LIPC, TCF7L2, WFS1, HNF1A, KCNJ11, GCK, HNF4A, ENPP1, PDX1, NEUROD1, IL6, SLC30A8
GO CC	vesicle	534	VEGFA, SLC30A8
GO BP manual	regulation of biological quality	563	TCF7L2, WFS1, HNF1A, KCNJ11, GCGR, HNF4A, VEGFA, ENPP1, PDX1, IL6, SLC30A8
GO MF manual	transporter activity	575	ABCC8, SLC30A8

Table B.13: Autosomal recessive deafness-1A. Annotation terms shared by GJB3 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
GO MF	gap junction channel activity	7	GJB2
GO MF	wide pore channel activity	8	GJB2
Pfam architecture	Connexin, Connexin_CCC	18	GJB2
InterPro	Connexin, N-terminal	20	GJB2
GO CC	connexon complex	21	GJB2
Pfam family	Connexin	21	GJB2
Pfam family	Gap junction channel protein cysteine-rich domain	21	GJB2
Keyword Cellular component	Gap junction	26	GJB2
GO CC	gap junction	27	GJB2
Keyword Disease	Deafness	98	GJB2
GO CC	cell-cell junction	156	GJB2
Keyword Cellular component	Cell junction	371	GJB2
GO MF	channel activity	394	GJB2
GO MF	passive transmembrane transporter activity	394	GJB2
Mammalian Phenotype	no phenotypic analysis	424	GJB2
GO CC	cell junction	461	GJB2
GO MF	transmembrane transporter activity	868	GJB2
Mammalian Phenotype	embryogenesis phenotype	992	GJB2

Table B.14: Autosomal recessive dyskeratosis congenita. Annotation terms shared by NHP2 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	NOP10
GO BP	pseudouridine synthesis	14	NOP10
GO CC	small nucleolar ribonucleoprotein complex	16	NOP10
GO BP	RNA modification	39	NOP10
Keyword Biological process	Ribosome biogenesis	45	NOP10
Keyword Biological process	rRNA processing	54	NOP10
GO BP	rRNA processing	83	NOP10
GO BP	rRNA metabolic process	86	NOP10
GO BP	ribosome biogenesis	113	NOP10
GO BP	ncRNA processing	157	NOP10
GO BP	ribonucleoprotein complex biogenesis	164	NOP10
GO BP	ncRNA metabolic process	201	NOP10
Keyword Molecular function	Ribonucleoprotein	305	NOP10
GO CC	nucleolus	399	NOP10
GO CC	ribonucleoprotein complex	422	NOP10
GO BP	RNA processing	505	NOP10
GO BP	RNA metabolic process	838	NOP10

Table B.15: Orofacial cleft-1. Annotation terms shared by MTHFR and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
HIMAP	Indirect interaction	6	MTR
HIMAP	Indirect interaction	10	MTR
HumanCyc	folate transformations	11	MTR
GO BP	methionine metabolic process	12	MTR
HIMAP neighbors	Indirect interactionspermine synthase	14	MTR
KEGG	One carbon pool by folate	16	MTR
GO BP	sulfur amino acid metabolic process	22	MTR
GO BP	aspartate family amino acid metabolic process	23	MTR
HIMAP	Indirect interaction	25	MTR
GO BP	sulfur metabolic process	89	MTR
GO BP	cellular amino acid metabolic process	218	MTR
GO BP	cellular amino acid and derivative metabolic process	292	MTR
GO BP	amine metabolic process	338	MTR
GO BP	nitrogen compound metabolic process	387	MTR
OrthoMCL	Thermotoga maritima MSB8	439	MTR
OrthoMCL	Wolinella succinogenes DSM 1740	447	MTR
GO BP	carboxylic acid metabolic process	483	MTR
GO BP	organic acid metabolic process	487	MTR
OrthoMCL	Chlorobium tepidum TLS	516	MTR

Table B.15: *(continued)*

annotation source	description	size	genes
OrthoMCL	Synechococcus sp. WH 8102	543	MTR
OrthoMCL	Geobacter sulfurreducens PCA	618	MTR
OrthoMCL	Brucella suis 1330	705	MTR
OrthoMCL	Yersinia pestis CO92	733	MTR
GO CC manual	cytosol	736	MTR
OrthoMCL	Vibrio cholerae O1 biovar eltor str. N16961	743	MTR
OrthoMCL	Shigella flexneri 2a str. 301	758	MTR
OrthoMCL	Salmonella enterica subsp. enterica serovar Typhi str. CT18	781	MTR
OrthoMCL	Ralstonia solanacearum GMI1000	812	MTR
OrthoMCL	Escherichia coli W3110	817	MTR
OrthoMCL	Agrobacterium tumefaciens str. C58	830	MTR
GO CC	cytosol	892	MTR
OrthoMCL	Rhodopirellula baltica SH 1	894	MTR

Table B.16: Alzheimer disease. Annotation terms shared by APP and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	desc	size	genes
IntAct	Protein interaction	2	APBB2
IntAct	Protein interaction	2	APBB2
DIP	Protein interaction	2	A2M
STRING	Protein interactions	2	SORL1
STRING	Protein interactions	2	BLMH
STRING	Protein interactions	2	APBB2
STRING	Protein interactions	2	A2M
HPRD	Indirect interaction	16	A2M
HPRD	Indirect interaction	19	PLAU
GO CC	platelet alpha granule lumen	33	A2M
GO CC manual	platelet alpha granule lumen	33	A2M
GO CC	cytoplasmic membrane-bounded vesicle lumen	35	A2M
GO CC	vesicle lumen	35	A2M
GO CC manual	cytoplasmic membrane-bounded vesicle lumen	35	A2M
GO CC manual	vesicle lumen	35	A2M
GO CC	platelet alpha granule	45	A2M
GO CC manual	platelet alpha granule	45	A2M
HPRD	Indirect interaction	53	A2M
Reactome pathway	Exocytosis of Alpha granule	59	A2M
GO BP	extracellular matrix organization	60	APBB2
Reactome pathway	Platelet degranulation	61	A2M
Reactome pathway	Response to elevated platelet cytosolic Ca++	65	A2M
GO CC manual	secretory granule	68	A2M
HPRD	Indirect interaction	70	NOS3

Table B.16: *(continued)*

annotation source	desc	size	genes
GO BP	axon guidance	78	APBB2
Reactome pathway	Platelet Activation	80	A2M
GO CC manual	cytoplasmic vesicle part	80	A2M
Reactome pathway	Formation of Platelet plug	83	A2M
Keyword Molecular function	Serine protease inhibitor	84	A2M
Keyword Biological process	Endocytosis	86	SORL1
GO CC	apical part of cell	92	NOS3
GO MF	serine-type endopeptidase inhibitor activity	93	A2M

Table B.17: Susceptibility to atypical hemolytic uremic syndrome-1. Annotation terms shared by CFHR1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	CFH
HIMAP	Indirect interaction	32	CFH, CD46
HIMAP	indirect interaction	35	CFH, CD46
GO BP	activation of plasma proteins involved in acute inflammatory response	36	CFH, CD46
GO BP	complement activation	36	CFH, CD46
HIMAP	Indirect interaction	37	CFH, CD46
HIMAP	Indirect interaction	39	CFH, CD46
HIMAP	Indirect interaction	40	CFH, CD46
HIMAP	Indirect interaction	41	CFH, CD46
HIMAP	Indirect interaction	49	CFH, CD46
Pfam family	Sushi domain (SCR repeat)	49	CFH, CD46
InterPro	Sushi/SCR/CCP	54	CFH, CD46
InterPro	Complement control module	55	CFH, CD46
Keyword Domain	Sushi	56	CFH, CD46
GO BP	activation of immune response	66	CFH, CD46
GO BP	humoral immune response	68	CFH, CD46
GO BP	acute inflammatory response	74	CFH, CD46
GO BP	positive regulation of immune response	96	CFH, CD46
GO BP	immune effector process	100	CFH, CD46
GO BP	positive regulation of response to stimulus	137	CFH, CD46
GO BP	regulation of immune response	152	CFH, CD46
GO BP	positive regulation of immune system process	166	CFH, CD46
GO BP	regulation of response to stimulus	245	CFH, CD46
GO BP	regulation of immune system process	270	CFH, CD46
GO BP	inflammatory response	274	CFH, CD46
GO CC manual	extracellular space	292	CFH
GO CC manual	extracellular region part	401	CFH
GO BP	response to wounding	408	CFH, CD46

Table B.17: *(continued)*

annotation source	description	size	genes
GO BP	defense response	539	CFH, CD46
GO CC	extracellular space	587	CFH
GO BP	immune response	616	CFH, CD46
GO BP	response to external stimulus	677	CFH, CD46
GO CC manual	extracellular region	726	CFH
GO CC	extracellular region part	860	CFH
GO BP	immune system process	870	CFH, CD46

Table B.18: Endometrial cancer. Annotation terms shared by MHL3 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	MSH6
Keyword Disease	Hereditary nonpolyposis colorectal cancer	7	MSH6
GO MF	mismatched DNA binding	16	MSH6
GO BP	mismatch repair	22	MSH6
KEGG	Mismatch repair	23	MSH6
GO BP	meiosis I	33	MSH6
GO MF	double-stranded DNA binding	62	MSH6
GO CC	nuclear chromosome part	71	MSH6
GO BP	M phase of meiotic cell cycle	79	MSH6
GO BP	meiosis	79	MSH6
GO BP	meiotic cell cycle	81	MSH6
GO BP	DNA recombination	92	MSH6
GO CC	nuclear chromosome	102	MSH6
GO MF	structure-specific DNA binding	105	MSH6
GO MF	chromatin binding	115	MSH6
Keyword Biological process	DNA repair	160	MSH6
Keyword Biological process	DNA damage	178	MSH6
GO BP	DNA repair	254	MSH6
GO BP	M phase	265	MSH6
GO BP manual	DNA metabolic process	275	MSH6
GO BP	cellular response to DNA damage stimulus	288	MSH6
GO CC	chromosomal part	295	MSH6
GO BP	response to DNA damage stimulus	327	MSH6
GO BP	cell cycle phase	336	MSH6
GO BP	cellular response to stress	340	MSH6
GO BP	cellular response to stimulus	363	MSH6
GO CC	chromosome	366	MSH6
GO BP	chromosome organization	390	MSH6
GO BP	DNA metabolic process	445	MSH6
GO BP	cell cycle process	476	MSH6

Table B.18: *(continued)*

annotation source	description	size	genes
GO BP	cell cycle	671	MSH6
GO MF manual	DNA binding	818	MSH6
GO BP manual	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	882	MSH6
GO BP	organelle organization	974	MSH6
GO MF manual	nucleic acid binding	1096	MSH6
GO MF	ATP binding	1314	MSH6
GO BP	response to stress	1329	MSH6
GO MF	adenyl ribonucleotide binding	1329	MSH6
GO CC	nuclear part	1377	MSH6
GO MF	adenyl nucleotide binding	1409	MSH6

Table B.19: Susceptibility to atypical hemolytic uremic syndrome-1. Annotation terms shared by CFHR3 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	3	CFH
STRING	Protein interactions	3	CFH
Pfam architecture	Sushi, Sushi, Sushi, Sushi	5	CD46
HPRD neighbors	Indirect interaction	24	CFH, CD46
Pfam family	Sushi domain (SCR repeat)	49	CFH, CD46
InterPro	Sushi/SCR/CCP	54	CFH, CD46
InterPro	Complement control module	55	CFH, CD46
Keyword Domain	Sushi	56	CFH, CD46
GO CC manual	extracellular space	292	CFH
GO CC manual	extracellular region part	401	CFH
GO CC	extracellular space	587	CFH
GO CC manual	extracellular region	726	CFH
GO CC	extracellular region part	860	CFH

Table B.20: Mitochondrial neurogastrointestinal encephalopathy syndrome. Annotation terms shared by POLG and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	TYMP
GO BP	mitochondrial genome maintenance	5	TYMP
GO BP	mitochondrion organization	93	TYMP
GO BP	DNA replication	170	TYMP

Table B.20: *(continued)*

annotation source	description	size	genes
GO BP manual	DNA metabolic process	275	TYMP
GO BP	DNA metabolic process	445	TYMP
GO BP	cellular biopolymer biosynthetic process	750	TYMP
GO BP	biopolymer biosynthetic process	751	TYMP
GO MF manual	transferase activity	794	TYMP
GO BP manual	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	882	TYMP
GO BP	cellular macromolecule biosynthetic process	945	TYMP
GO BP	organelle organization	974	TYMP
GO BP	macromolecule biosynthetic process	986	TYMP
ENZYME	Transferases	1109	TYMP
Keyword Disease	Disease mutation	1509	TYMP
Keyword Molecular function	Transferase	1535	TYMP
GO BP	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1596	TYMP
GO BP	cellular biosynthetic process	1603	TYMP
Mammalian Phenotype	immune system phenotype	1611	TYMP
GO MF	transferase activity	1645	TYMP
GO BP	biosynthetic process	1687	TYMP
GO BP manual	cellular biopolymer metabolic process	1716	TYMP
GO BP manual	cellular macromolecule metabolic process	1770	TYMP
GO BP	cellular component organization	1840	TYMP
Mammalian Phenotype	nervous system phenotype	1865	TYMP
GO BP manual	biopolymer metabolic process	1890	TYMP
GO BP manual	macromolecule metabolic process	1941	TYMP
GO CC manual	cytoplasmic part	2149	TYMP
GO MF manual	catalytic activity	2332	TYMP
GO BP manual	cellular metabolic process	2467	TYMP
GO BP manual	primary metabolic process	2500	TYMP
GO BP	developmental process	2638	TYMP
GO BP manual	metabolic process	2802	TYMP
GO BP	response to stimulus	2843	TYMP
GO CC manual	cytoplasm	3042	TYMP
GO BP	cellular biopolymer metabolic process	3304	TYMP
GO BP	cellular macromolecule metabolic process	3444	TYMP
GO BP	biopolymer metabolic process	3809	TYMP
GO BP	macromolecule metabolic process	3961	TYMP
GO CC	cytoplasmic part	4118	TYMP

Table B.21: Colorectal cancer. Annotation terms shared by CCND1 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
STRING	Protein interactions	2	APC
STRING	Protein interactions	2	EP300
STRING	Protein interactions	2	PIK3CA
STRING	Protein interactions	2	TP53
STRING	Protein interactions	2	AKT1
STRING	Protein interactions	2	AXIN2
GO BP	response to UV-A	3	AKT1
GO BP manual	response to UV-A	3	AKT1
GO BP manual	positive regulation of cyclin-dependent protein kinase activity	6	AKT1
GO BP	positive regulation of cyclin-dependent protein kinase activity	7	AKT1
IntAct neighbors	Indirect interactionSerine/threonine-protein phosphatase PP1-gamma catalytic subunit	16	TP53
HPRD	Indirect interaction	18	EP300
HPRD	Indirect interaction	21	EP300
MINT	Indirect interaction	22	TP53
GO BP	response to endoplasmic reticulum stress	23	TP53
GO BP	ER-nuclear signaling pathway	26	TP53
KEGG	Thyroid cancer	28	TP53, NRAS
GO BP manual	positive regulation of cell cycle	28	AKT1
GO BP manual	response to UV	31	AKT1
GO BP manual	G1/S transition of mitotic cell cycle	33	AKT1
HPRD	Indirect interaction	34	AKT1
HPRD	Indirect interaction	34	TP53
HPRD	Indirect interaction	36	EP300
HIMAP	Indirect interaction	40	AURKA, BUB1B
KEGG	Bladder cancer	42	TP53, NRAS
GO BP	positive regulation of cell cycle	43	AKT1
HPRD	Indirect interaction	45	TP53, EP300
GO BP	G1/S transition of mitotic cell cycle	46	AKT1
HPRD	Indirect interaction	46	AKT1
GO BP	response to UV	47	TP53, AKT1
GO BP manual	regulation of cyclin-dependent protein kinase activity	47	APC, AKT1

Table B.22: Osteogenic sarcoma. Annotation terms shared by TP53 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
MINT		2	CHEK2
STRING	Protein interactions	2	RB1
STRING	Protein interactions	2	CHEK2
PDB complexes	Indirect interaction	3	RB1

Table B.22: *(continued)*

annotation source	description	size	genes
Keyword Disease	Li-Fraumeni syndrome	3	CHEK2
IntAct	Indirect interaction	4	RB1
MINT	Indirect interaction	5	RB1
HPRD	Indirect interaction	6	RB1
HPRD	Indirect interaction	7	CHEK2
IntAct	Indirect interaction	11	CHEK2
GO BP	response to gamma radiation	14	CHEK2
GO BP manual	DNA damage response, signal transduction resulting in induction of apoptosis	15	CHEK2
OMIM	Breast cancer	16	CHEK2
HPRD	Indirect interaction	16	RB1, CHEK2
MINT	Indirect interaction	17	RB1, CHEK2
IntAct	Indirect interaction	22	RB1
GO BP	DNA damage response, signal transduction resulting in induction of apoptosis	23	CHEK2
GO CC manual	PML body	24	RB1, CHEK2
GO CC	PML body	25	RB1, CHEK2
GO BP manual	induction of apoptosis by intracellular signals	28	CHEK2
GO BP	DNA damage checkpoint	32	CHEK2
GO BP	response to ionizing radiation	34	CHEK2
GO BP	DNA integrity checkpoint	35	CHEK2
KEGG	Bladder cancer	42	RB1
GO BP	induction of apoptosis by intracellular signals	42	CHEK2
GO BP manual	DNA damage response, signal transduction	42	CHEK2

Table B.23: Mitochondrial complex I deficiency. Annotation terms shared by NDUFA11 and the disease-associated genes. The column 'size' gives the total number of genes sharing the respective annotation.

annotation source	description	size	genes
CORUM	Respiratory chain complex I (lambda subunit) mitochondrial	16	NDUFS6, NDUFV1, NDUFS1, NDUFS2, NDUFS4
CORUM	Respiratory chain complex I (holoenzyme), mitochondrial	44	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
HumanCyc	NAD/NADH phosphorylation and dephosphorylation	44	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
Keyword Biological process	Respiratory chain	58	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
Reactome pathway	Electron Transport Chain	76	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
Keyword Biological process	Electron transport	101	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO BP	electron transport chain	113	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
KEGG	Oxidative phosphorylation	128	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
Keyword Cellular component	Mitochondrion inner membrane	186	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	mitochondrial inner membrane	280	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4

Table B.23: *(continued)*

annotation source	description	size	genes
GO CC	organelle inner membrane	297	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO BP	generation of precursor metabolites and energy	306	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	mitochondrial membrane	359	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	mitochondrial envelope	378	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	mitochondrial part	535	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	organelle envelope	548	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	envelope	549	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO BP	oxidation reduction	604	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
Keyword Cellular component	Mitochondrion	793	NDUFAF4, NDUFS6, NDUFA1, NDUFV1, NDUFAF2, NDUFS1, NDUFS2, NDUFS4
GO CC	organelle membrane	933	NDUFS6, NDUFA1, NDUFV1, NDUFS1, NDUFS2, NDUFS4
GO CC	mitochondrion	989	NDUFAF4, NDUFS6, NDUFA1, NDUFV1, NDUFAF2, NDUFS1, NDUFS2, NDUFS4

C

List of own publications

Journal Publications

- Ramírez, F., Lawyer, G., and Albrecht, M. (2011). Novel search method for the discovery of functional relationships. *Bioinformatics, in press.*

- Reiss, S., Rebhan, I., Backes, P., Romero-Brey, I., Erfle, H., Matula, P., Kaderali, L., Poenisch, M., Blankenburg, H., Hiet, MS., Longerich, T., Diehl, S., Ramírez, F., Balla, T., Rohr, K., Kaul, A., Bühler, S., Pepperkok, R., Lengauer, T., Albrecht, M., Eils, R., Schirmacher, P., Lohmann, V., and Bartenschlager, R. (2011). Recruitment and activation of a lipid kinase by hepatitis C virus NS5A is essential for integrity of the membranous replication compartment. *Cell Host Microbe*, 9(1):32-45.

- Ramírez, F. and Albrecht, M. (2010) Finding scaffold proteins in interactomes. *Trends in Cell Biology*, 20(1):2-4.

- Blankenburg, H., Ramírez, F., Büch, J., and Albrecht, M. (2009). DASMIweb: online integration, analysis and assessment of distributed protein interaction data. *Nucleic Acids Research*, 37(Web Server issue), W122-8.

- Blankenburg, H., Finn, R., Prlic, A., Jenkinson, A., Ramírez, F., Emig, D., Schelhorn, S., Büch J., Lengauer T., and Albrecht, M. (2009). DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics*, 25(10):1321-8.

- Assenov, Y., Ramírez, F., Schelhorn, S., Lengauer, T., and Albrecht, M. (2008). Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282-4.

- Tress, M., Martelli, P., Frankish, A., Reeves, G., Wesselink, J., Yeats, C., Olason, PI., Albrecht, M., Hegyi, H., Giorgetti, A., Raimondo, D., Lagarde, J.,

Laskowski, R., López, G., Sadowski, M., Watson, J., Fariselli, P., Rossi, I., Nagy, A., Kai, W., Størling, Z., Orsini, M., Assenov, Y., Blankenburg, H., Huthmacher, C., Ramírez, F., Schlicker, A., Denoeud, F., Jones, P., Kerrien, S., Orchard, S., Antonarakis, S., Reymond, A., Birney, E., Brunak, S., Casadio, R., Guigo, R., Harrow, J., Hermjakob, H., Jones, D., Lengauer, T., Orengo, C., Patthy, L., Thornton, J., Tramontano, A., and Valencia, A. (2007). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America*, 104(13):5495-500.

- Schlicker, A., Huthmacher, C., Ramírez, F., Lengauer, T., and Albrecht, M. (2007). Functional evaluation of domain-domain interactions and human protein interaction networks. *Bioinformatics*, 23(7), 859-65.

- Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T., and Albrecht, M. (2007). Computational analysis of human protein interaction networks. *Proteomics*. 7(15):2541-52.