Knowledge Management and Discovery for Genotype/Phenotype Data

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

im Fach Informatik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät II

von

M.Sc. Bioinf. Philip Groth

geboren am 15.12.1978 in Berlin

Präsident der Humboldt Universität zu Berlin

Prof. Dr. Christoph Markschies

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II

Prof. Dr. Peter Frensch

Gutachter/Gutachterin

1. Prof. Dr. Ulf Leser

2. Prof. Dr. Hanspeter Herzel

3. Dr. Bertram Weiss

Tag der Verteidigung: 26. November 2009

“If you try and take a cat apart to see how it works, the first thing you have on your hands is a non-working cat. Life is a level of complexity that almost lies outside our vision…”

- Douglas Adams

Abstract

Phenotypes often visibly reflect the health state of organisms. Especially in diseases with a genetic component, examination of the phenotype can aid understanding the underlying genetics. Many technologies to generate phenotypes systematically in a high-throughput manner, such as RNA interference (RNAi) or gene knock-out, have been developed to decipher functions for genes. This ongoing large-scale characterization of genes in model systems will increase phenotypic information exponentially in the near future.

It is still a major challenge to interpret the results of large-scale functional screens, even more so if heterogeneous data sets are to be combined. Furthermore, there have been relatively few efforts to make use of phenotype data beyond the single genotype-phenotype relationship. In this thesis, methods are presented for knowledge discovery in phenotypes across species and screening technologies.

A thorough survey is conducted of the available phenotype resources and various approaches to analyzing their content are reviewed, including a discussion of hurdles yet to be overcome, e.g. lack of data integration, inadequate phenotype ontologies and shortage of appropriate analytical tools.

PhenomicDB version 2, a multi-species genotype/phenotype database, is an approach to integrate and show genotype and phenotype data on a large scale, using orthologies to show phenotypes across species. Here, the focus lies on the incorporation of quantitative and descriptive RNAi screening data and ontologies of phenotypes, assays and cell-lines.

Furthermore, as the heart of this thesis, the results of a study are presented in which the large set of phenotype data from PhenomicDB is taken to predict gene annotations. Here, text clustering is utilized to group genes based on their phenotype descriptions. It is shown that these clusters correlate well with several indicators for biological coherence in gene groups, such as functional annotations from the Gene Ontology (GO) and protein-protein interactions. The clusters are then used to predict gene function by carrying over annotations from well-annotated genes to less well-characterized genes in the same cluster.

Finally, the prototype PhenoMIX is presented, showing the integration of genotype and phenotype data with clustered phenotypes, orthologies, interaction data and other similarity measures. These data, grouped by their similarity measures are evaluated for predictiveness in gene functions and phenotype terms.

Zusammenfassung

Häufig spiegeln Phänotypen die Gesundheit von Organismen sichtbar wider. Die Untersuchung des Phänotyps bringt daher insbesondere bei genetischen Krankheiten ein Verständnis der zugrunde liegenden genetischen Mechanismen mit sich. Aufgrund dessen wurden neue Technologien entwickelt, so zum Beispiel RNA-Interferenz (RNA interference – RNAi) oder Gen-knock-out Verfahren, um unbekannte Genfunktionen zu entschlüsseln. Diese Experimente führen zu einem starken Anstieg der phänotypischen Daten.

Es bleibt eine große Herausforderung, Ergebnisse von großen Versuchen zu interpretieren, insbesondere bei heterogenen Daten. Nur wenige Ansätze haben bisher solche Daten über die einzelne Verknüpfung von Genotyp und Phänotyp hinaus interpretiert. In dieser Dissertation werden neue Methoden gezeigt, um Entdeckungen in Phänotypen über die Grenzen von Spezies und Methodik hinweg zu ermöglichen.

Es erfolgt eine gründliche Erfassung der verfügbaren Phänotypen-Ressourcen und einiger Ansätze zur Analyse ihres Inhalts. Die Grenzen und Hürden, die noch bewältigt werden müssen, beispielsweise fehlende Datenintegration, lückenhafte speziesübergreifende Ontologien und der Mangel an angemessenen Methoden zur Datenanalyse, werden diskutiert.

Der Ansatz zur Integration von Genotyp- und Phänotypdaten in großem Maßstab, PhenomicDB Version 2, wird präsentiert. Diese Datenbank assoziiert Gene mit Phänotypen mittels Orthologie über Spezies hinweg. Im Fokus liegen die Integration von RNAi-Daten und die Einbindung von Ontologien für Phänotypen, Experimentiermethoden und Zelllinien.

Ferner wird als Herzstück dieser Arbeit eine Studie präsentiert, in der die Phänotypendaten aus PhenomicDB genutzt werden, um Genfunktionen vorherzusagen. Dazu werden Gene aufgrund ihrer Phänotypen mittels Textclustering gruppiert. Diese Gruppen zeigen hohe biologische Kohärenz, da sich viele gemeinsame funktionale Annotationen aus der Gen-Ontologie (Gene Ontology – GO) und viele Protein-Protein-Interaktionen (PPi) innerhalb der Gruppen finden, was zur Vorhersage von Genfunktionen durch Übertragung von Annotationen von gut annotierten Genen zu Genen mit weniger Annotationen genutzt wird.

Zuletzt wird der Prototyp PhenoMIX präsentiert, in dem Genotypen und Phänotypen mit geclusterten Phänotypen, PPi, Orthologien und weiteren Ähnlichkeitsmaßen integriert wurden. Diese Daten werden aufgrund ihrer Ähnlichkeitsmaße gruppiert und zur Vorhersage von Genfunktionen, sowie von phänotypischen Wörtern genutzt.

Content

Dedication and Acknowledgements

I acknowledge and highly appreciate the scientific support of my doctoral adviser, Prof Ulf Leser at the chair for Knowledge Management in Bioinformatics of the Humboldt University of Berlin. He brought into this thesis the informatics portion and gave me an office and intellectual nourishment for many years.

I am grateful to Dr Bertram Weiss and the Global Drug Discovery Unit of Bayer Schering Pharma AG who supported this work in too many ways to list here. Especially, I thank Dr Weiss for his mentorship and for his support in the past years. I am indebted for the advice he gave me.

I thank Prof Herzel of the Institute for Theoretical Biology of the Humboldt University of Berlin for agreeing to be the third reviewer of this thesis.

The metalife AG is collaborating with the Bayer Schering Pharma AG on the production and maintenance of PhenomicDB. I am thankful to the PhenomicDB staff of the metalife AG for their valuable support in maintaining this resource.

I also acknowledge the work of Abdullah Kahraman, the first student who worked on MSP and PhenomicDB and whose valuable work brought forth this wonderful resource.

I acknowledge the work of Holger Pirk who coded programs that helped in the visualization of phenotype and PPi networks and the visualization of PhenoMIX and its translation into ORACLE.

I thank Mr Richard Homes for his valuable comments on the English language in this thesis.

I admire the work of Charles Darwin who enabled so many scientific achievements including this doctoral thesis and whose 200th birthday was celebrated this year on February 12th.

Last but not least I thank my wife Katrin. She supported me all the way and I am especially grateful that she actually became my wife even though I had a full time job and an unfinished doctoral thesis. I hope she still remembers how I look like.

Listing of Abbreviations

Abbreviation	Meaning
AG	Aktiengesellschaft (= Inc.)
ANN	Artificial Neural Networks
API	Application Programming Interface
apoE	apolipoprotein E
ASCII	American Standard Code for Information Interchange
ATCC	American Type Culture Collection
BLAST	Basic Local Alignment Search Tool
CLUTO	Clustering Toolkit
dbGaP	Database of Genotypes and Phenotypes
DictyB	DictyBase
DILS	Data Integration in the Life Sciences
DNA	De(s)oxyribonucleic Acid
DRSC	Drosophila RNAi Screening Center
dsRNA	Double-stranded RNA
EMS	Ethylmethanesulphonate
ENU	N-Ethyl-N-Nitrosourea
EUROFAN	European Functional Analysis Network
FN	False Negative
FP	False Positive
GO	Gene Ontology
GT	Genotype
HGVS	Human Gene Variation Society

Abbreviation	Meaning
HP	Human Phenotype ontology
ID	Identifier
Inc.	Incorporated (= AG)
ISMB	Intelligent Systems on Molecular Biology
KB	Kilo bytes (= 1,024 bytes)
kNN	k-Nearest Neighbor
KO	Knock-out
MAGIC	Multisource Association of Genes by Integration of Clusters
MeSH	Medical Subject Headings
MIPS	Munich Information Center for Protein Sequences
MP	Mammalian Phenotype ontology
MSP	Multi-Species Genotype/Phenotype Database
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
nt	Nucleotides
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
PHA	Phenylalanine hydroxylase
PHAdb	Phenylalanine hydroxylase database
PKU	Phenylketonuria
PMC	PubMed Central
PPi	Protein-protein interaction
PT	Phenotype

Abbreviation	Meaning
RGD	Rat Genome Database
RNA	Ribonucleic acid
RNAi	RNA interference
sd	Standard Deviation
sim	Similarity
siRNA	Small Interfering RNA
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
SVM	Support Vector Machines
TF	Term Frequency
TFIDF	Term Frequency – Inverse Document Frequency
TN	True Negative
TP	True Positive
TRIPLES	Database of Transposon-Insertion Phenotypes, Localization and Expression in *Saccharomyces cerevisiae*
UMLS	Unified Medical Language System
WB	WormBase

Preface

This thesis deals with the knowledge management about phenotypes in the context of biomedical sciences. It has been supported by a stipend from the Department for Global Drug Discovery of the BayerSchering Pharma AG and by the chair for Knowledge Management in Bioinformatics at the Humboldt University of Berlin.

In accordance with Section 5, Article 2b) of the doctorate regulations of the faculty of natural sciences at the Humboldt University of Berlin, parts of this thesis have been published as follows:

- Groth, P.; Weiss, B.; Pohlenz, H. D. and Leser, U. (2008): Mining phenotypes for gene function prediction, BMC Bioinformatics (vol. 9), p. 136.
- Groth, P.; Pavlova, N.; Kalev, I.; Tonov, S.; Georgiev, G.; Pohlenz, H. D. and Weiss, B. (2007): PhenomicDB: a new cross-species genotype/phenotype resource, Nucleic Acids Res (vol. 35), No. Database issue, pp. D696-9.
- Groth, P. and Weiss, B. (2006): Eine speziesübergreifende Ressource für RNAi-Daten und klassische Phänotypen, Laborwelt (vol. 6), No. 7, pp. 14-16.
- Groth, P. and Weiss, B. (2006): Phenotype Data: A Neglected Resource in Biomedical Research?, Current Bioinformatics (vol. 1), No. 3, pp. 347-358.

Further parts have been published in the context of the following talks, posters and interviews:

- P. Groth and B. Weiss, Interview on Phenotypes with BioInform Journalist Vivien Marx: Marx, V.; "Phenotype Lookup and Linkup: New Methods Arise to Mine Phenotypes for Gene Function", BioInform, Vol. 12, No. 35, 2008.
- P. Groth, U. Leser, H. D. Pohlenz and B. Weiss; „*Mining Phenotypes for Protein Function Prediction*", Invited Talk and Poster, Special Interest Group meeting on Automated Function Prediction, Intelligent Systems on Molecular Biology, Toronto, Canada, July 2008.
- P. Groth and B. Weiss; "Using Phenotypes for Gene Function Prediction", Poster, Human Genome Variation Society Scientific Meeting, Barcelona, Spain, May 2008.

- P. Groth and B. Weiss; "*Relationship between mouse phenotype and human disease from the perspective of PhenomicDB*", Invited Talk, CASIMIR First Networking Meeting, Corfu, Greece, October 2007.
- P. Groth, B. Weiss, U. Leser; "*Making use of Phenotype Data - an Ongoing Challenge*", Poster, Intelligent Systems on Molecular Biology, Vienna, Austria, July 2007.
- P. Groth, N. Pavlova, I. Kalev, S. Tonov, G. Georgiev, H. D. Pohlenz, B. Weiss; "*A Multi-Species Genotype/Phenotype Database for Comparative Phenomics*", Poster, Data Integration in The Life Sciences, Cambridge United Kingdom, July 2006.
- P. Groth and B. Weiss; "*Phenotypic data: An important resource in biomedical research*", Invited Talk and Poster, Special Interest Group meeting on Bioinformatics and Disease, Intelligent Systems on Molecular Biology, Detroit, USA, July 2005.

Exploiting phenotype information in a systematic manner and across studies and species is an important concept in the fields of comparative phenomics and drug discovery. Its importance has been widely recognized and studied from various aspects. This thesis adds value to these fields by reviewing and extending data integration and text-mining methods, making use of cross-species genotype-phenotype data.

1 Introduction

1.1 Motivation

Phenotypes are traceable changes or variations in behavior or appearance, differentiating one individual of a species from another on all but the genetic levels. They are thus a highly valuable information resource at the interface of medicine and biology. They can be used to dissect the relationships between genetic diseases and their responsible genes. However, they are not limited to superficial physical observations, e.g. the skin, of an organism and are usually the result of a long-term interaction between genes and environment, so they are a highly complex concept.

Phenotypes have been a subject of research ever since ancient Greek and Roman physicians such as Hippocrates (460-370 BC), Celsus (25 BC – 50 AD), and Galen (130 - 201 AD) took an interest in meticulously describing and studying the human body and associated illnesses with physical causes [Delvey and Barbara, 2005]. It was in the 19th century, however, when scientists started to systematically examine phenotypes for their origins. By minutely describing the differences and common traits of different species, Charles Darwin (1809 - 1882) postulated his theory of evolutionary selection which states that variation within species occurs randomly and that the survival or extinction of each organism is determined by that organism's ability to adapt to its environment [Bowler, 1996]. Although his renowned book was entitled ‘The Origin of Species’, he never explained the actual origin of species or how heritable changes were passed on to subsequent generations, a topic heavily disputed at the time. By examining successive generations of peas, Gregor Mendel (1822 - 1884) observed that specific ‘traits’ of peas were passed on from one generation to the next and recurred in certain numerical ratios. To explain his results, he distinguished between the internal state (‘genotype’) and the external appearance (‘phenotype’). He came up with the ideas of dominance and segregation, postulating that offspring receive different sets of discrete ‘hereditary factors’ (‘genes’) from parents, and thus founded modern genetics. Although published in 1866 already, Mendel’s findings were only really understood in the early 20th century, when they were independently rediscovered by Hugo de Vries, Carl Correns and Erich von Tschermak [DeVries, 1900; Klare, 1997; Rheinberger, 1995; Tschermak, 1900]. It was the Danish botanist Wilhelm Johannsen who in 1909 coined the terms genotype and phenotype (derived from Greek *genein* meaning ’to give birth’ and

phanein 'to show' respectively) and later, in 1911, introduced the distinction between genotype as a descriptor of the genome (describing the process of inheritance) and phenotype as a descriptor of the phenome (describing the process of development) [Johannsen, 1911]. Thus, in the most general sense, a phenotype is the expression of an organism's individual genetic blueprint under varying environmental influences. In the same sense, a genotype comprises that organism's entire genetic blueprint.

It is now well accepted that there is a close relationship between genotype and phenotype. We are accustomed to defining diseases by a sum of symptoms and try to trace changes in phenotype back to their genetic origin or to environmental properties. It is common practice that researchers define a phenotype in terms of the very small set of phenotypic characteristics that differ at the clinical, cellular, or molecular level from a fictitious average within a species. Our limited ability to fully describe all similarities and differences between several individual's entire genomes and phenomes leaves us no choice but to learn to live with partial data in this area. This is also why the term genotype is often used incorrectly for a single genetic change at a certain site in the genome (as if compared to a 'reference genome') and phenotype as a synonym for a certain phenotypic characteristic that is different as if compared with a hypothetical average individual of that species. Researchers typically do not report all other observable changes obviously unrelated to their specific interest. Another limitation is that for most phenotype data collected so far, the environmental contributions are either neglected or, in the case of model organisms, kept to a homogeneous minimum using standardized laboratory conditions. Still, enormous efforts have been undertaken to decipher genes and their functions, and to find the phenotypes corresponding to them. Most are motivated by the fact that an understanding of genotypes, phenotypes and their relationships will lead to new cures for diseases and give further insights into the connection between the molecular and systemic functions of organisms.

Driven by this motivation, new methods have been developed in order to find more functional relationships between genotypes and phenotypes in less time. These efforts have culminated in the development of high-throughput phenotype screening methods such as RNA interference (RNAi) [Tuschl and Borkhardt, 2002] and in combination with public databases [Gunsalus, et al., 2004; Sonnichsen, et al., 2005], phenotypes have become an acknowledged and widely used component of functional genomics. Mostly, these data are interpreted for a gene-by-gene functional annotation, using each single genotype-phenotype relationship, since it is the most immediate result of such a screen. This simple

biological evaluation can already uncover the involvement of genes in diseases and may lead to novel therapeutic approaches. However, such genotype-phenotype relationships, especially generated in large amounts as is currently the case, may yield more than just the information about a single gene. As is the case in other large-scale whole-genome approaches (such as microarray analyses), data mining can be applied in order to analyze groups of phenotypes, and thus groups of genes. Thus, information can be gained beyond the single genotype-phenotype relationship. These comprehensive studies of phenotype data aim at helping to understand the genotype-phenotype relationship across datasets, methods and even species, and form the field of research termed comparative phenomics.

Phenotype studies in biomedical research have driven the creation of public phenotype data repositories for many different kinds of data types, species and purposes (for more details, see section 4.2.4 and the survey by Groth and Weiss [Groth and Weiss, 2006a]), and have in consequence helped to increase the amount of available phenotype data in the recent years. Almost all phenotype data are stored in textual form. Entries range from descriptions of the outcome of RNAi knock-down experiments in model organisms such as *Caenorhabditis elegans* with single terms from a controlled vocabulary, like 'lethal' or 'reduced egg size' in WormBase [Rogers, et al., 2008], knock-out studies in *Mus musculus* described with free-text supplemented with terms from the Mammalian Phenotype ontology (MP) [Smith, et al., 2005] in the Mouse Genome Database (MGD) [Bult, et al., 2008], all the way to clinicians' free-text descriptions of genetic diseases in *Homo sapiens*, such as Diabetes mellitus or Alzheimer's disease in the Online Mendelian Inheritance in Man (OMIM) [McKusick, 2007].

Methods for data integration, data mining and discovery are being developed to make this wealth of information readily accessible and comparable and to provide the means for new discoveries. For example, the outcomes of RNAi experiments supplemented with a controlled vocabulary mentioned above have been be used to create a 'disease map', a graphical display of 45 disease categories [Sonnichsen, et al., 2005], which in turn has led to the development of 'PhenoBLAST', a tool to compare knock-down phenotypes by absence or presence in any of these 45 disease categories [Gunsalus, et al., 2004]. These approaches have culminated in the integration of data from protein interactions, gene expression clusters, and phenotypic RNAi profile similarities into one large gene/protein network with 'a high predictive value for novel gene functions' [Gunsalus, et al., 2005] (for more details, see section 4.2.4 and the survey by Groth and Weiss [Groth and Weiss, 2006a]).

In light of the advances in utilizing genotypes and phenotypes in the functional discovery process, and given the possible benefits in biomedical science, it is feasible to explore further the knowledge discovery on genotype/phenotype data. The present thesis deals with this. In the following, I shall explore how to efficiently gather and integrate large amounts of genotype-phenotype data across species in the database PhenomicDB [Groth, et al., 2007; Kahraman, et al., 2005], use text clustering on the textual phenotype descriptions in order to predict gene functions [Groth, et al., 2008] and develop an integrative system for large-scale genotype-phenotype data mining and prediction, named PhenoMIX.

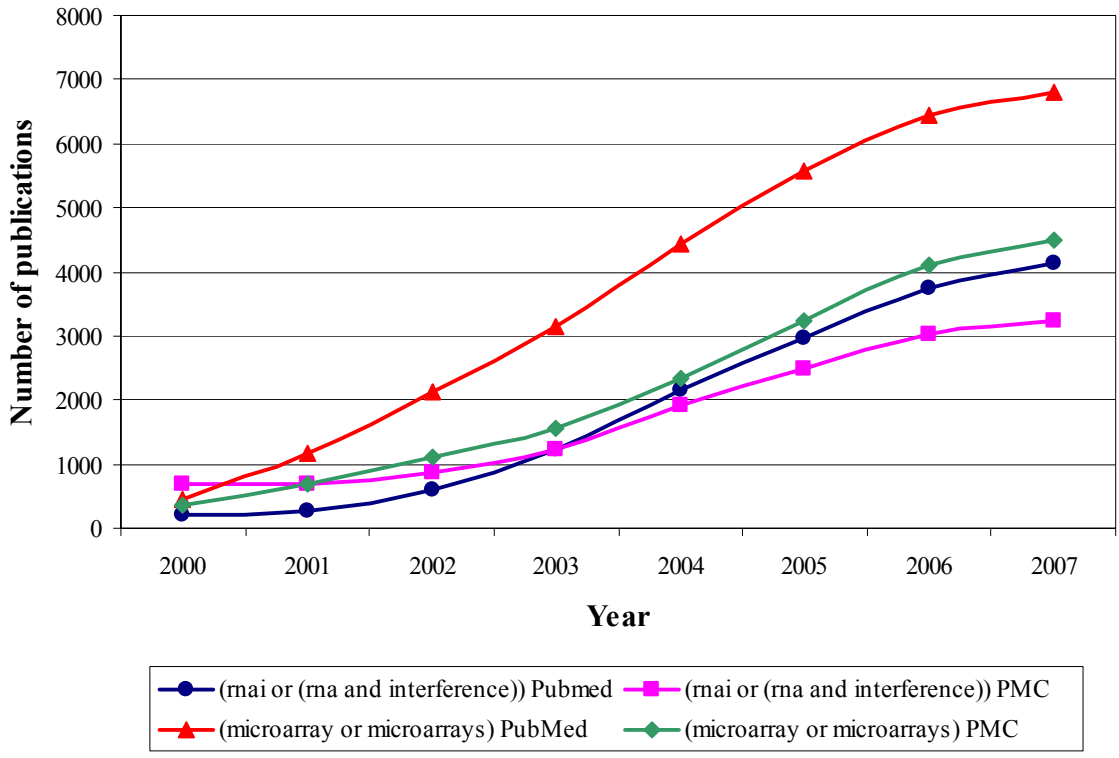


Figure 1:

The number of new publications each year in PubMed and PubMed Central (PMC) that can be found with the search terms '(rnai or (rna and interference))' and ' (microarray or microarrays)'.

In short, comparative phenomics approaches can help to identify gene functions and select candidate genes for widespread genetic disorders, like diabetes. The public awareness of this topic can also be traced with Google™ Trends (http://www.google.com/trends). Google™ Trends provides a basic graphical indication for the number of searches in

Google™ with a specific keyword over time. Here, the number of searches with the term 'microarray' (another popular method for high-throughput gene function analysis) has been decreasing steadily for the past three years, while the number of searches with 'rnai' is approaching that number. In 2006, the number of searches with 'rnai' is actually higher than those for 'microarray' for a short period of time, corresponding to the announcement of the 2006 Nobel Prize for medicine to be awarded to Craig Mello of the University of Massachusetts Medical School in Worcester, MA, and Andrew Fire of Stanford University for their discovery that strands of RNA can selectively silence genes [Couzin, 2006]. This is a clear indication that phenotype experimentation is an advancing hot topic. Looking up the number of new publications each year since 2000 in PubMed and PubMed Central (PMC) with the search terms '(rnai or (rna and interference))' and '(microarray or microarrays)' shows that the number of new publications about RNA interference each year is still chasing the number of new publications on microarrays (see Figure 1). Thus, there is a keen scientific interest on phenotype information for the use of biomedical knowledge.

1.2 Genotypes

1.2.1 Definitions and concepts

For practical use, both genotypes and phenotypes are vague concepts. In order to be able to work with them it is necessary to narrow down both definitions to more practical concepts, such that they can be pinpointed to actual traceable and comparable entities. Only then can they be used to represent a 'genotype' or a 'phenotype' on a practical level, e.g. in a database or in a computational analysis.

Figure 3 shows different concepts for a genotype at different levels of resolution of the genetic composition. Any of these can be considered a 'genotype' from a certain point of view. From this, it becomes clear that a genotype is regarded as anything ranging from the sum of all genetic properties of an organism (which is closest to its actual definition, see section 1.1) to a single variant in a single nucleotide in the genetic code of that same organism. Despite its clear definition, this heterogeneity of concepts for a genotype needs to be overcome for practical use.

Another limiting factor for using the 'genotype' even in its strictest definition in practice is that it is virtually impossible to determine the entire genetic composition of an individual organism. In fact, each individual's genotype differs from all others, even within a species. It is thus necessary to work with more practical structures than a 'genotype'.

One idea is to use the concept of genes rather than that of a genotype in order to have a more clear-cut definition. However, in a recent publication, Gerstein et al. give eight different definitions of a gene, showing that the perception of the concept has changed dramatically over time [Gerstein, et al., 2007]. Derived from Gregor Mendel's 'hereditary factors', the concept for a gene was specified by Thomas Morgan as an 'abstract entity whose existence is reflected in the way phenotypes were transmitted between generations' [Gerstein, et al., 2007; Morgan, et al., 1915]. After Watson and Crick had solved the three-dimensional structure of DNA in 1953 [Watson and Crick, 1953], a molecular view of the gene developed as 'a code residing on nucleic acid that gives rise to a functional product' [Gerstein, et al., 2007]. The genetic code, using letters 'A', 'C', 'G' and 'T' in sequencing and cloning technologies, along with algorithms developed to discover functional sequences in genomes meant that a DNA sequence could be used to infer structure and function for the gene and its products. Thus, in the 1970s and 1980s, a gene was viewed as some part of a (predicted) sequence rather than as a genetic locus responsible for a phenotype [Gerstein, et al., 2007; Griffiths and Stotz, 2006]. The sequence view is still widely accepted and the Human Genome Nomenclature Organization defines a gene as 'a DNA segment that contributes to phenotype/function. In the absence of demonstrated function, a gene may be characterized by sequence, transcription or homology' [Wain, et al., 2002].

Clearly, the concept of a gene is under constant development. For practical use, other definitions are necessary for gene or genotype. In this thesis, the term genotype is used synonymously with the term gene and is regarded solely as an entry in the NCBI's Entrez gene database (see Figure 2) [Maglott, et al., 2007] with its unique identifier. The composition of such an entry is explained in section 1.2.3 and the gene data used here is shown in section 2.1.1.

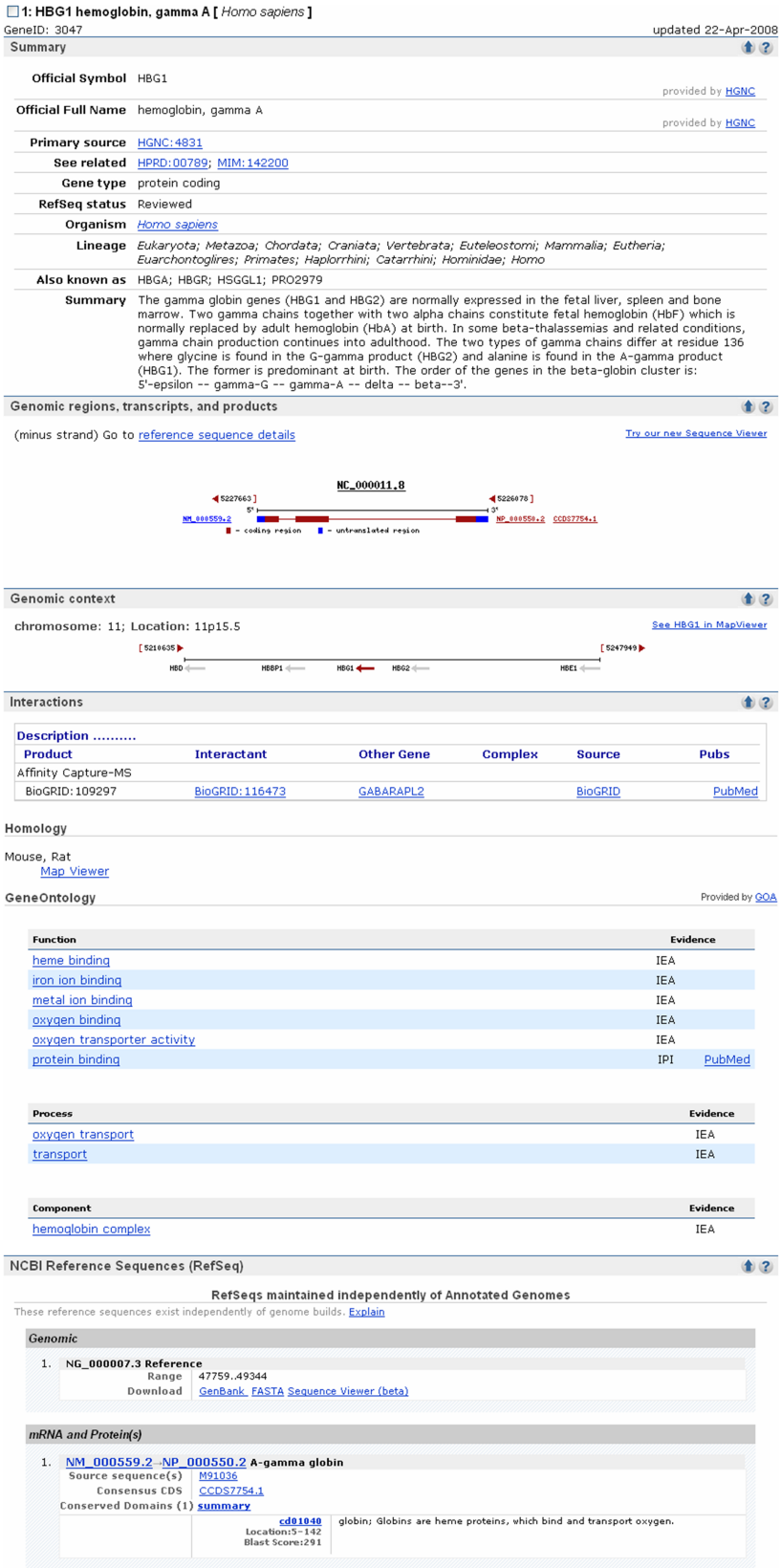


Figure 2:

NCBI Entrez gene entry (abridged) on the gene coding for the human *HBG1* protein, depicting information on location, function, sequence, interactions and orthologies (from: [Maglott, et al., 2007] / adapted: P. Groth, 2008).

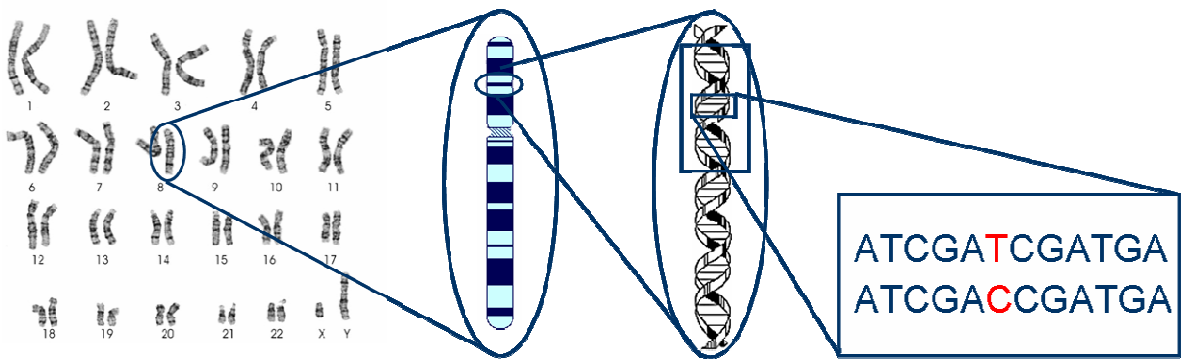


Figure 3:

Different concepts for a genotype at different levels of granularity of the genetic composition of an individual from A to E. Even a single change in one of the nucleotides represented in E gives a different genotype, termed single-nucleotide polymorphism (SNP).

A: Shows a karyogram, the entire genetic composition of the diploid chromosome set of a female human individual (from: [Manske, 1995]);

B: Depicts a schematic view on the human chromosome 8 (from: [Rankinen, et al., 2006]);

C: Schematic view of a part of the double-helix DNA strand within this chromosome (from: [Felsenfeld and Groudine, 2003]);

D: Depicts the possible borders of a gene on this DNA strand;

E: Shows the nucleotide sequence of a part of this gene. Each letter of that sequence represents one of the four nucleotides Adenine, Cytosine, Guanine or Thymine.

1.2.2 From genes to proteins

By a process called ‘protein biosynthesis’, proteins are synthesized from genes by transcription of DNA sequence to mRNA and translation into their amino acid sequence. It is by this mechanism that genes encode for proteins. Furthermore, many genes encode for more than one protein by effects occurring during the protein biosynthesis, i.e. post-transcriptional (e.g. ‘alternative splicing’) and post-translational modifications (e.g. ‘phosphorylation’) [Campbell and Reece, 2008].

It is beyond the scope here to discuss this ‘central dogma of molecular biology’ as it was postulated by Crick [Crick, 1970], the criticism of its application [Werner, 2005] or the importance of RNA editing [Maydanovych and Beal, 2006] and small non-coding RNA in biological processes [Mattick, 2003]. It is sufficient to know that proteins are essential for every organism and are participating in many processes in the cells. They form enzymes and are thus vital parts of the metabolism by catalyzing biochemical reactions. Furthermore, proteins have structural and mechanical functions, e.g. in muscles or in the cy-

toskeleton, maintaining the shape of cells. Proteins also play significant roles in cell signaling, immune responses, cell adhesion, and in the cell cycle [Campbell and Reece, 2008].

This relationship between genes and proteins is used in sections 3.2.2.1, 3.2.2.5, and 3.3, where physically interacting proteins are treated as groups of their encoding genes.

1.2.3 Gene-centered information

The National Center for Biotechnology Information (NCBI – http://www.ncbi.nlm.nih.gov) is a suitable source for gene-centered information. Wheeler et al. provide a thorough survey of available resources at the NCBI [Wheeler, et al., 2008]. Most gene-centered information (or links thereto) can be found in the NCBI's Entrez gene database (see Figure 2 for an entry) [Maglott, et al., 2007]. Major aspects that are important to understand genes are gathered there. Gene-centered information includes, but is not limited to: Gene name and symbol, species of origin, chromosomal location and genomic context, orthologies, interactions, functional annotation, and sequence(s). In the context of this thesis, the central pieces of information for a gene that are extracted and kept for each gene are orthologies, physical interactions of gene products, functional annotations and the sequence (see section 2.1.1 for details). These four data points contain both necessary information to understand the function of a gene and possibilities to compare one gene with another.

Of course, there are other repositories for gene-centered information, e.g. Ensembl [Flicek, et al., 2008]. While the data presentation and the structure of the underlying database may vary, most of the contents are the same. It is therefore reasonable to choose one of the major gene databases with daily updates, such as Entrez gene (see also section 2.1.1 for more details). The many different gene information repositories are discussed by Furey [Furey, 2006].

1.2.4 Genotype-genotype relationships

1.2.4.1 Orthologies

Knowledge of a group of homologous genes (or homologs) can be highly useful for transferring information from one member of such a group to another (see section 2.1.2 for details on the data). Homology has been used to search for novel genes [Cui, et al., 2007], to predict protein function [Friedberg, 2006] or protein structure [Ginalski, 2006], for prediction of proteins with specific locations and properties, e.g. membrane proteins [Punta, et al., 2007] and for prediction of protein-protein interactions (PPi) [Shi, et al., 2005]. There

are two types of gene homology: orthology and paralogy (see Figure 4 for an overview). Whereas paralogs may only partially share function, orthologous genes, which are defined to be the same genes on functional level in two different species derived from a common ancestor, are of interest in disease modeling and function prediction. Rubin et al. have identified 177 genes of *Drosophila melanogaster* as ortholog to human disease genes [Rubin, et al., 2000]. This is a highly interesting finding since by definition these orthologs should have a similar function and may thus be useful for finding cures for the associated diseases. In *Caenorhabditis elegans*, 12% of the species' genes encode for proteins whose biological roles are highly similar to their putative orthologs in *Saccharomyces cerevisiae* (~27% of all yeast genes) [Chervitz, et al., 1999].

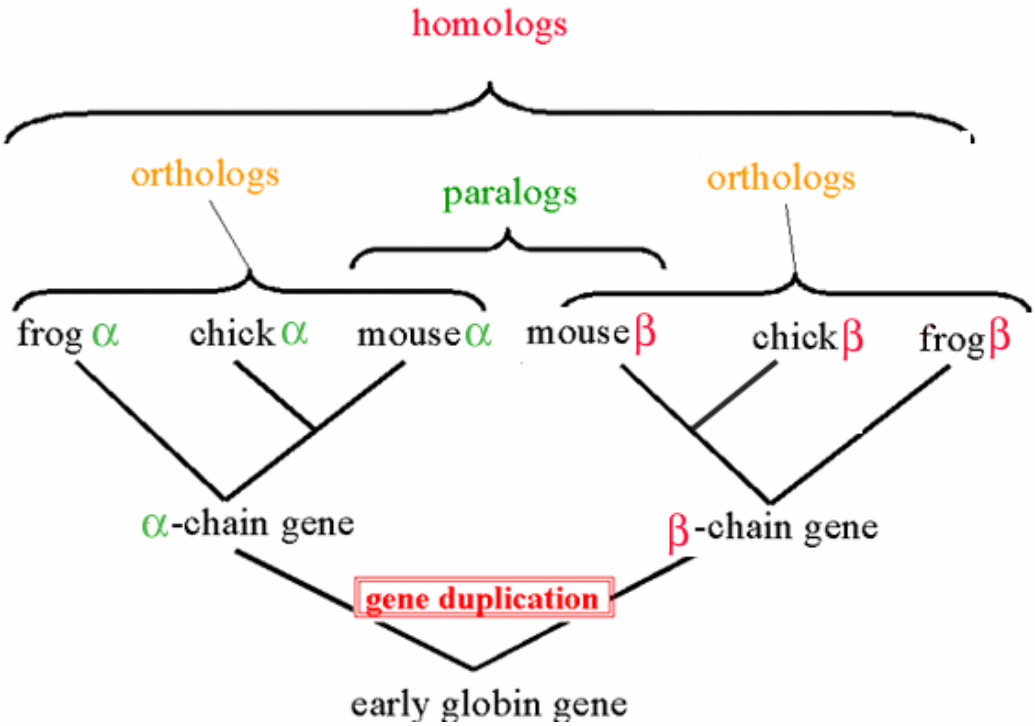


Figure 4:

Depiction of the types of gene homology: orthology and paralogy. An ancestral gene (in this example from the globin family) is separated into two types (alpha and beta) by gene duplication. Each of the resulting alpha and beta chain genes are *orthologs* in the different species. Two different genes deriving from one ancestor gene in a single species (here: alpha and beta chain genes in mouse) are *paralogs* (from: [Sezen, 2000]).

1.2.4.2 Interactions

Another very useful way to find functionally related genes is by looking at groups of interacting gene products, i.e. protein-protein interactions (PPi, see section 2.1.3 for details). Physically interacting proteins have a higher chance to be part of the same biological process or pathway than non-interacting proteins [Guo, et al., 2006]. Furthermore, such groups are essential for all cellular functions and they work in concert by physical interaction with other biomolecules [Mathivanan, et al., 2006]. Thus, groups of interacting proteins have highly interesting biological properties. Furthermore, they can be regarded as networks of genes and have been subject to intensive studies in the past (see for example: [Jaeger and Leser, 2007; Kemmer, et al., 2005; Riley, et al., 2005; Schuster-Bockler and Bateman, 2007; von Mering, et al., 2007] and section 4.2.3.2 for further details). The importance of these networks have brought forth a large number of PPi databases, many of which have been reviewed by Mathivanan, et al. [Mathivanan, et al., 2006].

1.2.5 Functional annotation: The Gene Ontology (GO)

Many eukaryotic gene products have been found to share core biological functions [Rubin, et al., 2000]. The Gene Ontology (GO) was created in 1999 with the goal to 'produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge of gene and protein roles in cells is accumulating and changing' [Ashburner, et al., 2000]. GO is structured, which means that all terms are organized in one of three sub-ontologies, i.e. molecular function, biological process and cellular component, representing a directed acyclic graph with a number of successors and predecessors for each term. In short, GO provides a structured controlled vocabulary to coherently annotate genes with functions, locations and processes. GO is widely recognized as the most comprehensive functional classification system and has become a *de facto* international standard for functional annotation and to control predictions [Daraselia, et al., 2007; Pandey, et al., 2006; Rison, et al., 2000]. As of January 2008, 199,703 unique IDs from Entrez gene are associated with 12,425 (of 25,928) unique IDs from GO. The structure and style used to represent gene functional annotations and associations to genes within GO is illustrated in Figure 5 [Ashburner, et al., 2000].

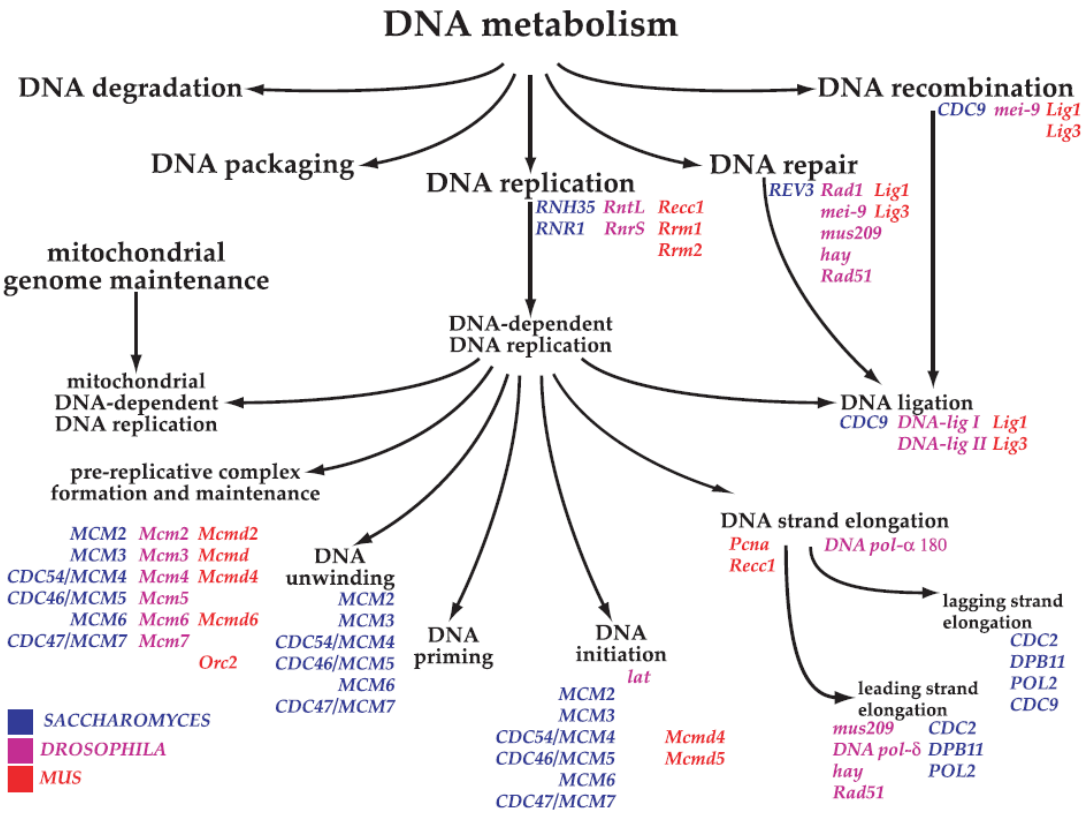


Figure 5:

Example from the Gene Ontology. This example illustrates the structure and style used to represent gene function annotations and associated genes within GO. The depicted part of GO illustrates a small portion of the biological process sub-ontology describing DNA metabolism (from: [Ashburner, et al., 2000]).

1.3 Phenotypes

1.3.1 Differences in concepts and how to overcome them

Our growing knowledge about the complexity of genetic interactions and the inter-individual genetic variance reflects our daily experience of the high degree of phenotypic individuality within the species *Homo sapiens*. Therefore, the concept of a phenotype is even more difficult to grasp than that for a genotype. In the frame of the present work, it is necessary to reduce a phenotype to reflect the following: Only a *change* in appearance, i.e. a deviation from the so-called wild-type will be considered a phenotype. Furthermore, such a change of a wild-type needs to be associated with a traceable change in a genotype, either by mutation, disease or genetic interference. Even in this narrow view of a phenotype, it could still comprise any observable characteristic of an organism, the description of a dis-

ease, the characterization of natural mutations, results of a gene knock-out or knock-down experiment or an artificially induced mutation, etc. (refer to the review by Groth and Weiss [Groth and Weiss, 2006a] for more examples on the definition of a phenotype). Unfortunately, there is no common vocabulary to describe these observations. Instead, researchers use either home-grown or domain-specific vocabularies or plain English text. Due to the resulting heterogeneity in descriptions, and in order to maintain the ability to use all the available phenotype data, this thesis uses the broadest common denominator for a phenotype, i.e. its textual description. Thus, phenotype descriptions used here comprise all of the concepts named above. In summary, the scope of what can be considered 'phenotype data' is poorly defined (if at all) and can range from data at the molecular level to clinical patient records at organism or even population level.

Currently, the phenotype community places particular emphasis on phenotype data from RNA interference screens (RNAi). The concept of introducing one RNA sequence into a cell to knock-down one single gene and thus the abundance of its protein derivative at a time is groundbreaking [Tuschl and Borkhardt, 2002]. Its high specificity (a single base mismatch prevents the silencing effect [Brummelkamp, et al., 2002]) makes its application cost-effective and thus gives ground to the possibility of knocking down the expression of mutated alleles e.g. in cancer and neurodegenerative diseases [Shi, 2003]. The development of RNAi technologies for mammals, where now the immune response against long dsRNA is avoided by using short 21-23nt long siRNAs [Shi, 2003], has enabled generating large amounts of mammalian in-vitro data [Downward, 2004; Kolfschoten, et al., 2005; Stephens, et al., 2005; Westbrook, et al., 2005]. As RNAi is applicable at the cellular level, it overcomes the limiting generation time of genetically modified higher mammals (e.g. knock-out mice) and has proven to be useful in filling gaps in our understanding of genotype-phenotype relationships [Shi, 2003; Tuschl, 2003; Tuschl and Borkhardt, 2002]. The 'range of biological read-outs that can be used to infer function' [Wheeler, et al., 2005] is actually not limited and is one of the most important aspects of large RNAi screens. Due to the availability of whole genome sequences for many model organisms as well as for humans, the number of projects relating phenotypes with genotypes using RNAi is rising steadily (see the review by Friedman and Perrimon [Friedman and Perrimon, 2004]). From the number of RNAi-based phenotypes elucidated in the last few years, a large number of RNAi-based phenotype data can be extrapolated: for instance, 25 RNAi phenotypic assays

in a genome-wide screen (~ 20,000 genes) in 8 of the most important model organisms would lead to the accumulation of 4 million cellular phenotypes.

The recently started mutagenesis programs using chemical agents such as ethylmethanesulphonate (EMS) or N-ethyl-N-nitrosourea (ENU) add to this wealth of phenotype data. Both EMS and ENU induce point mutations in DNA, leading to a variety of genetic lesions that are expressed as complete loss of function, partial loss of function, or gain of function alleles [O'Brien and Frankel, 2004]. Three NIH mutagenesis centers are equipped to screen over 13,000 mice each year and have generated ~500 mouse strains of interest for neuroscience mutagenesis alone [Frankel, et al., 2008]. The NCBI and the National Human Genome Research Institute have announced they will commit US$ 50 million to collect and analyze all genetic mutations found in human cancers.

These low-level cellular 'phenotypes' are blessing and curse in our understanding of phenotype data. Their abundance in almost all sequenced model organisms, each adding their part to a mosaic, will eventually give a phenotypic picture in unparalleled high resolution. On the other hand, they are very simple, often quantitative data points in such vast amounts but low granularity that makes a comparison to 'classical' phenotypes e.g. from knock-out mice or human genetic diseases notoriously difficult. Yet, they are too valuable to be ignored. The co-existence of qualitative (i.e. 'classical' phenotype data) and quantitative phenotype data leads to a heterogeneity that can explain why there have been few attempts to integrate phenotype data, leaving us with a large number of data resources. Given the plethora of species-specific and locus-specific databases with phenotype content, genome-wide and cross-species databases are the next logical step towards comparative phenomics. Large-scale projects like the Mouse Phenome Project [Bogue, 2003; Bogue and Grubb, 2004] and Eumorphia [Brown, et al., 2005] have started to coordinate world-wide efforts for the generation of standardized phenotype data published in purpose-orientated databases. Recently, the European Union granted € 12 million in order to support GEN2PHEN, a collaborative Genotype-Phenotype effort involving 19 institutions across Europe [Brookes, 2008]. The call for an analogous project for human phenotypes, namely the Human Phenome Project [Freimer and Sabatti, 2003] was made in 2003.

Three stages have been identified to enable general data handling independent of data sources [Groth and Weiss, 2006a]. In the first stage, data sources need to be gathered (for optimal coverage), aligned (e.g. to remove redundancy), mapped onto a gene index and ultimately integrated at a semantic level such that equivalent data is eventually found in the

same place. The next stage aims at making the data comparable by introducing ontologies and other data structures that relate the data points (e.g. orthologies). Only then, in the final stage, can robust data mining methods be employed systematically to exploit the data based on statistical analyses and/or direct or cross-species functional comparison. In this thesis, all three stages are implemented in order to overcome the hurdles that have been raised by the current phenotype data diversity.

1.3.2 The Mammalian Phenotype ontology (MP)

The most consistent way to structure knowledge domains based on large heterogeneous data is the creation of a useful ontology. This has been proven true for genotypes by the creation of GO and it is also true for phenotypes. The strength of ontologies and controlled vocabularies is their use beyond one system, especially when instances or patterns reoccur. The Mammalian Phenotype ontology (MP) [Smith, et al., 2005], containing in September 2008 6,182 phenotype terms [MGI, 2008], and the very recently published Human Phenotype ontology (HP) [Robinson, et al., 2008] (~8,000 terms in April 2009) are the major phenotype ontologies available for mammals today (see Figure 6 for an example of phenotype annotations from MP). Even though HP was developed for humans (and is mostly focused on anatomic abnormalities) and MP was mainly developed for rats and mice, their potentials go beyond these species and even beyond mammals. MP, for example, contains cellular phenotypes that can partially be applied to other model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans*, opening up new perspectives for the use of this resource. Such efforts, however, could certainly not be maintained by the main HP and MP contributors alone. Thus, there have been calls for a larger community effort to drive forward more general phenotype ontologies [Brookes, 2008; Groth and Weiss, 2006a; Hancock, et al., 2007], the creation of which using automated methods has been studied by Boehm [Boehm, 2008] (see sections 4.3 and 4.4 for a discussion) and could be supported by phenotype annotation software like Phenote (available from http://www.phenote.org).

Why are ontologies important for comparative phenomics? The free-text description used by researchers to document their phenotypic observations may differ considerably and allow comparison only manually or by text-mining approaches as shown in this thesis. Phenotypic results from one screen need to be abstracted to a level where comparisons with results from other screens are enabled. An RNAi readout for apoptosis can be achieved with a TUNEL, a PARP cleavage, or an anoikis assay (see for example the review by Karaflou

et al. [Karaflou, et al., 2008]). Although the absolute values of each result type are not directly comparable, they can at least qualitatively be compared if abstracted to an ontological instance like ‘decreased apoptosis (MP:0006043)’. Ontologies also allow for unambiguous identification of biological objects when synonyms are used as query terms. For these reasons, they are highly useful for cross-species comparisons.

Figure 6:

Abridged excerpt from the Mammalian Phenotype ontology (MP) showing most of its level 1 classes and for ‘pigmentation phenotype’ also its subclasses. The terminology that is used to describe a phenotype strongly depends on the species and on the community of researchers studying it.

1.4 Relationships between genotypes and phenotypes

1.4.1 Mendelian phenotypes

So-called ‘Mendelian’ or ‘monogenic’ diseases are traceably inherited and thus, positional cloning techniques have led to the identification of roughly 1,200 disease-causing genes in humans [Botstein and Risch, 2003]. Unfortunately, their genotype-phenotype relationships are not always obvious. Different single nucleotide polymorphisms (SNPs) of the same genetic locus have been shown to cause different phenotypes (e.g. varying mutations in the *Drosophila's* shaker gene lead to a reduced sleep rate or to shaking legs after etherization [Cirelli, et al., 2005]) or, if less pronounced, may lead to subtle sub-phenotypes of a disease (e.g. mild or severe muscular dystrophy [Botstein and Risch, 2003]). Furthermore, combinations of correlated SNPs (so-called haplotypes) may ‘fine-tune’ the final phenotypic outcome of a disease-causing SNP or explain differential disease susceptibility [Botstein and Risch, 2003; Crawford and Nickerson, 2005; Shastry, 2003]. For example, phenylketonuria (PKU) is considered a classic monogenic disease which can be caused by several different mutations in the enzyme phenylalanine hydroxylase (PHA). However, even siblings who share an identical PHA genotype show ‘widely differing phenotypes’ [DiSilvestre, et al., 1991; Scriver and Waters, 1999; Treacy, et al., 1996; Weatherall, 1999]. For PHA, there is a locus-specific mutation database (PHAdb) listing 498 mutations for this locus alone [Scriver, et al., 2003; Scriver, et al., 2000], most of which are presumed to be disease-causing [Scriver and Waters, 1999; Waters, 2003]. There are hints that the ‘remarkable phenotypic variability’ [Weatherall, 1999] within this monogenic disease may also be influenced by other genetic factors [Scriver, et al., 1994; Weatherall, 1999]. Another such example for a complex genotype/phenotype relationship of a monogenic disease is thalassaemia, reviewed by Weatherall [Weatherall, 1999]. If sets of SNPs are inherited together as haplotypes, the individual phenotypic contribution of each SNP within this haplotype is further obscured. Nevertheless, haplotype analysis is an opportunity to measure the effects en-bloc [Crawford and Nickerson, 2005] and detailed elucidation of haplotypes is currently underway within the International HapMap Project [HapMapConsortium, 2003]. In short, ‘simple Mendelian inheritance is often not so simple’ [Botstein and Risch, 2003].

1.4.2 Complex traits

In comparison with the monogenic diseases, phenotypes of complex diseases are even harder to link unequivocally to the relevant variant genomic sites, as the signal spreads over several loci. Accordingly, for many multifactorial and 'complex diseases' like diabetes, Alzheimer's disease, stroke, psychiatric disorders, or obesity, the complete picture of the genotype-phenotype relationships remains largely unsolved. Here, the contributions of a gene to the disease are usually detected through studies of larger populations and are rather termed 'association' or 'susceptibility' underlining the lack of a true understanding of the contribution. Variability in two phenotypes despite identical sequences of the phenotype-causing gene(s) are often declared to be a consequence of the 'different genetic background' hereby referring to unexplained effects from the rest of the genome (e.g. epigenetic influences). Association or susceptibility data of complex diseases are difficult and expensive to measure and not the subject here. Even in a perfect setting, two organisms with an identical genotype having experienced development in an ideal environment would still differ slightly due to the stochastic nature of the underlying processes (so-called phenotypic plasticity or polyphenism [Nijhout, 2003]). Here, however, the focus shall be more on phenotype data from genotype-phenotype relationships where the genetic component plays a proportionally much more pronounced role than e.g. the epigenetic or environmental influences on the phenotype.

1.4.3 The genotype-phenotype association

Despite the difficulties of mapping a phenotype to the underlying genotype and the challenges of describing phenotypes consistently and in a highly standardized manner, many efforts have been undertaken to collect and generate phenotype data [Goldowitz, et al., 2004; Kuttenkeuler and Boutros, 2004; Page and Grossniklaus, 2002; Peters, et al., 2003]. The fruit fly is a good example for this, where one genetic screen to identify mutations with developmental phenotypes in *Drosophila* [Nusslein-Volhard and Wieschaus, 1980] was awarded the Nobel Prize. A battery of methodologies (reviewed by Carroll et al. [Carroll, et al., 2003]) has been employed over the past few decades to collect and describe mutants in great detail. These data have helped to work out the genotype-phenotype relationship with the help of so-called forward genetics, where one starts with a mutation phenotype and works toward identifying the mutated gene [Peters, et al., 2003]. In contrast, analogous studies in higher mammals have been hampered by much longer life spans, a

lack of sophisticated methods, or for ethical reasons. Here, systematic examinations of transgenic or knock-out animals as well as comprehensive SNP analyses have been successful but limited in number, especially since there was a lack of high-throughput methods to generate large quantities of data.

This situation has changed through the advent of RNAi screening methods and other potent screening technologies (see section 1.3.1) like the recently started mutagenesis programs using chemical agents such as ethylmethanesulphonate (EMS) or N-ethyl-N-nitrosourea (ENU) for uncovering genotype-phenotype relationships in high numbers.

1.5 Managing genotype-phenotype data

Knowledge management comprises methods for the identification, extraction and storage of information and knowledge. In this thesis, data integration, clustering and prediction have been applied in order to drive forward the use of phenotypes in biomedical sciences and to make use of them beyond the single genotype-phenotype association.

In the past, prior to any analysis, the status of phenotype resources required difficult manual intervention and curation of data subsets in order to allow for a systematic analysis workflow. Thus, many analytical methods have been developed and applied to such tediously hand-curated phenotype data sets.

Data integration is an essential first step when dealing with diverse data such as cross-species phenotypes. In order to create a system such as presented here, integrating heterogeneous data from many different sources, several aspects of data integration have to be considered, e.g. whether the data should be materialized or virtual, which integration technologies to use, etc. [Busse, et al., 1999]. Usually, and as is the case here, large-scale data integration is only a semi-automated process, where the mappings of fields between databases have to be done manually. Furthermore, there are some cases where automated data acquisition is not supported by the data source, leaving ‘screen scraping’ (i.e. data download by programming a script against the interface of the data source) as the only method for data retrieval. Thus, all data used here were integrated in a materialized manner. To demonstrate diversity in phenotype data, Figure 7 shows available phenotype information for the *cdk-7* gene of *Caenorhabditis elegans* in WormBase [Rogers, et al., 2008], PhenoBank [Sonnichsen, et al., 2005], and RNAi Database [Gunsalus, et al., 2004], and the corresponding yeast and mouse orthologs.

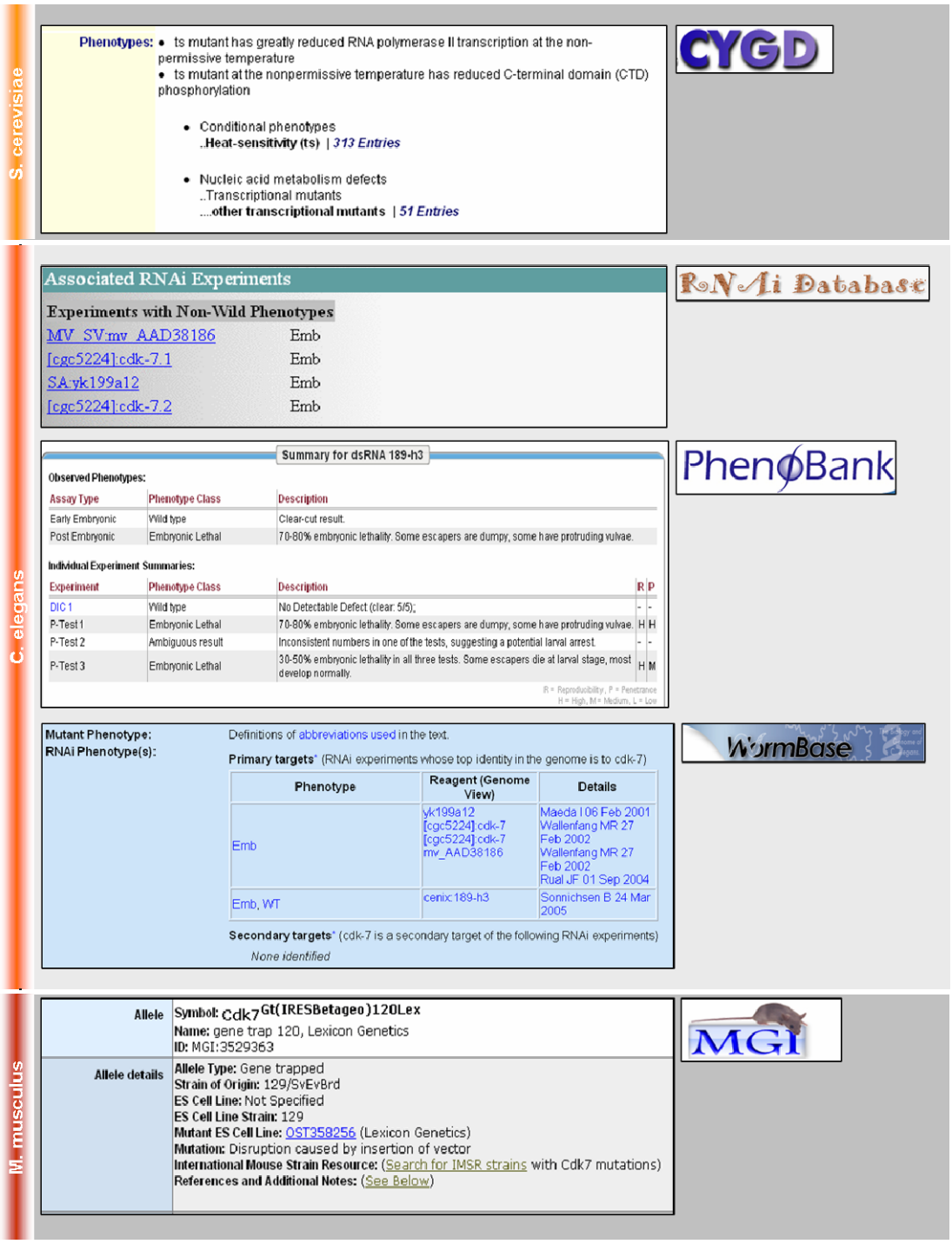


Figure 7:

Even equivalent phenotype information for *Caenorhabditis elegans* cdk-7 gene is presented in diverse ways across relevant databases, here RNAi Database [Gunsalus, et al., 2004], WormBase [Rogers, et al., 2008], and PhenoBank [Sonnichsen, et al., 2005]. Phenotype description for cdk-7 in other species (here *Saccharomyces cerevisiae*, CYGD [Guldener, et al., 2005] and *Mus musculus*, MGD [Bult, et al., 2008]) is characterized by species-specific terminology and a layout unique to each resource. The phenotype information is partially complementary, but mostly redundant. Assembling available phenotype information of the ortholog group for cross-species comparison is a time-consuming manual task.

1.6 Cross-species phenotype clustering

Any comparison or clustering of genes and phenotypes is based on similarity, e.g. sequence similarity for genes or textual similarity for phenotypes. Genes that are considered 'similar' should in some way show a consistent picture of their relatedness. For example, they should encode for proteins from the same family, or have physical protein-protein interactions (PPi). Such a group shows a 'biological coherence' as a meaningful biological statement of the relatedness of its members.

It has been shown that similar phenotypes can reflect such biological coherence. Thus, they can show functional coherence, and could even be used to identify new members of known pathways on the genetic level [Eggert, et al., 2004; Piano, et al., 2002; van Driel, et al., 2006]. However, 'comparing genotypes or even phenotypes between organisms as different as yeast and humans may involve serious scientific hurdles' [Kahraman, et al., 2005]. Thus, it is hypothesized within this thesis that clusters of phenotypes (= phenoclusters, see section 2.1.6 and Definition 5 in section 2.2.2), even across species and methods, should reflect a common biological theme. Since each phenotype used here is directly associated with a gene, this method is a novel way to group genes. In order to understand the methods for phenotype clustering (see section 2.2.2) and the results (see section 3.2.2), this section gives an overview over the possibilities for text clustering and some aspects of the evaluation of results.

Generally, there are two types of clustering algorithms: hierarchical and partitional clustering. These methods differ regarding approach, results and interpretation of results. After hierarchical clustering, the remaining question is at which level to set a cut-off to receive a sensible partitioning of the data. In a partitional clustering, this is determined by the *a priori* choice of the number of clusters with the parameter k. The success of the clustering can then only be estimated by evaluating the coherence of the resulting clusters. In the worst case, non-coherence will imply a re-computation with a different choice of k (see section 4.3.2.3 on the choice of k).

In the hierarchical clustering algorithm, the divisive implementation repeats a sequence of binary partitions on one single cluster of all elements until all these elements are left in single sets at the bottom (leafs) of the resulting tree (dendrogram) and the initial cluster containing all samples at the top (root) of the tree. The agglomerative implementation (see

Algorithm 1) starts with the single sets, merging two smaller clusters into one larger cluster until all elements are merged into a single cluster.

> ***Algorithm 1:*** *Hierarchical clustering with the agglomerative method (from: [Steinbach, et al., 2000]).*
>
> 1. *Start with a set of N clusters, each containing exactly one element.*
> 2. *Calculate the pair-wise similarity of all clusters, i.e. calculate an NxN similarity matrix M in which the ij^{th} entry reflects the similarity between the i^{th} and the j^{th} cluster.*
> 3. *Merge the two clusters with the highest pair-wise similarity.*
> 4. *Recompute M, such that it now contains the similarities between the old clusters and the new one and remove the two single clusters that have been merged.*
> 5. *Repeat steps 2 and 3 until only one cluster is left.*

In the partitional clustering method, the result consists of k unnested and distinct clusters. The most common partitional clustering method is k-means. This algorithm uses centroids, which are (in most cases) virtual points, one in the centre of each cluster. Thus, the centroid is a vector that is calculated as the mean or median of all (real) data points.

The k-means algorithm can be implemented either as basic k-means (see Algorithm 2A.) or bisecting k-means (see Algorithm 2B.). In basic k-means, the centroids will be set randomly into the feature space and then re-centered step by step according to their nearest data point until a steady state is reached. The bisecting k-means algorithm starts from one single cluster containing all samples. This cluster is partitioned into two clusters (of similar size) by a repeated search for the highest similarity. From the resulting clusters, one (the largest remaining cluster) is selected and partitioned again until there are k clusters.

The main difference between the two implementations is that the basic k-means can result in empty clusters, which is not possible with bisecting k-means. Furthermore, the bisecting k-means approach generally creates more coherent clusters in a shorter computation time and outperforms any hierarchical clustering approach in a high-dimensional feature space [Steinbach, et al., 2000]. For these reasons, the bisecting k-means is used (see section 2.2.2 for details of the software). Further discussions of the advantages and disadvantages or

suitable applications of either method can be found elsewhere (see chapter 20 of the book by MacMay [MacKay, 2003] and especially chapter 8 of the book by Backhaus et al. [Backhaus, et al., 2000]).

Algorithm 2: *Partitional clustering algorithms for* ***A.*** *basic k-means, and* ***B.*** *bi-secting k-means (from: [Steinbach, et al., 2000]).*

A

1. *Randomly select k points (initial centroids) in the feature space.*
2. *Assign all samples to their nearest centroid.*
3. *Re-compute the centroid vector from the sample vectors in the emerged cluster.*
4. *Repeat steps 2 and 3 until all centroids have reached a steady state.*

B

1. *Select a cluster for partitioning (e.g. by random, or the largest available cluster).*
2. *From this cluster, form 2 clusters according to basic k-means rules.*
3. *Repeat step 2 n times and keep the partitioning with the highest similarity of the cluster members.*
4. *Repeat steps 1 through 3 until k clusters have been formed.*

1.7 Objectives

The objectives of the present work are:

1. To identify the current state of the art in the field of comparative phenomics, especially the current approaches in phenotype data integration, clustering of textual phenotype descriptions and data mining. Also, the possibilities involved in application of these techniques to textual phenotype data and the implications on gene function prediction shall be explored.
2. To extend PhenomicDB as a means for integrating phenotype data from more diverse sources and to include means for structuring these data, i.e. ontologies and controlled vocabularies.

3. To create a workflow for text clustering of phenotypes from PhenomicDB with the goal of assessing possibilities and benefits of gene function prediction from phenotype data, also in comparison to other methods.
4. To apply this workflow and implement an integrative system for genotype/phenotype data analysis and prediction in a usable and useful fashion for computational life scientists.

1.8 Contributions

The present work mainly contributes to the fields of comparative phenomics and gene function prediction, with special focus on the practical development and advancement of the field. Since large-scale data integration and data mining requires less of a theoretical advancement of algorithms and formulas, this work's central aspects lie in exploring the implications and benefits of clustering a large-scale cross-species data set, such as in PhenomicDB.

In particular, the main contributions of this thesis are:

1. The extensions to PhenomicDB, making it one of the world's largest integrated repositories of cross-species phenotype data and enabling a more structured view of the data using phenotype ontologies and controlled vocabularies. This will be shown in an in-depth exploration of the benefits coming from this integration. See section 3.1.
2. The application of k-means text-clustering to the entire corpus of PhenomicDB's textual phenotype descriptions and the exploration of the resulting phenotype clusters. Since each phenotype in PhenomicDB is directly associated with a gene, this method provides a novel way to group genes. These groups of genes are the basis for gene function prediction from functional annotation with competitive precision in comparison to other methods. Benefits and limitations of this approach will be shown in the course of this work. See section 3.2.
3. The development of PhenoMIX, a system to access cross-species genotype-phenotype data and to enable comprehensive data mining possibilities for all interested scientists, e.g. grouping similar phenotypes by textual description or common annotation with phenotype ontology terms. It can also be used to group genes with many pair-wise interactions, common functional annotations, high sequence similarities or cross-species homologies. As proof of concept, a set of human and mouse genes and phenotypes de-

rived through the use of this resource are evaluated for their predictiveness of functional annotations and phenotype terms. See section 3.3.

4. This work further contributes to the field of comparative phenomics with a comprehensive survey of the current standing of the field, giving insight into its most urgent needs and most promising developments. See chapters 1 and 4.

1.9 Structure of this thesis

Chapter 2 first gives an overview of materials and methods that have been used to produce the result of this thesis. These are genotype-phenotype data, protein-protein interactions, and the preparation of these data for clustering and grouping, as well as data for the extensions in PhenomicDB. In the second part of Chapter 2, I present the methods applied in text-clustering and prediction, as well as cross-validation and evaluation of biological coherence. In Chapter 3 the development and outcome of genotype-phenotype data integration, clustering and mining are presented. In this chapter I show how text clustering is used to group genes based on their phenotype descriptions. Such gene groups correlate well with several indicators for biological coherence, e.g. functional annotations from the Gene Ontology (GO) and protein-protein interactions (PPi). Grouped genes are used for predicting gene function by carrying over annotations from well-annotated genes to other, less well-characterized genes. For a subset of groups selected by applying objective criteria, GO-term annotations are predicted from the biological process sub-ontology with up to 72.6% precision and 16.7% recall, as evaluated by cross-validation. Furthermore, this chapter contains the results of the application of the integration and clustering into a system named PhenoMIX enabling comprehensive data mining possibilities. It is shown that the integrated data can be used for prediction of phenotypes from groups of similar genes as well as prediction of gene function annotations. Chapter 4 gives an overview over the achievements and contributions within this thesis, along with a thorough survey of related works and of the current status of the field of comparative phenomics, followed by a discussion of results, conclusions and outlook.

2 Materials and Methods

2.1 Materials

2.1.1 Gene-centered data

In PhenomicDB, gene-centered data are taken from NCBI Entrez gene for reasons stated by Kahraman et al. [Kahraman, et al., 2005], i.e. due to the availability of a common cross-species gene index and due to the existence of NCBI HomoloGene [Wheeler, et al., 2008] which is being used to interlink genes across species. As PhenomicDB is the main data source for this thesis, the genotype data was taken from PhenomicDB (Version 2.4, Source release date: 2007-10-04) downloaded at http://www.phenomicdb.de/downloads.html and supplemented, where necessary, with the data from Entrez gene.

Nucleotide sequences (see sections 2.2.4) were downloaded from the NCBI using RefSeq nucleotide IDs [Pruitt, et al., 2009] and the NCBI efetch tool (http://www.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=<identifier>&rettype=fasta) [Wheeler, et al., 2008]. To retrieve sequences by RefSeq nucleotide ID, each Entrez gene ID in PhenomicDB also found in column 2 of the file gene2refseq.gz (75,146 KB, downloaded on 2008-01-2 from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA), having one or more RefSeq nucleotide ID entries in column 4 of the same file was used. Where an Entrez gene entry referred to more than one RefSeq nucleotide entry, the longest sequence was taken.

For supplementing genotypes with GO-terms [Harris, et al., 2004] (used in sections 2.1.4, 2.2.3, 2.2.5, and 2.2.7), the GO-term IDs with reference to an Entrez gene ID in the file gene2go.gz (7,675 KB, downloaded on 2008-01-20 from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA) were used.

To match PPi data from IntAct [Kerrien, et al., 2007] (see section 2.1.3) with genotypes from Entrez gene, the file gene2refseq.gz mentioned above (75,146 KB, downloaded on 2008-01-2 from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA), was used again, this time matching Entrez gene IDs in column 2 with RefSeq protein IDs [Pruitt, et al., 2009] in column 6 of that file. Then, these RefSeq protein IDs were mapped to UniprotKB IDs [UniProtConsortium, 2009] (the format in which IntAct PPi data is stored, see section 2.1.3) using the file gene_refseq_uniprotkb_collab.gz (11,727 KB, downloaded on 2008-01-20 from ftp://ftp.ncbi.nih.gov/refseq/uniprotkb), in which the first column is a RefSeq

protein ID and the second column the corresponding UniProtKB ID. Thus, one pair of interacting proteins from IntAct is mapped to one or more Entrez gene IDs (see Table 1).

Table 1:
Example for referencing two interacting proteins from IntAct to Entrez gene IDs via RefSeq protein IDs and UniProtKB IDs.

IntAct UniprotKB	RefSeq protein IDs	Entrez gene IDs	IntAct UniprotKB	RefSeq protein IDs	Entrez gene IDs
P67662	NP_417710	947760	P76142	NP_416033	945418
P0A6N1	NP_417798	947838	P0A8I3	NP_414547	944749
	NP_418407	948482			
P53619	XP_612353	533078	P35604	NP_776707	281707

To associate Entrez gene IDs to taxonomy information (used for additional information on genotypes and species), supplementary genotype information from the NBCI's taxonomy database [Wheeler, et al., 2008] was taken from the file gene_info.gz (65,436 KB, downloaded on 2008-06-8 from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA). Further taxonomy information (such as Latin species names, trivial species names, etc.) was taken from the file taxdump.tar.gz (11,411 KB, downloaded on 2008-06-15 from ftp://ftp.ncbi.nih.gov/pub/taxonomy).

In order to remain up-to-date with Entrez gene IDs throughout the work on this thesis, the NCBI's gene history file gene_history.gz was downloaded and consulted regularly (last download: 2,734 KB, on 2008-06-10 from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA).

2.1.2 Orthology data

In this work, orthology of genes is one of the available measures for genotype-genotype relationship (similarity) to derive groups of functionally related genes (see section 1.2.4.1). Orthology information has been retrieved from the NCBI's HomoloGene database [Wheeler, et al., 2008], using the file homologene.data (10,114 KB, downloaded on 2007-11-1 from ftp://ftp.ncbi.nih.gov/pub/HomoloGene/current).

2.1.3 Data of protein-protein interactions

PPi data are used for evaluation of the biological coherence of gene groups (derived by clustering), and as a measure for similarity in PhenoMIX (see sections 1.2.4.2 and 2.1.1).

PPi data have been retrieved from both BioGrid (21,934 genes in 203,051 pair-wise interactions) [Breitkreutz, et al., 2008; Stark, et al., 2006] and IntAct (27,083 genes in 128,807 pair-wise interactions) [Kerrien, et al., 2007] due to quantity and quality of content and due to their marginal overlap. Evaluation of phenoclusters was based on the number of PPis between their associated genes. PPi data was obtained from the files BIOGRID-ALL-2.0.36.tab.zip and BIOGRID-IDENTIFIERS-2.0.36.tab.zip from the BioGrid website (downloaded on 2007-12-15 from http://www.thebiogrid.org/downloads.php). As PPi information, these files contain lists of Entrez gene IDs of genes encoding for interacting proteins. The binary protein interactions in UniprotKB-format from IntAct (file: intact.zip, downloaded on 2008-01-04 from ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab) was transformed into this format. Details on how IntAct data was transformed into lists of Entrez gene IDs in the same format as BioGrid data are shown in section 2.1.1.

2.1.4 Functional data from GO

The Gene Ontology [Harris, et al., 2004] data used here (see sections 1.2.5, 2.1.1, 2.2.3, 2.2.5, and 2.2.7) were taken from the file gene_ontology_edit.obo (in OBO v1.2 format, cvs version 5.658, downloaded on 2008-01-21 from http://www.geneontology.org/GO.downloads.ontology.shtml). GO-terms were matched to genotypes as described above (see section 2.1.1).

2.1.5 Phenotype annotations from MP

The Mammalian Phenotype ontology [Smith, et al., 2005] (see sections 1.3.2 and 2.2.3) is not only used for its annotations. Like GO, it is used to calculate the similarity of phenotypes from their MP-annotations. The methods are the same as described for GO (see section 2.2.3). Since the number of phenotypes annotated with MP-terms is an order of magnitude smaller than the number of genotypes annotated with GO-terms, MP-terms from the descriptive text of phenotypes were extracted by exact matching with an ontology term or a synonym thereof using a script written in the Perl programming language. Extracted terms were then stored as part of the MP-annotation of the phenotype.

Associations of genes from *Mus musculus* with GO-terms were taken from the file gene_association.mgi (version 1.38, downloaded on 2008-01-23 from ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno). MP-terms were taken from the file mammalian_phenotype.obo (in OBO v1.2 format, downloaded on 2008-01-18 from http://obofoundry.org/cgi-bin/detail.cgi?id=mammalian_phenotype). Associations of MGI

gene identifiers with gene identifiers from Entrez gene were taken from the file MGI_EntrezGene.rpt (downloaded on 2008-01-23 from ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno). Associations of genes with MP-terms were taken from the files MGI_PhenoGenoMP.rpt, MGI_PhenotypicAllele.rpt and MRK_Ensembl_Pheno.rpt (all three files downloaded on 2008-01-23 from ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno).

2.1.6 The cross-species phenotype data set

PhenomicDB [Groth, et al., 2007; Kahraman, et al., 2005] is a cross-species genotype/phenotype database integrating data from the Online Mendelian Inheritance in Man database (OMIM), the Mouse Genome Database (MGD), WormBase, FlyBase, the Comprehensive Yeast Genome Database (CYGD), the Zebrafish Information Network (ZFIN), DictyBase and the MIPS Arabidopsis thaliana database (MAtDB).

All phenotypes from PhenomicDB were considered. Of 428,150 (347,689) phenotype entries from PhenomicDB (version 2.1, Source release date: 2006-03-01 and in parentheses entries from version 2.4, Source release date: 2007-10-04, both downloaded from http://www.phenomicdb.de/downloads.html), 411,102 (327,201) entries are directly associated to at least one gene; only those were considered for this study. For each entry, its Entrez gene ID and the available text from all corresponding phenotype entries using the PhenomicDB fields ‘names’, ‘descriptions’, ‘keywords’ and ‘references’ were collected. Phenotypes with less than 200 characters were removed, as they are expected to be too short to deliver reasonable results in textual comparison and clustering (see section 4.3.2.2 for a discussion on this issue). All words were then stemmed using the stemming algorithm from the doc2mat package (implementing Porter’s stemming algorithm [Porter, 1980]) which is part of the clustering toolkit CLUTO version 2.1.1 (see section 2.2.2) [Zhao and Karypis, 2005]. Also, HTML-tags were removed, as well as so-called ‘stop-words’, which are words of such high frequency that they will not add to the distinctiveness of any feature vector (see Appendix A7 for the full list). This stop-word list comprises 348 unique words derived from the lists of the 319 most common words in the English language (downloaded from the Department of Computing Science, University of Glasgow, on 2006-10-16 from http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words) and the 133 most common words in PubMed (from the NCBI help pages, downloaded on 2006-10-16 from http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp). All

texts were then concatenated into a single string, called phenotype document or phenodoc (see Definition 1).

Definition 1: *Let p be a phenotype and T be the set of descriptive texts associated to that phenotype. A phenodoc is defined as the concatenation* $t_1 \circ t_2 \circ ... \circ t_n \in T$ *after removal of HTML-tags, stop-words and subsequent stemming of the remaining words in each* $t \in T$.

For this study, all 511 phenotypes linked to more than one gene (forming 1,227 'complex' genotype-phenotype relationships) were removed. Since PhenomicDB is not normalized with respect to replicate phenotype entries, there was no filter for those phenotypes associated with various genes (see Appendix A3 for some numbers about the phenotypes and phenodocs, see Figure 8, Figure 9, Figure 10 and Figure 11 for examples for phenotypes and their transition into gene-specific phenodocs and see Table 2 for how features translate into words). This resulted in a data set of 39,610 (38,656) phenodocs associated with 15,426 (15,431) genes from 7 species: 1.7% (1.9%) *Danio rerio* (zebrafish), 19.9% (20.0%) *Caenorhabditis elegans* (earth worm), 1.7% (1.7%) *Dictyostelium discoideum* (slime mold), 24.1% (23.9%) *Drosophila melanogaster* (fruit fly), 15.6% (15.8%) *Homo sapiens* (human), 28.7% (28.7%) *Mus musculus* (mouse) and 8.3% (8.0%) *Saccharomyces cerevisiae* (yeast). This strong data reduction (only ~9% (~11%) of all phenotypes in PhenomicDB passed the filters) will be discussed later (see section 4.3).

Table 2:

Excerpt from the feature translation table used to document the correspondence between features and words.

1 rnai	26 larval	44 drosophila	66 pupal	150 downstream
2 lethal	27 stage	45 melanogast	67 littl	151 compon
3 rang	28 wildtyp	…	68 increas	152 induct
...	…	60 systemat	69 mitot	153 nonleth
18 function	36 map	61 genom	70 index	154 invers
19 caus	37 orfeome-bas	63	71 mutant	155 arisen
20 non-viabl	38 librari	63 10a	…	156 vde2
21 phenotyp	39 fbcv0000351	64 heterozygot	148 absentia	157 clear
...	…	65 die	149 tramtrack	…

3354888 6511032 Mutant phenotype: rl⁹/rl^{10a} heterozygotes die in the pupal stage. In contrast to wild type, there is very little increase in the mitotic index in either rl⁹/rl^{10a} or rl¹⁰/rl^{10a} mutants after colchicine treatment for 3 hrs, and over-condensation of the chromosomes is not seen.

Mutant phenotype: rl⁹/rl² flies lack the R7 photoreceptor cell and some outer photoreceptors. This phenotype is dominantly suppressed by aop^{pok-1}.

Mutant phenotype: Hemizygous larvae completely lack imaginal discs and die. Individuals die as pupae when heterozygous with rl¹ or rl⁶.

Mutant phenotype: L3 larval lethal as hemizygote no imaginal discs pupal lethal when heterozygous with rl¹ or rl⁶ (Hilliker, 1976).

Mutant phenotype: Mutants have a have a mild rough eye phenotype. FBcv0000351:lethal FBcv0000351:lethal | FBdv00005336:larval stage | FBcv0000298:recessive FBcv0000354:visible The Drosophila Ral GTPase regulates developmental cell shape changes through the Jun NH(2)-terminal kinase pathway. Genetic analysis of rolled, which encodes a Drosophila mitogen-activated protein kinase. rugose (rg), a Drosophila A kinase anchor protein, is required for retinal pattern formation and interacts genetically with multiple signaling pathways. Cytogenetic analysis of the second chromosome heterochromatin of Drosophila melanogaster. The Drosophila rolled locus encodes a MAP kinase required in the sevenless signal transduction pathway. The sevenless signaling cassette mediates Drosophila EGF receptor function during epidermal development. Genetic analysis of the centromeric heterochromatin of chromosome 2 of Drosophila melanogaster: deficiency mapping of EMS-induced lethal complementation groups. Involvement of the rolled/MAP kinase gene in Drosophila mitosis: interaction between genes for the MAP kinase cascade and abnormal spindle. Genetic interactions of pokkuri with seven in absentia, tramtrack and downstream components of the sevenless pathway in R7 photoreceptor induction in Drosophila melanogaster. The new Drosophila melanogaster nonlethal inversion, arisen from the In(2R)bw^{VDe2}.

Figure 8:

Example of a phenotype as it is retrieved from PhenomicDB in textual form. The first integer is the Entrez gene ID associated with this phenotype, next is the PhenomicDB phenotype ID, and then follows its description.

6511032 10a heterozygot die pupal stage wildtyp littl increas mitot index 10a 10a mutant colchicin hr over-condens chromosom seen fly lack photoreceptor outer photoreceptor phenotyp dominantli suppress aop pok-1 hemizyg larva complet lack imagin disc die individu die pupa heterozyg larval lethal hemi zygot imagin disc pupal lethal heterozyg hillik mutant mild rough eye phenotyp fbcv0000351 lethal fbcv0000351 lethal fbdv00005336 larval stage fbcv0000298 recess fbcv0000354 visibl drosophila ral gtpase regul shape jun termin kinas pathwai genet roll encod drosophila mitogen-activ kinas rugos drosophila kinas anchor requir retin pattern format interact geneticli multipl signal pathwai cytogenet second chromosom heterochromatin drosophila melanogast drosophila roll locu encod map kinas requir sevenless signal transduct pathwai sevenless signal cassett mediat drosophila egf receptor function epiderm genet centromer heterochromatin chromosom drosophila melanogast defici map ems-induc lethal complement involv rolled/map kinas drosophila mitosi interact map kinas cascad abnorm spindl genet interact pokkuri seven absentia tramtrack downstream compon sevenless pathwai photoreceptor induct drosophila melanogast drosophila melanogast nonleth invers arisen vde2

Figure 9:

The same phenotype as in Figure 8 after stemming and removal of tags, stop words and Entrez gene ID. The phenotype description has now the phenodoc format as defined in Definition 1.

6511032 2 5 18 1 21 2 26 2 27 2 28 1 36 3 39 2 40 3 44 10 45 4 63 3 64 1 65 3 66 2 67 1 68 1 69 1 70 1 71 2 72 1 73 1 74 1 75 3 76 1 77 1 78 2 79 3 80 1 81 1 82 1 83 1 84 1 85 1 86 1 87 1 88 2 89 2 90 1 91 1 92 2 93 1 94 1 95 1 96 1 97 1 98 1 99 1 100 1 101 1 102 1 103 1 104 1 105 1 106 1 107 1 108 1 109 1 110 6 111 4 112 2 113 2 114 1 115 1 116 1 117 2 118 1 119 1 120 1 121 3 122 1 123 1 124 3 125 1 126 1 127 2 128 1 129 3 130 1 131 1 132 1 133 1 134 1 135 1 136 1 137 1 138 1 139 1 140 1 141 1 142 1 143 1 144 1 145 1 146 1 147 1 148 1 149 1 150 1 151 1 152 1 153 1 154 1 155 1 156 1

Figure 10:

The same phenotype as in Figure 9 represented as vector showing term frequencies. Each word has been translated into a numeric feature and has received its frequency (TF-) score.

6511032 2 0.08460 18 0.00983 21 0.02114 26 0.03343 27 0.03441 28 0.00507 36 0.05651 39 0.05698 40 0.08022 44 0.17934 45 0.10405 63 0.26168 64 0.04040 65 0.11750 66 0.09284 67 0.05253 68 0.03795 69 0.04489 70 0.05860 71 0.04908 72 0.08393 73 0.07311 74 0.09344 75 0.06396 76 0.03538 77 0.03722 78 0.07061 79 0.13850 80 0.05561 81 0.06376 82 0.04011 83 0.07985 84 0.09870 85 0.049511 86 0.04228 87 0.03877 88 0.09206 89 0.08387 90 0.04445 91 0.05650 92 0.08142 93 0.05375 94 0.04676 95 0.09569 96 0.04654 97 0.04934 98 0.03292 99 0.04669 100 0.02929 101 0.02698 102 0.03575 103 0.03439 104 0.08179 105 0.05542 106 0.02872 107 0.04494 108 0.06235 109 0.04460 110 0.22430 111 0.13932 112 0.13059 113 0.06509 114 0.06578 115 0.07755 116 0.06244 117 0.05220 118 0.04803 119 0.03343 120 0.03452 121 0.10136 122 0.05171 123 0.03634 124 0.09472 125 0.04608 126 0.03967 127 0.11870 128 0.03202 129 0.18175 130 0.05008 131 0.06823 132 0.03837 133 0.05299 134 0.03493 135 0.05090 136 0.05917 137 0.03960 138 0.06878 139 0.04471 140 0.03413 141 0.09014 142 0.05193 143 0.06056 144 0.02665 145 0.04308 146 0.08723 147 0.05871 148 0.08465 149 0.07280 150 0.05402 151 0.04378 152 0.04452 153 0.08828 154 0.05554 155 0.06547 156 0.08628

Figure 11:

The same phenotype as in Figure 10, now in a vector format, where term frequencies have been multiplied by their respective inverse document frequencies (i.e. TFIDF-score).

2.1.7 Phenotype and disease-specific vocabularies

Specialized phenotype and disease vocabularies were used in section 3.2.3 to over-weight TF- and TFIDF-scores of vocabulary terms in phenotype vectors (see 2.2.1 for methods). Utilized vocabularies were GO-terms (see section 2.1.4), the Medical Subject Headings (MeSH) [Nelson, et al., 2004] (Files: Descriptors: d2007.bin, Qualifiers: q2007.bin, Supplementary Concept Records: c2007.bin, source release date for all files: 2007-08-01, downloaded from http://www.nlm.nih.gov/mesh/filelist.html) and MP-terms (see section 2.1.5).

2.1.8 Data for phenocopies

Usually, a phenocopy is a condition produced by an environmental effect that mimics the condition produced by a gene. However, there are phenocopies which are independently induced by two or more different genes. In an extensive manual search of Medline literature on such phenocopies induced by different genes, 27 such phenocopies, induced by 57 genes in total, were identified and used here (see Appendix A4 for details of the phenocopies and the literature references).

2.1.9 Data integration

Data integration for PhenoMIX has been done in the same way as for PhenomicDB by Kahraman et al. [Kahraman, et al., 2005]: the physical (i.e. materialized) integration of phenotype data from various sources, i.e. the Online Mendelian Inheritance in Man database (OMIM) [McKusick, 2007], the Mouse Genome Database (MGD) [Bult, et al., 2008], WormBase [Rogers, et al., 2008], FlyBase [Wilson, et al., 2008], the Comprehensive Yeast Genome Database (CYGD) [Guldener, et al., 2005], the Zebrafish Information Network (ZFIN) [Sprague, et al., 2008], and the MIPS Arabidopsis thaliana database (MAtDB) [Schoof, et al., 2004], where each data field from each data source was mapped manually (by 'course-grained semantic mapping' [Kahraman, et al., 2005]) to the data fields of the target database. Furthermore, the genotypes directly associated to each phenotype were manually mapped back to a common gene index, namely the NCBI's Entrez gene index [Maglott, et al., 2007; Wheeler, et al., 2008]. Finally, genotypes where grouped by species using orthology information taken from the NCBI's HomoloGene [Wheeler, et al., 2008].

2.1.10 Extensions in PhenomicDB

In the first part of this thesis, PhenomicDB was extended by over 250,000 RNAi phenotypes, now making up roughly two thirds of the phenotypes in PhenomicDB (see section 3.1 for more results from the extension effort). In addition to the data sources mentioned above (see section 2.1.9 and section 4.2.4), new data sources have been made accessible (all data sources are listed in section 4.2.4 with more details). These new data sources are:

1. FlyRNAi, maintained by the Drosophila Resource Screening Centre (DRSC) (13,900 targeted genes; available at http://www.flyrnai.org) [Flockhart, et al., 2006],

2. PhenoBank (24,671 RNAi phenotypes for 20,981 genes; available at http://www.phenobank.org) [Gonczy, et al., 2000; Gonczy, et al., 1999; Sonnichsen, et al., 2005],
3. RNAiDB (59,991 RNAi phenotypes; available at http://www.rnai.org) [Gunsalus, et al., 2004],
4. PharmGKB (608 targeted genes; available at http://www.pharmgkb.org) [PharmGKB, 2008].

Extensive RNAi data was also derived from previously integrated data sources that had to be updated due to their novel content, namely WormBank (74,602 RNAi phenotypes on 77,763 genes; available at http://www.wormbase.org) and FlyBase (150,000 phenotypic statements on 14,029 genes; available at http://www.flybase.org) [Tweedie, et al., 2009; Wilson, et al., 2008].

To properly host these data, the PhenomicDB database scheme was updated (see section 3.1 for details and results). The schemes were tested and populated with the above mentioned data using two simple Java classes (one for database handling and one for database queries) to connect to, update and query the database. The finished MSSQL scheme was presented to database experts for revision and was then sent to the collaboration partner metalife AG for implementation along with the SQL statements and the raw data. Subsequently, the data were mapped by hand to the fields of the newly created tables (see also the integration methods described in section 2.1.9).

In order to present the data to the user in a comprehensive and structured format, several mock-up screenshots were written in HTML using the PhenomicDB style sheets and the newly integrated data. These mock-ups were then presented to biologists and database experts for revision and then sent to the metalife AG for implementation.

2.2 Methods

2.2.1 Pair-wise similarity measures for phenotypes

The textual descriptions of phenotypes were used as the basis for feature vectors, in which each feature is equal to a word within the phenotype description. The features were then weighted according to their importance within the description and within the set of documents and thus, phenotypes could be compared in their feature space. There are several possibilities for weighting terms by their importance. The most common and most broadly

applicable weighting schemes are term frequency (TF) and term frequency-inverse document frequency (TFIDF, see Definition 2). These weighting schemes are very common, since they account for the importance of each term within the document. The multiplication of the term frequency by inverse document frequency (=TFIDF) 'discounts frequent words with little discriminating power' [Steinbach, et al., 2000] and thus is a better weighting scheme when there are many documents under consideration, which is the reason why this is the weighting scheme of choice here. There are many other weighting schemes [Cummins and O'Riordan, 2005; Manning, et al., 2008], e.g. an (over-) weighting of terms on their occurrence within a specialized dictionary, also applied here in order to test whether over-weighted phenodocs have a better correlation of similarity to biological coherence than those with the TFIDF-weighting scheme (see section 3.2.3).

For this, I used the phenotype- and disease-specific vocabulary described previously (see section 2.1.7). With the help of a script written in the Perl programming language, each term (and its synonyms where applicable) was parsed from each of the term lists and matched to the phenotype documents with simple exact matching. Wherever there was such a match in the phenotype description, the term would be weighted tenfold in contrast to its original frequency (TF x 10). Stemming (where applicable) was performed after this step. The results of these efforts are described in section 3.2.3.5.

Definition 2 *(from: [Steinbach, et al., 2000]): Given an alphabet* $\Sigma = \{a, b, \ldots, z\}$ *and the set* $\Sigma^* = \{w \mid w = x_1, \ldots, x_n, x_i \in \Sigma, i, n \in N\}$ *of all words from Σ, a set of documents* $D = \{d \mid d = w_1, \ldots, w_n, w_i \in \Sigma^*, i, n \in \mathrm{N}\}$ *and a set of terms* $T \subseteq \Sigma^* : t \in d, d \in D$*, where t is the* i^{th} *of k different terms of a document d and n the number of occurrences of term t in d, then there are weighting schemes* $TF_{t_i} := \frac{n_i}{\sum_{j=1}^{k} n_j}$ *and* $IDF_{t_i} := log\left(\frac{|D|}{\{d \mid t_j \in d\}}\right)$*, with* $TFIDF_{t_i} := TF_{t_i} \times IDF_{t_i}$*, where* $i, k, j \in N$.

Once a document (or rather a set of documents) has been transferred into numerical feature vectors, similarity measures can be applied in order to find sets of similar documents

within the entire set. The similarity measure for comparing document vectors used is the cosine similarity, a measure of correlation between length-normalized vectors (see Definition 3) [Steinbach, et al., 2000; Zhao and Karypis, 2002].

Definition 3 *(from: [Steinbach, et al., 2000; Zhao and Karypis, 2002]): The cosine similarity* $sim(\vec{x},\vec{y})$ *of two length-normalized vectors* $\vec{x}$ *and* $\vec{y}$ *of length* n *is equal to* $1-cos(\vec{x}\cdot\vec{y})$*, representing the angle between them* $(i,n\in N, i\le n)$*:*

$$\begin{aligned} sim\ (\vec{x},\vec{y}) &= 1 - cos\ (\vec{x},\vec{y}) \\ &= 1 - \frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|} \\ &= 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\cdot\sqrt{\sum_{i=1}^{n} y_i^2}} \end{aligned}$$

These similarity values range from 0 (no similarity) to 1 (identity). The normalization of the vectors over their lengths results in the property that only the direction of the two vectors is an indication of their similarity. In a feature space of terms, this direction correlates with the 'theme', or the textual content of the document represented by the vector. Thus, two vectors pointing in the same direction (with a small angle between them) contain many similar words, making it a reasonable and useful similarity measure.

2.2.2 Phenotype clustering with CLUTO

The software used for clustering was CLUTO (Clustering Toolkit, version 2.1.1, 9.3 MB, file cluto-2.1.1.tar.gz downloaded on 2006-10-3 from http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download) [Zhao and Karypis, 2003]. It is up-to-date, fast and simple open-source software for clustering and was thus used for all clustering applications here. Specifically, the scalable implementations of the bisecting k-means algorithm from the CLUTO package, called *vcluster* and *scluster*, were used. From the available clustering algorithms in the public domain, CLUTO has been chosen here for

its reliability and availability. Furthermore, it has been proven to work well on textual data sets from the life science domain on several occasions [Steinbach, et al., 2000; Tagarelli and Karypis, 2008; Zhao and Karypis, 2003; Zhao and Karypis, 2005]. Also, it should be noted that quality and coherence of a clustering depend more on the choice of parameters than on the algorithm's implementation (assuming this has been done with reasonable care).

Definition 4 *(from: [Zhao and Karypis, 2002]): For a subset A of documents* $d \in D$*, represented by their vectors, the composite vector* $K_A = \sum_{d \in A} d$ *is the sum of the vectors of documents from A and the centroid* $C_A = \frac{K_A}{|A|}$ *is the mean of the features of A. The criterion function* $I_2 = \sum_{r=1}^{k} \sum_{d_i}^{D} cos(d_i, C_r) = \sum_{r=1}^{K} \|K_r\|$ *maximizes the similarity between document and centroid for each available document, where* $i, k, r \in N$.

There are three main parameters to the k-means algorithm: k (the command-line parameter *nclusters*, i.e. the number of clusters), the similarity (or distance) measure (the command-line parameter *colmodel*, e.g. 'idf' when the input features' values are given as term frequencies and the TFIDF scoring scheme should be calculated, or 'none' when the given values should be accepted as they are) and the criterion function (the command-line parameter *crfun*, typically I_2, see Definition 4). The choice for the criterion function, i.e. the function that uses the similarity measure in order to assign samples to their best (i.e. nearest) centroid is highly dependent on the choice of similarity measure, which in turn is dependent on the clustered data [Zhao and Karypis, 2002]. The cosine similarity measure is the best choice for document clustering (see section 2.2.1) for which CLUTO's criterion function I_2 (see Definition 4) has been suggested [Zhao and Karypis, 2002].

Clustering with *vcluster* was done as follows. Prior to clustering, the phenotypes had to be prepared as phenodocs (see Definition 1 in section 2.1.6) and transferred into vectors (see section 2.1.6), stored in a single file (one line for each phenodoc) containing just the pairs of feature identifiers and frequencies for each vector (omitting the vectors' identifiers,

since they are identified by the order of occurrence within the file). A single line was added to this file as header, consisting of three separate integers denoting the number of vectors (total number of lines), the number of unique features (words) and the total number of features in all phenodocs, respectively. This file, called the *matrix* file, was stored and given as the command-line parameter *filename.mat*. Two other very useful optional files could be created, namely the *rlabel* file (giving the phenodocs' identifiers in order of appearance in the *matrix* file) and the *clabel* file, containing all unique words as represented by feature identifiers in the *matrix* file. For convenience, the vectors could also be created by the doc2mat package which is part of CLUTO (see also section 2.1.6). The program takes as input a file of phenotypes or phenodocs with identifiers and textual descriptions (one per line) and returns the three files named above. When given phenotypes, doc2mat can be used to create phenodocs, as it optionally allows for stemming of words with Porter's stemming algorithm (see 2.1.6 and [Porter, 1980]) and also offers the option to remove stop-words (either generic or from a list). It offers, however, no cut-off for character or feature count (as it has been applied, see section 2.1.6) and offers no options for word weighting and can only calculate term frequencies. The clustering program was then run with '*vcluster -colmodel=<string> -crfun=<string> <filename.mat> <nclusters>*' on command-line.

In another setting, where instead of vectors, there was only a matrix of pairwise similarities available (e.g. for clustering of sequences by their pairwise sequence similarities), the utilized program was *scluster*, also from the CLUTO package. The data input for *scluster* is called a graph file, in which each line represents a vertice (or node) in a graph and each entry consists of a pair of values, one integer value as the identifier of a vertice connected to the vertice represented by the current line in the file followed by whitespace and a floating point value representing the value (or weight) of the edge between the two vertices followed by one or more such pairs of vertice identifier and edge weight. The file also contains a header consisting of two integers, denoting the number of vertices and the number of edges in the file respectively. The program is also called on command-line with 's*cluster -crfun=<string> <filename.graph> <nclusters>*'.

The result of the *vcluster* and *scluster* is the file *filename.mat.clustering.nclusters* containing an integer between 0 and *nclusters*-1 in each line, representing the cluster identifier number for the vector in the corresponding line from the input file. Now, using the pheno-

type identifiers correspondingly stored in the *rlabel* file, each cluster identifier is associated with at least one phenotype identifier, resulting in *phenoclusters* (see Definition 5).

> ***Definition 5:*** *Phenoclusters are defined as the result of applying a k-means clustering to n phenodocs resulting in k clusters, each associated with at least 1 and at most n-k+1 phenotypes,* $n,k \in N, n \geq k$ *. A phenocluster is any one of those k clusters.*

2.2.3 Gene similarity based on GO-annotation

The similarity of two genes can be calculated effectively using their GO-annotations. Generally, there are two different approaches to measuring the similarity of pairs of sets of GO-terms, namely to analyze the graphs induced by the terms of either set, or to measure the frequency of term occurrence in samples (reflecting information content of both terms). The latter approach has been declared most effective in a study by Guo et al., comparing many similarity measures of either class and was therefore used here to calculate GO-term similarities [Guo, et al., 2006]. Based on the two general approaches, there are several other similarity measures for two groups of GO-terms, resulting in similarity measures for genes (see section 4.2.2 for an overview).

Here, a variation of the approach suggested by Lord et al. was used [Lord, et al., 2003]. Instead of considering all pairs of terms from either set, only the k best-scoring term pairs (with k being the size of the smaller term group) were used, where each term from the smaller set contributes exactly its best similarity score with one of the terms from the larger set. In case of an equal number of terms in both sets, all of them were paired and all pairwise similarity scores were used (see Definition 6). For example, if one gene has ten associated GO-terms and another gene has only three associated GO-terms, the GO-similarity of the two genes is the mean of the three highest similarity scores from the three terms of the smaller set with any one of the terms from the larger set. This score better reflects the nature of GO-annotations, where genes are unequally well annotated or genes only partially share functions with each other. For calculating the similarity of two GO-terms, the similarity measure proposed by Lin was used here (see section 4.2 or [Lin, 1998]). This score is intuitive, because resulting scores range between 0 (when the terms are connected

only via the root) and 1 (if both sets of terms are equal). Since beginning this thesis, this definition has been independently described and published as the optimal way to measure gene similarity by Frohlich et al. (see section 4.2.2 or [Frohlich, et al., 2007] for more details), with the variation that they suggest applying maximum weighted bipartite matching (see e.g. the book by West for details [West, 1999]) when both sets of terms are of equal size.

As is the case for genes, phenotype similarity can also be calculated from ontology terms, here from MP (see sections 1.3.2 and 2.1.5). Calculation is carried out as described in this section. These resulting similarities were used in section 3.3.

> ***Definition 6:*** *For any term t that is part of the set of terms T making up the biological process sub-ontology GO_{BP}, a gene g is associated with one or more terms $t \in T$. For two genes g_h associated with the set T_k of n terms $t \in T$ and g_j associated with the set T_l of m terms $t \in T$, the similarity $sim(g_h,g_j)$ is defined as the sum of the highest similarity score (calculated using the formula by Lin [Lin, 1998]) from each term of the smaller of the two term sets with any term from the larger set. If both sets are of equal size, the highest pair-wise similarity score is summed for each term of both sets. To adjust for differing sizes of term sets, the resulting sum is divided by the number of total term pairs considered.*

2.2.4 Gene similarity based on sequence

The most commonly used similarity measure for two genes is the similarity of their respective sequences, also utilized here. The three most popular algorithms to calculate this similarity are BLAST [Altschul, et al., 1990], the Smith-Waterman local alignment algorithm [Smith and Waterman, 1981] and the Needleman-Wunsch global alignment algorithm [Needleman and Wunsch, 1970]. Global alignment algorithms are designed to align every letter in every sequence and are most useful for similar and equally sized sequences. Local alignments are commonly used for dissimilar sequences containing several shorter stretches of similarity within the entire sequence. With sufficiently similar sequences, there is basically no difference between local and global alignments [Brudno, et al., 2003].

BLAST compares a query sequence with many sequences from a database and calculates the statistical significance of local matches between query and database entries. This makes BLAST a very useful tool to find the best matches for a sequence. Its great advantage is that it outperforms the Smith-Waterman algorithm on large databases of sequences in terms of time by at least an order of magnitude [Altschul, et al., 1990]. The main disadvantage of BLAST is that it is a heuristic, thus giving not the optimal local alignment like the Smith-Waterman algorithm does. However, when pair-wise comparisons between many sequences are needed, especially when looking for all pair-wise similarity scores between a set of sequences, the performance advantages of BLAST are enormous. Thus, BLAST (version 2.2.9, 27,868 KB, file blast-2.2.18-ia32-linux.tar.g downloaded on 2008-07-10 from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/2.2.9) was the algorithm of choice applied here for sequence similarity calculations. It is freely available from the NCBI BLAST webpage (http://www.ncbi.nlm.nih.gov/BLAST/download.shtml).

The gene sequences used for similarity calculations (see section 2.1.1) were stored in a file and formatted into a BLAST-able database file (*seqs.fasta* used by BLAST with command-line option *-d*) by a program named *formatdb* delivered with BLAST. Then, a script written in the Perl programming language consecutively extracted each sequence from the original file, stored it in a temporary file *tmp.txt* (command-line option *-i*) and applied BLAST on command-line (using command-line option *–p blastn* for nucleotide sequences) with '*blastall -p blastn -i tmp.txt -d seqs.fasta -v 0 -b 100*' to retrieve all pair-wise alignments of this sequence with all other sequences in the original file. The results were stored when query and target sequence showed an identity value greater than or equal to 0.1.

2.2.5 Assembling groups of similar genes

Groups of genes are assembled from phenoclusters (used in section 3.2) but also from other similarity measures (used in section 3.3):

1. Interactions in Intact and BioGrid (see sections 1.2.4.2 and 2.1.3),
2. Orthology (see sections 1.2.4.1 and 2.1.2),
3. GO-similarity (see sections 1.2.5, 2.1.4 and 2.2.3), measured separately in all of the three sub-ontologies,
4. Nucleotide sequence similarity (see sections 2.1.1 and 2.2.4),

5. Using groups of similar phenotypes, measured by cosine similarity (see sections 2.1.6 and 2.2.1) MP-similarity (see sections 1.3.2, 2.1.5 and 2.2.1), calculated analogously to gene similarity from GO-terms (see section 2.2.3).

For phenoclusters, groups of genes were derived by looking up the associated genes to the phenotypes in PhenomicDB (see section 2.1.7) and joining all genes together into one group for which associated phenotypes are together in the same phenocluster.

For Boolean similarity measures, i.e. interactions and orthology, all genes belonging together were joined into one group, e.g. when gene A interacts with B and gene B interacts with gene C, genes A, B and C form one group. In any other case, groups were assembled analogously, defining as threshold a similarity measure of 0.8. Where groups of phenotypes were assembled by this method, i.e. using the cosine similarity measure and MP-similarity, gene groups were built the same way as it was done for phenoclusters.

2.2.6 Correlation between GO-similarity and phenodoc similarity

For each gene pair from a 'phenocluster', two measures were calculated: the GO-similarity score (see section 2.2.3) and the mean pair-wise similarities of phenodocs associated with the genes (see section 2.2.1). From these measures, the mean cluster GO-similarity and phenodoc similarity was computed, resulting in two functional measures of which the degree of correlation was calculated using the Pearson correlation coefficient r (ranging from $r = -1.0$, with perfect inverse linear correlation, over $r = 0.0$, no correlation, to $r = 1.0$, perfect linear correlation). This correlation coefficient was used for cluster coherence assessment.

2.2.7 Prediction of functional annotation

One of the observations made during the clustering experiment showed that the resulting groups of genes are highly enriched in terms of coherence of their functional annotation (see section 3.2.2.2), suggesting that genes within a phenocluster have a high chance of sharing gene function. This led to the hypothesis that phenoclusters can be used for function prediction by transferring the function of badly characterized or un-annotated genes in a cluster from well characterized genes in the same cluster (see Figure 12). To estimate precision and recall (see section 2.2.8) of this approach, all gene groups with at least three members were considered, but groups with no GO-terms common to at least 50% their members were disregarded. Each resulting group was then randomly partitioned into a

training set of at most 90% of its genes and a test set of at least one gene and at most 10% of genes. The terms associated to the training set were 'predicted' as new annotations to all genes in the test set of the same group. Then, these predictions were compared to the real annotation of the test genes to judge prediction correctness. This procedure was repeated 200 times (with different training / test sets) and the means of precision and recall values of the suggested terms were computed (see section 2.2.8). To measure an empirical threshold (p-value ≤ 0.05), randomly populated gene groups of equal size were used.

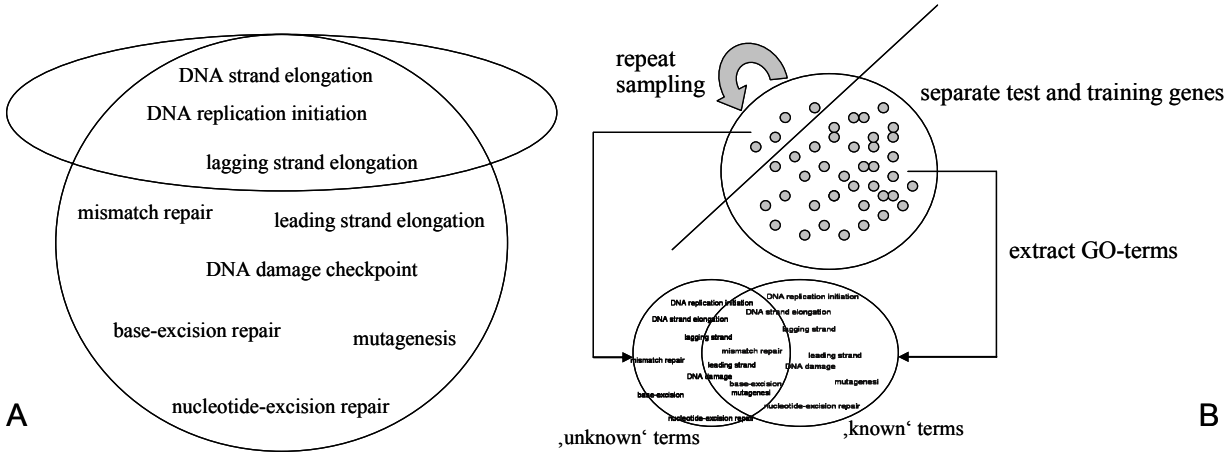


Figure 12:

Schematic view of gene function prediction for a group of genes.

A. Genes of a group annotated with several different GO-terms (large circle). To predict functional annotations, only GO-terms commonly annotated to at least 50% of members are considered (small oval).

B. From a group of genes, a training set of 90% of those genes is separated from a test set of 10%. The GO-terms associated with the genes are divided into a set of 'known' terms (the prediction) and 'unknown' terms (the predicted). From the resulting overlap of these two groups, precision and recall can be calculated.

Analogously, groups of genes assembled as described in section 2.2.5 were used to predict GO-terms (as described above for phenoclusters), and also phenotype terms from their associated phenotypes (see sections 2.1.1 and 2.1.6 for details on the data set). For this, genes associated with at least one phenotype and only groups with at least three members were considered. From each associated phenotype, the entire phenotype description was extracted (after stemming and stop-word-removal as described in section 2.1.6) and unique words were merged (when there was more than one associated phenotype) and stored as

annotation to the gene (as if they were GO-terms). The subsequent prediction of (phenotype) terms and cross-validation were done exactly as described above.

2.2.8 Precision and recall

Precision and recall (see Definition 7 and Definition 8) are important parameters to evaluate search and prediction methods. Prerequisite is the knowledge of the amounts of relevant data that can be found in a data set. The amount of data that actually will be found in this set should overlap to some degree with the relevant data and depends on the quality of the applied search or prediction method (see Figure 13) [Jizba, 2000]. In absence of a 'gold standard', where a set of genes is fully annotated with all applicable GO-terms, any prediction may be correct, but cannot (currently) be verified. As a result, TP=FP may hold for both definitions below (see also the comment in section 3.2.2.4).

Definition 7: *Precision is the relation between the correctly predicted annotations (TP) and the annotations that have been predicted (TP+FP). High precision means a small fraction of wrong predictions. In terms of Figure 13, where M_1 represents all available annotations, M_2 represents annotations that should be predicted and M_3 represents annotations that have been predicted.* $Precision := \frac{|M_2 \cap M_3|}{|M_3|}$.

Definition 8: *Recall is the relation between the correctly predicted annotations (TP) and the annotations that should have been predicted (TP+FN). High recall means a high percentage of correct predictions. In terms of Figure 13, where M_1 represents all available annotations, M_2 represents annotations that should be predicted and M_3 represents annotations that have been predicted.* $Recall := \frac{|M_2 \cap M_3|}{|M_2|}$.

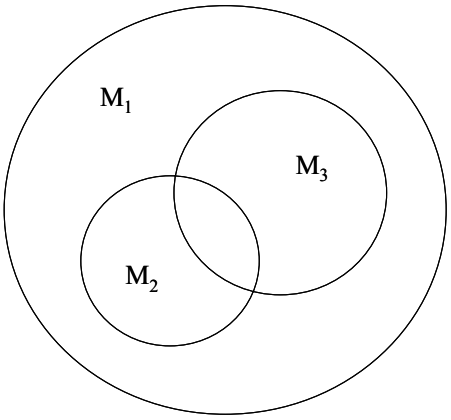


Figure 13:

In a search for relevant data (M_2) within a total set of data (M_1), a set of data (M_3) overlapping with the relevant data will be retrieved. The degree of overlap is dependent on the quality of the applied method (from: [Jizba, 2000] / adapted: P. Groth, 2008).

3 Results

3.1 PhenomicDB: Integration of genotype/phenotype data

3.1.1 Background

Almost all of the available genotype-phenotype resources that have been reviewed for this thesis (see section 4.2.4 and the review by Groth and Weiss [Groth and Weiss, 2006a]) are limited in scope and in content. It is common practice for a research group to keep their results in a (more or less public) database limited to their species of interest. The situation is worse for RNAi data, where the published data is locked up (in terms of availability for automatic processing) as supplementary material or in a specific database for one single screen (as in PhenoBank). A result of this one-sided view is the abundance of community-specific vocabulary used by the different research groups to describe their species.

In 2004, the first version of PhenomicDB was created by Kahraman et al., 'in order to remedy to this situation' that there is no integrative system of genotypes and phenotypes across species and screening methods [Kahraman, et al., 2005]. They gathered data from the different public resources and mapped them into a single data model (see Appendix A1.1) initially termed 'Multi-Species Genotype/Phenotype Database' (MSP). With the help of metalife AG they created the database's first productive version, since then termed PhenomicDB (see Appendix A1.2 for the database scheme of PhenomicDB version 1.x).

3.1.2 New structures and data for PhenomicDB

The publicly available genotype-phenotype data has increased steadily since 2004, when especially large amounts of RNAi data came into the public domain. This development can be traced by the near-7-fold increase of available genotype-phenotype information of the data content from the Drosophila Resource Screening Centre (DRSC), the only resource dedicated to RNAi data (see Figure 14 and section 4.2.4 for details, also on the data increase for other resources). Thus, PhenomicDB had to be extended in order to host these new data and make them available in a semi-structured way. One of the results of this thesis is therefore the remodelling of the database scheme, update and extension of data content and scope as well as a restructuring of data presentation, moving PhenomicDB towards its current version 2.6 (April 2009).

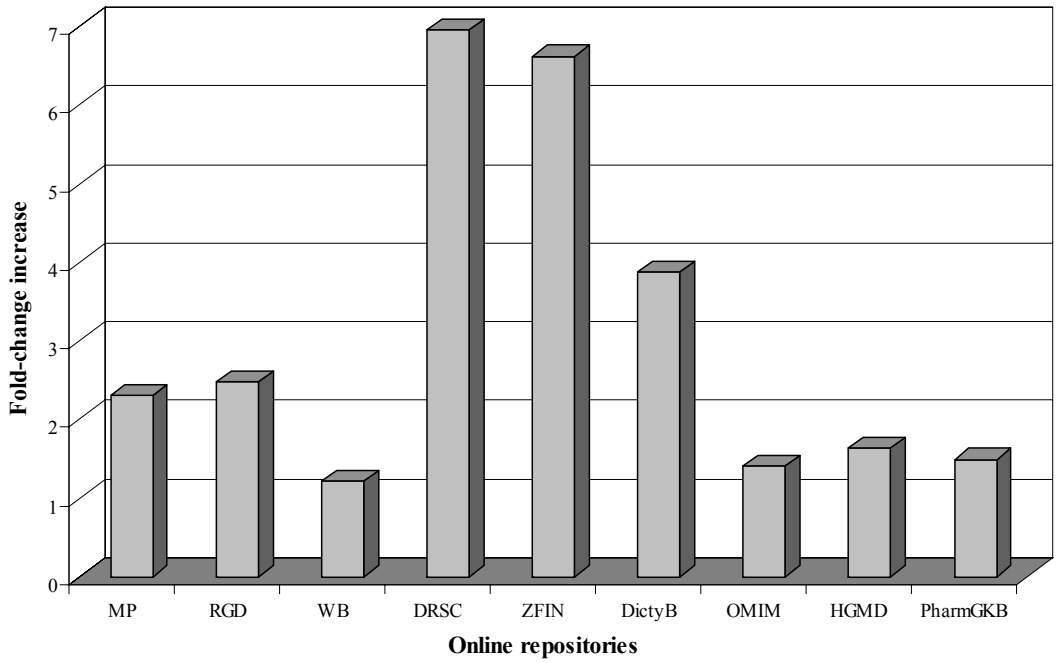


Figure 14:

Fold-change of the increase in the number of genotypes with available phenotype information from 2006 (data taken from [Groth and Weiss, 2006a]) to 2008 (data taken from section 4.2.4) for 9 different online genotype-phenotype resources. It can be seen for example that data from the DRSC, the only dedicated RNAi database in this list has increased almost 7-fold in that period.

MP = Mammalian Phenotype ontology [Smith, et al., 2005];

RGD = Rat Genome Database [de la Cruz, et al., 2005];

WB = WormBase [Rogers, et al., 2008];

DRSC = Drosophila Resource Screening Centre [Flockhart, et al., 2006];

ZFIN = Zebrafish Information Network [Sprague, et al., 2008];

DictyB = DictyBase [Chisholm, et al., 2006];

OMIM = Online Mendelian Inheritance in Man [McKusick, 2007];

HGMD = Human Genome Mutation Database [Cooper, et al., 2006];

PharmGKB = Pharmacogenetics and Pharmacogenomics Knowledge Base [Hodge, et al., 2007]

3.1.3 Home for large-scale RNAi data

The extension of PhenomicDB towards version 2.x has primarily been targeted towards the inclusion of RNAi phenotypes. Previously, it was already possible to insert such phenotypes. However, these data typically consist of a very short description, like e.g. ‘embryonic lethal’ or ‘binucleated cell’. Thus, one of the main goals of the extension was to enable supplementing these data with more information on the cellular conditions during the

screen, on the type of screen itself, as well as the siRNA sequence used for the silencing. Another novel feature in PhenomicDB version 2.x is the possibility to describe these short phenotypes by a controlled vocabulary in order to avoid inconsistency in information, when for example the same phenotype observed in one screen is labeled 'lethal' and 'non-viable' in a second screen. For this purpose, the phenotype terms from MP were added as a controlled vocabulary. All these data mentioned above are now stored in a newly designed database scheme with new tables hosting siRNA sequences and information, controlled vocabulary terms for phenotypes and cell types, tables for experimental design and conditions, as well as descriptions for experiments and phenotypes (see section 2.1.10 and Appendix A1.3 for the new tables within the database scheme of PhenomicDB version 2.x).

Table 3:

Examples of a controlled vocabulary of RNAi assays added to PhenomicDB version 2.x

Assay Name	Assay Type
Mitotic index	proliferation
BrdU incorporation	proliferation
Metabolic activity	metabolism
Cell number	apoptosis
Apoptosis TUNEL	apoptosis
Alamar blue	viability
Centrosome duplication	morphology
Cell cycle affected	cell cycle
Cytokine creation	immunology

With this, PhenomicDB was redesigned to accept large datasets of over 250,000 RNAi phenotypes in total (now making up roughly two thirds of the phenotypes in PhenomicDB version 2.x, see sections 2.1.10 and 4.2.4 for details on these data sources). Finally, the controlled vocabularies for RNAi screens (from the scientists at Bayer Schering Pharma AG, see Table 3), phenotypes (from MP, see sections 1.3.2 and 2.1.5 for details) and cell cultures (from the ATCC, Table 4) ere integrated and associated with these RNAi phenotypes where possible.

Table 4:

List of cell lines from the online catalogue of ATCC (http://www.atcc.org), some of which were added as controlled vocabulary in PhenomicDB version 2.x for the appropriate RNAi screens.

Tumor Cell Line			
ATCC No.	**Name**	**Cancer Type**	**Tissue Source**
CCL-256	NCI-H2126	carcinoma; non-small cell lung cancer	lung
CRL-5868	NCI-H1395	adenocarcinoma	lung
CRL-5872	NCI-H1437	adenocarcinoma	lung
CRL-5911	NCI-H2009	adenocarcinoma	lung
CRL-5985	NCI-H2122	adenocarcinoma	pleural effusion
CRL-5922	NCI-H2087	adenocarcinoma	lymph node (metastasis)
CRL-5886	NCI-H1672	carcinoma; classic small cell lung cancer	lung
CRL-5929	NCI-H2171	carcinoma; small cell lung cancer	lung
CRL-5931	NCI-H2195	carcinoma; small cell lung cancer	lung
CRL-5858	NCI-H1184	carcinoma; small cell lung cancer	lymph node (metastasis)
HTB-172	NCI-H209	carcinoma; small cell lung cancer	bone marrow (metastasis)
CRL-5983	NCI-H2107	carcinoma; small cell lung cancer	bone marrow (metastasis)
HTB-120	NCI-H128	carcinoma; small cell lung cancer	pleural effusion
CRL-5915	NCI-H2052	mesothelioma	pleural effusion
CRL-5893	NCI-H1770	neuroendocrine carcinoma	lymph node (metastasis)
HTB-126	Hs 578T	ductal carcinoma	mammary gland; breast
CRL-2320	HCC1008	ductal carcinoma	mammary gland; breast

After its November 2008 data update, PhenomicDB hosted 280,533 phenotypes, connected to 77,400 eukaryotic genes. All data entries are cross-referenced by links to their original data sources. It is kept up-to-date on a regular schedule and is freely accessible without restrictions. By the end of 2006 (data from Groth and Weiss [Groth and Weiss, 2006b]), the percentage of the Entrez gene index with a phenotype was approximately 99% for *Drosophila melanogaster*, 79% for *Caenorhabditis elegans*, 21% for *Saccharomyces cerevisiae*, approximately 16% for *Mus musculus* (this number was estimated on the basis of the human Entrez gene entries, as Entrez gene index for mouse (62,907 gene IDs) was still in progress and therefore had not collapsed then) and 8% for *Homo sapiens*. 84% of all available phenotypes in PhenomicDB came from *Drosophila melanogaster* and *Caenorhabditis elegans*. 16.2% of phenotypes were associated with a gene having no orthologs, and less than 1.5% of the phenotypes were associated to a gene that could not be mapped to the Entrez gene index. 40,299 eukaryotic orthology groups were registered and a third of them (13,695) had at least one phenotype in any of the species. For *Homo sapiens*, 2,850 genes were linked to 4,009 phenotypes and for another 7,592 human genes there was at least one 'orthologous phenotype' available. Thus, the percentage of human genes with phenotypic information was raised from 8% of the Entrez gene index (without orthologous informa-

tion) to 31% with orthologous information. For *Mus musculus*, 'orthologous phenotypes' increased available phenotypic information for mouse genes to over 30% of the gene index (see Figure 15 for more details, also on other species). These figures clearly show that integrating disparate phenotype data from different species can generate new information with worthwhile implications for each of these species.

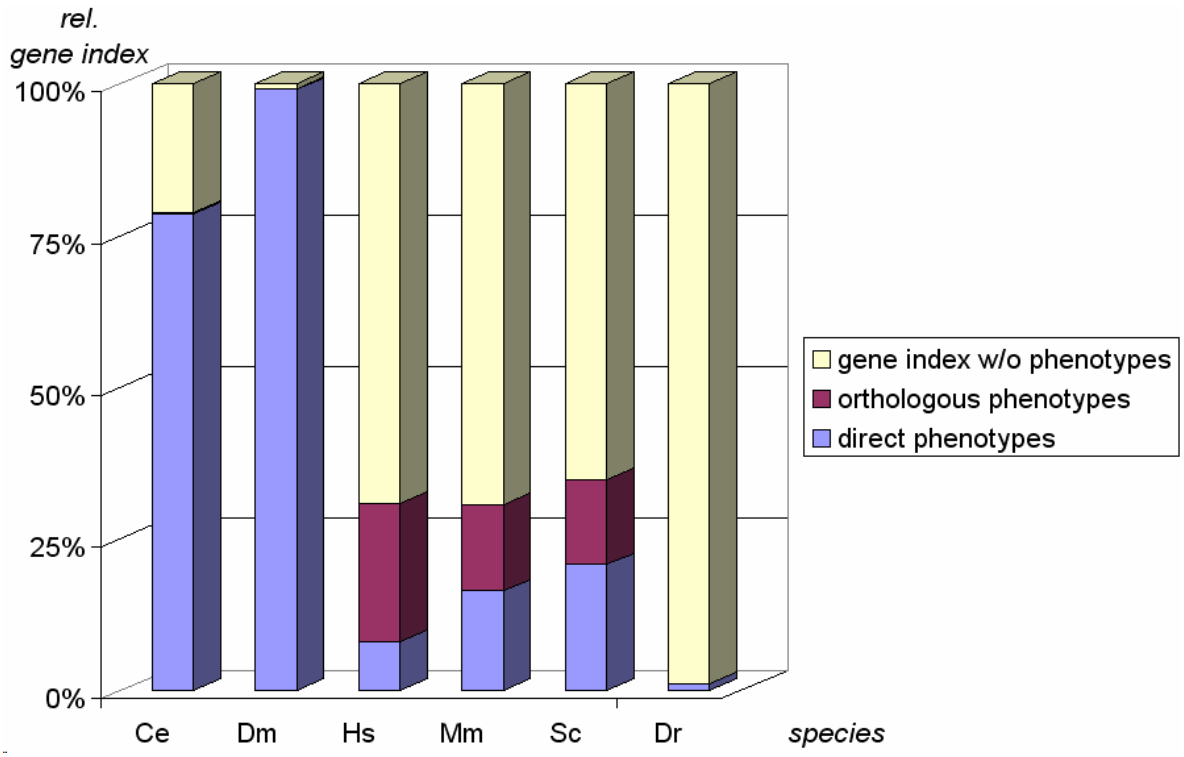


Figure 15:

Percentage of NCBI Entrez gene indices with phenotypic information in PhenomicDB for 5 model organisms and *Homo sapiens*. The percentage of genes with one or more phenotype from the given species is shown in blue ('direct phenotypes'), of genes with one or more phenotype associated only by orthology are shown in red ('orthologous phenotypes'), and of those genes that have no phenotype associated are shown in yellow. The red bars show the direct benefit from cross-species integration in PhenomicDB. The high coverage of *Caenorhabditis elegans* and *Drosophila melanogaster* gene indices with phenotypic information is mainly due to integrated RNA interference data (data and figure taken from [Groth and Weiss, 2006b]).

Ce = *Caenorhabditis elegans*; **Dm** = *Drosophila melanogaster*; **Hs** = *Homo sapiens*; **Mm** = *Mus musculus*; **Sc** = *Saccharomyces cerevisiae*; **Dr** = *Danio rerio*

3.1.4 Extensions to the user interface

In PhenomicDB, genotype and phenotype data have been organised in a single database scheme. Having all genes annotated and also indexed over orthology groups allows searching orthologous genotype and phenotype data with a single database query. The advent of RNAi data required the scheme to be extended (as described in sections 2.1.10 and 3.1.3) in order to cope with a 'qualitative' phenotype, e.g. the description of a visual inspection via microscopy, but also with a 'quantitative' phenotype, i.e. a floating point number expressing an absolute or relative deviation from an expected 'normal' or average phenotype. The update in the database scheme also required improvements in the interface (see section 2.1.10). This interface can now show these details, as well as the important aspects of RNAi study design that have been specially addressed in the new database scheme, e.g. assay, cell line, time point, mRNA knock-down efficiency, phenotype penetrance, siRNA sequence, etc. have been addressed adequately. Furthermore, PhenomicDB was enriched with tables holding MGI's Mammalian Phenotype ontology and controlled vocabulary for cell lines and RNAi assays.

PhenomicDB's graphical user interface has been designed to be as simple and as effective as possible. A basic query can be started intuitively by entering any search term (e.g. apoptosis, BUB1) or identifier (e.g. NM_001211). Users can configure the output data fields to be shown individually, e.g. gene symbol, phenotype name, ontology, chromosomal localization, etc. Queries allow wildcards and logical operators ('AND', 'NOT', 'OR') and can be further refined by limiting to data fields, data domains or organisms.

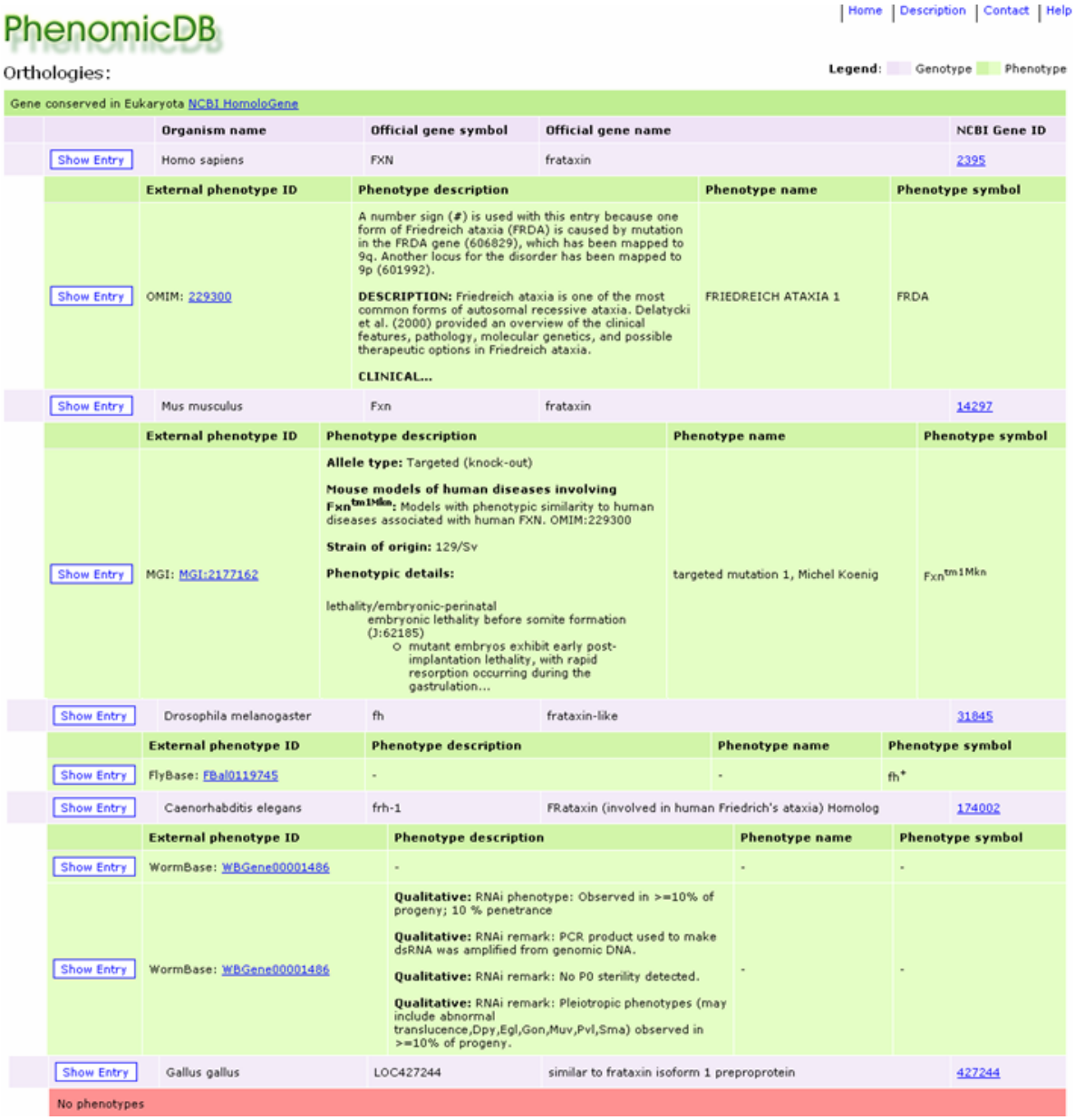


Figure 16:

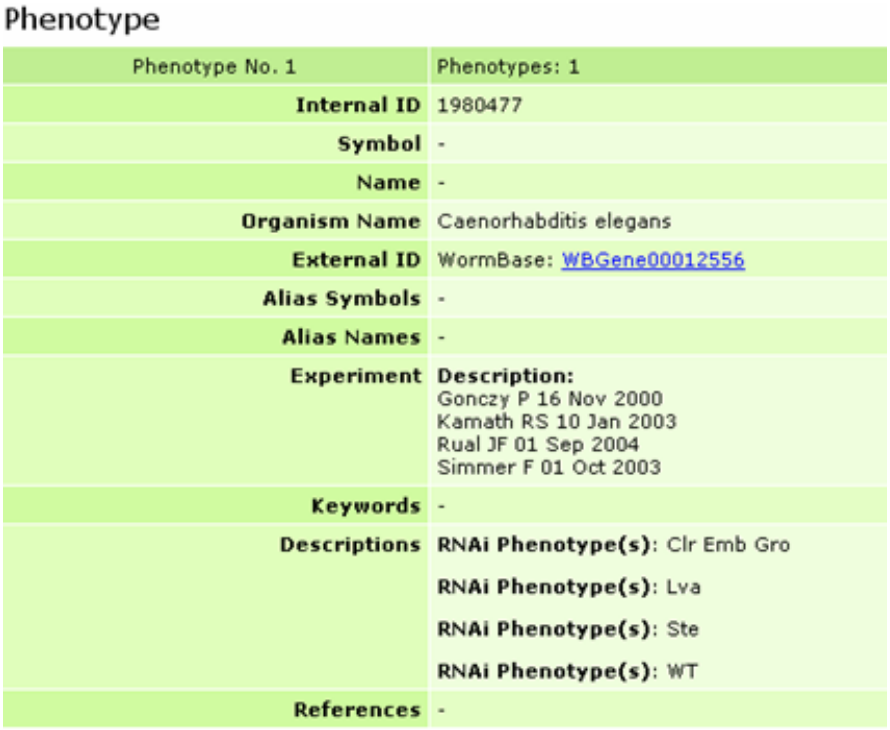
Typical result list, showing here the search for the *frataxin* orthology group (some entries omitted for simplicity). The *frataxin* genes from different species are shown with a marble background; indented and with green background, the corresponding phenotypes are shown. Hyperlinks lead to the source database. The 'Show Entry' button displays the full genotype/phenotype information. For *Gallus gallus*, no phenotype (with red background) is available.

The customizable results interface (see Figure 16) lists all hits organised by genes with their associated phenotypes indented and provides further links to more detailed views. The button 'Orthologies' enables the user to show all orthologous genes with their associated phenotypes for each hit. The 'Show entry' buttons lead to the full-length genotype or phe-

notype entries. Also, the entire hit list can be expanded to show the orthologs of all or selected genes as well as their corresponding phenotypes. All entries consistently link back to their original sources (e.g. entries derived from OMIM link back to OMIM) to make sure data will be properly referenced by users.

For convenient external access to PhenomicDB, static hyperlinks can be created to direct to any genotype or phenotype using e.g. the Entrez gene ID. Dynamic URLs using any query term behave as if the term was entered into the search mask of the homepage. A manual is available on the homepage. External linking to PhenomicDB is also featured in the browser task bar *BioBar* (http://biobar.mozdev.org).

Like other phenotype data, RNAi data was historically stored in a plain text field in the old version of PhenomicDB (see Figure 17). This changed with PhenomicDB version 2.x, where the additional phenotype information (see section 2.1.10 and 4.2.4 for details) is stored with the quantitative and qualitative RNAi phenotype data (see Figure 18).



Phenotype

Phenotype No. 1	Phenotypes: 1
Internal ID	1980477
Symbol	-
Name	-
Organism Name	Caenorhabditis elegans
External ID	WormBase: WBGene00012556
Alias Symbols	-
Alias Names	-
Experiment	**Description:** Gonczy P 16 Nov 2000 Kamath RS 10 Jan 2003 Rual JF 01 Sep 2004 Simmer F 01 Oct 2003
Keywords	-
Descriptions	**RNAi Phenotype(s)**: Clr Emb Gro **RNAi Phenotype(s)**: Lva **RNAi Phenotype(s)**: Ste **RNAi Phenotype(s)**: WT
References	-

Figure 17:

Entry for an RNAi phenotype from WormBase in the old PhenomicDB version 1.x. It can be seen that in comparison to Figure 18, the phenotype description was very sparse, with very short descriptions and domain-specific vocabulary.

Phenotype

Phenotype No. 3	Phenotypes: 1 2 3 4 5 6 7 show all			
Internal ID	1211503			
Symbol	-			
Name	-			
Organism Name	Homo sapiens			
External ID	ELN: 130650			
Alias Symbols	-			
Alias Names	-			
Descriptions	**Quantitative:** 5.66% +/- 0.97% **Qualitative:** RNAi experiment in HeLa cells yiedling decreased cell number **Keywords:** MP:0000329 (decreased cell number)			
Experiment	**Name:** Nuclei number **Description:** - **Type:** apoptosis/proliferation			
Experimental Conditions	**Type**	**Value**	**Classification**	**Timepoint**
	cell type	HeLa (ATCC: CCL-2)	**Organ:** cervix **Cell type:** epithelial **Disease:** adenocarcinoma	-
	knock-down mRNA level	85.9% / 86.0% / 86.0% (min/mean/max)	-	48h
RNA	**Accession No.:** siRNA00001_1 **External ID:** 213549 **Loop/Overhang:** TT **Sequence:** CCAAGGUUUUCGAUUGCUC			
References	**Submitter:** C.Merz			

Figure 18:

RNAi phenotype data entry in PhenomicDB version 2.x. More detailed information is given not only for the phenotype itself (with a link to MP), but also about other controlled vocabularies for screens and cell-lines.

3.1.5 Discovering knowledge with PhenomicDB

It becomes clear that PhenomicDB in its current version is ready for knowledge discovery, with its functionality to present phenotypes for orthologous genes on one single page. This is a valuable feature, for example in a pharmaceutical target discovery setting. Here, human genes are the actual targets, but there is much more information available for orthologous genes from model organisms and from various experiments or mutations that are at present impossible to obtain in such detail for any human gene. Also, PhenomicDB can be used to gain more insight into the nature of genetic diseases, due to its high-level integration of diverse phenotypes directly associated with responsible genes.

However, one question that could not be answered yet is whether similar phenotypes associated with different genes also yield new biological insights. The next part of this thesis (section 3.2) deals with the results of a study focusing on this question.

3.2 Cross-species phenotype clustering

3.2.1 PhenoDoc clustering: A new approach to group genes

Textual descriptions of phenotypes were obtained with reference to their associated gene from the PhenomicDB database (here: PhenomicDB version 2.1). For text-mining purposes, the descriptions had to be properly filtered, adapted and prepared, resulting also in a data reduction (see section 2.1.6 for details of methods and data). The term phenodoc is used in the following to refer to this adjusted form of phenotype description (see Definition 1). Phenocluster refers to a cluster of phenodocs (see Definition 5). The resulting 39,610 phenodocs associated to 15,426 genes from 7 different species (see section 2.1.6 for more information on the data set) were allocated to 1,000 clusters (and also other cluster sizes, see section 3.2.2.6) based on the cosine similarity between phenodocs using the k-means algorithm on a vectorized representation of the documents. From these clusters, gene groups were assembled as described in section 2.2.5. The resulting groups were studied from a number of perspectives to assess whether or not the grouping itself was biologically reasonable. Then, gene function was predicted within each cluster and evaluated using cross validation. Finally, the TFIDF weighting scheme for the vectorized presentation of the documents was challenged using other weighting schemes and the results were re-evaluated.

Of the 1,000 clusters, 90.4% were single species. Figure 19 shows the distribution of cluster sizes. Figure 20 details the distribution of genes by species (independent of the clustering) and the distribution of species in clusters (depending on the clustering).

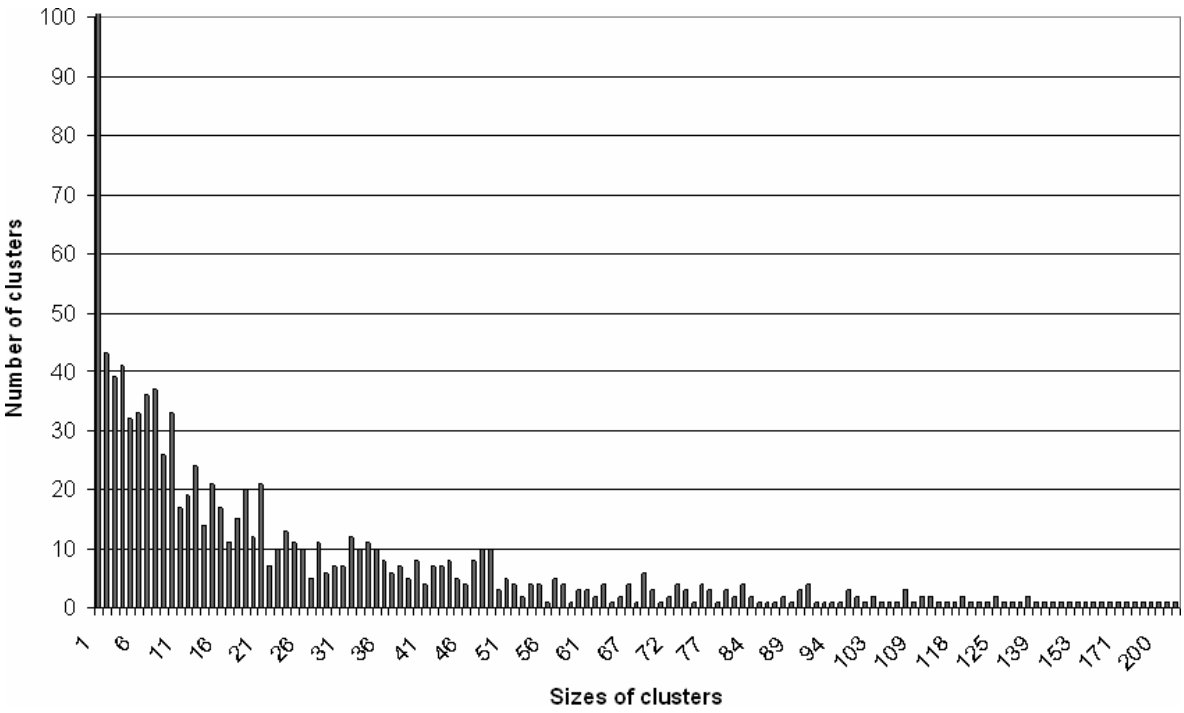


Figure 19:

The diagram shows the distribution of the number of clusters in different sizes.

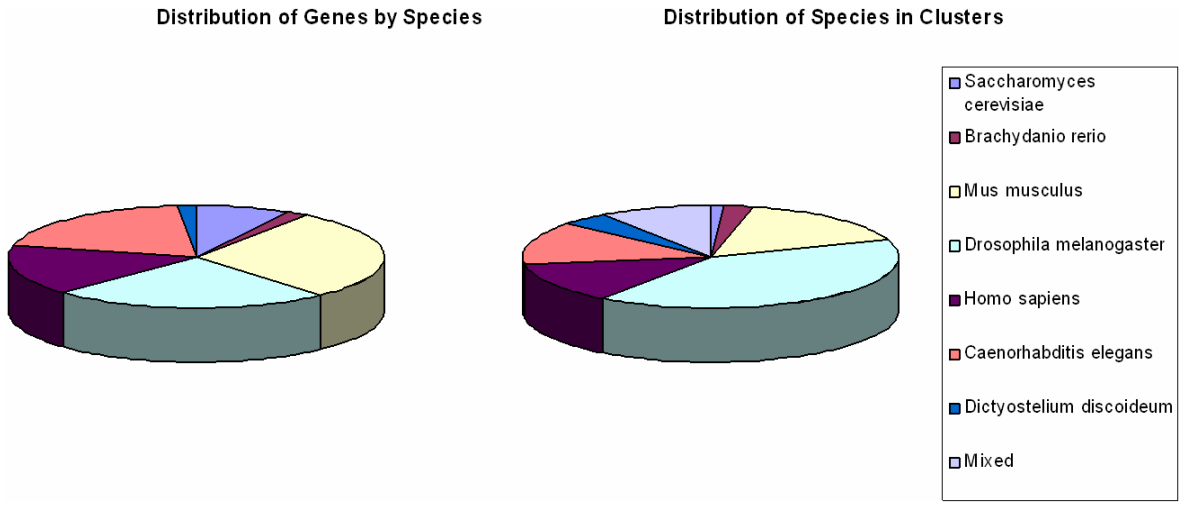


Figure 20:

The left pie chart depicts the distribution of genes by species, i.e. the relative number of genes in the gene set according to species affiliation. The right pie chart shows the distribution of clusters according to single species or 'mixed', if the cluster is made up of genes from more than one species.

3.2.2 PhenoCluster: Gene function prediction from phenotype data

3.2.2.1 Proteins within a 'phenocluster' intensively interact with each other

To test whether phenoclusters consist of genes with a high chance of being part of a common biological process, it was studied whether the proteins encoded by the genes within one cluster interact with each other more often than proteins in random control groups. This approach derives from the observation that physically interacting proteins have a higher chance to be part of the same biological process or pathway than non-interacting proteins [Guo, et al., 2006]. To test whether this is true for interacting proteins, PPi data was downloaded from the BioGrid database represented by pairs of Entrez gene IDs (see section 2.1.3 for details on the data set and see section 3.2.2.1 for an anecdotic inspection of phenoclusters and PPi). The degree of interactions among the members of a given phenocluster was then analyzed and those figures were compared to random gene groups of similar size.

In 60 phenodoc clusters (from 1,000) comprising 1,858 genes, all genes physically interact in a cell with at least 75% of the rest of the genes from the same group within at most two intermediates (empirical p-value smaller than 0.05). Thus, those clusters consist of genes which almost build cliques in the protein-protein-interaction network. Such quasi-cliques previously have been associated to functional modules [Spirin and Mirny, 2003]. In another 138 clusters, comprising a total of 4,322 genes, all genes interact with at least 33% of the rest of the genes in each group. In the mean, these groups have 30 members, which is approximately twice the expected mean size for groups (~ 15,000 genes in 1,000 clusters). However, they are still in the same size range of large 'functional modules' (as shown by e.g. Wu et al. [Wu, et al., 2005]). These numbers were compared to 200 repetitions of randomly sampled control groups. In this control dataset, only one group reached the threshold of 75% and two groups reached the threshold of 33%.

These figures show that clustering of phenodocs results in gene groups whose members much more often interact with each other than expected by chance and thus represent coherent biological knowledge. However, the interaction score of the rest of these clusters is not significantly higher than in the control groups. This difference is exploited in section 3.2.2.4 to sort clusters based on this score to see whether the prediction of function improved in highly interacting clusters.

The large number of those non-interacting clusters is mostly an artefact of the current incompleteness of PPi data sets in BioGrid, with the notable exception of *Saccharomyces cerevisiae*. Therefore, biologically coherent phenoclusters are only partially congruent with PPi networks. Even highly interacting phenoclusters will not necessarily mimic the PPi network due to the diverse nature of phenotypes or the lack of data, both on the PPi and the phenotype side (see findings in section 0).

Figure 21 shows genes from a phenocluster with many connected proteins (blue nodes), including interacting proteins with no phenotype described yet (added *a posteriori* and coloured in red).

Many of the proteins in Figure 21 are involved in the division of germ line stem cells or in the regulation of that process in *Drosophila melanogaster*. *Cyclin B* (blue node with Entrez gene ID 37618) is required for the division of germ line cells [Wang and Lin, 2005]. This protein associates with spindle microtubules throughout meiosis I and meiosis II [Swan and Schupbach, 2007]. *Profilin* (blue node with Entrez gene ID 33834) and *roughex* (blue node with Entrez gene ID 31535) both act independently as suppressors of *cyclin B* throughout the cell cycle [Foley, et al., 1999; Ji, et al., 2002]. *BetaTub85D* (blue node with Entrez gene ID 41124) forms microtubules and is a structural constituent of the cytoskeleton and the spindle [Goldstein and Gunawardena, 2000]. *NCD* (blue node with Entrez gene ID 43517) is responsible for the generation of sliding forces between adjacent microtubules and plays a role in spindle morphogenesis [Oladipo, et al., 2007]. Even though spindle assembly occurs in *NCD* mutants, its movement along microtubules is needed to stabilize interactions between chromosomes [Skold, et al., 2005]. *Subito* (blue node with Entrez gene ID 44870) is required for establishing and maintaining the meiotic spindle pole formation in oocytes [Giunta, et al., 2002; Jang, et al., 2005]. *Centrosomin* (blue node with Entrez gene ID 36491) targets and anchors gamma-tubulin to the centrosome and organizes microtubule-nucleating sites [Terada, et al., 2003]. One of the proteins with no known phenotype (red node) directly interacting with *centrosomin* is *sina* (Entrez gene ID 39884), which forms an ubiquitin ligase complex and is involved in ubiquitin-dependent protein catabolic process [Carthew and Rubin, 1990]. It may also be involved in regulating the levels of developmentally important transcription factors [Cooper, et al., 2008]. *Effete* (blue node with Entrez gene ID 41785), a ubiquitin-conjugating enzyme, acts as a suppressor of *sina* [Ryoo, et al., 2002], is involved in chromosome organization and meiosis [Cenci, et al., 1997], as well as regulation of R7 cell differentiation and protein ubiquitina-

tion [Bergmann, et al., 2003]. Interacting with *BetaTub85D* and *NCD*, the protein with no known phenotype (red node) is *sina homologue* (*sinaH*, Entrez gene ID 39885). It has been found that flies lacking *sinaH* are viable and that there is no redundancy in the function with *sina* [Cooper, 2007]. However, not much is known about this protein and it may still play an important regulatory role similar to that of *sina*. There are two further proteins for which no phenotype is known, but which interact with and actually connect two proteins from the same phenocluster. One is *brp* (Entrez gene ID 35977) connecting *cyclin B* with *BetaTub85D* and the other is *CG15631* (Entrez gene ID 33675) connecting *subito* and *effete*. This is evidence that their phenotype will eventually turn out to be similar to those of the proteins described above and that these proteins will be involved in the same processes.

There is a pair of proteins in Figure 21 that are not connected to the main network, three more which are not connected at all. Still, according to their phenocluster they show a similar phenotype. *Mei-P26* (blue node with Entrez gene ID 45775) restricts growth and proliferation in the ovarian stem cell lineage [Neumuller, et al., 2008] and is involved in meiosis and germ cell development [Page, et al., 2000]. *Meiotic* (blue node with Entrez gene ID 247151) is a protein with unknown molecular function, no known biological process involvements and no annotated transcripts. The observed mutant phenotype, however, is defective in its meiotic cell cycle [Ivy, 1981; Orr-Weaver, 1995]. *Fireworks* (blue node with Entrez gene ID 45977), is a protein with unknown molecular function, no known biological process involvements and no annotated transcripts. The observed mutant phenotype is meiotic cell cycle defective, manifesting in neuroblast and larva [Dean, et al., 2001]. *Asp* (blue node with Entrez gene ID 42946) plays a role in spindle pole organization [Morales-Mulia and Scholey, 2005]. Its mutation leads to severe defects in spindle microtubule within the germarium [Riparbelli, et al., 2004]. *Asterless* (blue node with Entrez gene ID 250471) is involved in the following biological processes: centrosome cycle; centrosome organization, and biogenesis. Furthermore, it is associated with defective meiotic cell cycle, manifesting in the larval brain, neuroblasts and spindle [Bonaccorsi, et al., 2007; Rusan and Peifer, 2007; Varmark, et al., 2007].

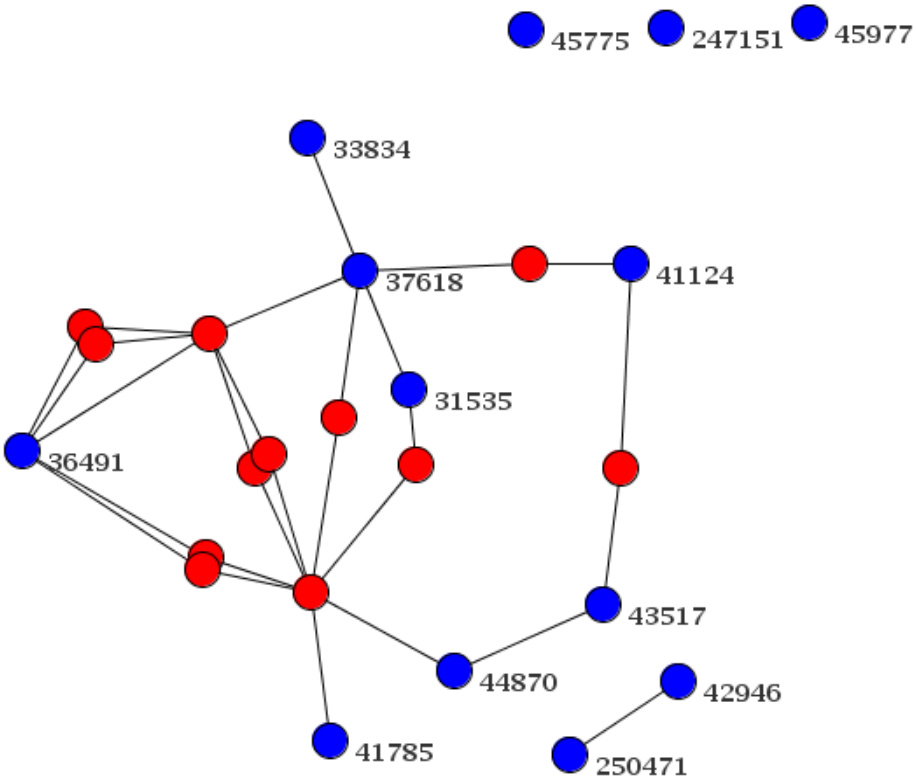


Figure 21:
The figure shows an example of interactions between proteins from genes in a single phenocluster. Depicted is a network with many genes from the same phenocluster (blue nodes with Entrez gene IDs) for which associated proteins are connected, while the genes of all proteins that are responsible for these connections are not in the initial set of genes due to lack of substantial phenotype data (red nodes).

There is a lot of biological evidence to show that this phenocluster indeed comprises a group of proteins acting in concert in closely related processes and functions. It is therefore very likely that for a growing number of connected proteins with unknown functions and no known phenotypes, information will soon be available with even more substance to this biologically coherent phenocluster.

In contrast to the phenocluster described above, there are other phenoclusters with seemingly unrelated members. In Figure 22, the clustered blue nodes are again supplemented by nodes from the PPi data. Here, connected proteins with available phenotype data (green nodes) that have not clustered into this phenocluster were added *a posteriori*. There also is a single unconnected node in this phenocluster. The proteins within this phenocluster are mostly involved in the regulation of the Hedgehog signaling pathway. This pathway is

mostly responsible for embryonic development in general and wing development in *Drosophila melanogaster* in particular [Cohen, 2003]. *Ci* (blue node with Entrez gene ID 43767) is an essential part of the hedgehog signaling complex [Bijlsma, et al., 2004; Stegman, et al., 2000]. Then there is the protein *hedgehog* itself which is an essential regulator of the pathway (blue node with Entrez gene ID 42737). Extracellular binding of *hedgehog* inhibits the proteolytic cleavage of *Ci* [Apionishev, et al., 2005]. *Tsc1* (blue node with Entrez gene ID 42862) is a (negative) regulator of cell proliferation, growth, size and cycle [Johnston and Gallant, 2002; Manning and Cantley, 2003; Potter, et al., 2002; Tapon, et al., 2001]. *Cos2* (blue node with Entrez gene ID 35653) plays an important role in regulating the amounts and activity of *Ci*, thus regulating the level of *hedgehog* to which cells are exposed [Farzan, et al., 2008]. The protein *fused* (blue node with Entrez gene ID 32855) induces activation of *hedgehog* targets [Ruel, et al., 2007] and is in fact vitally important for hedgehog signaling, as the signaling pathway requires an intramolecular association between two domains of *fused* [Ascano and Robbins, 2004].

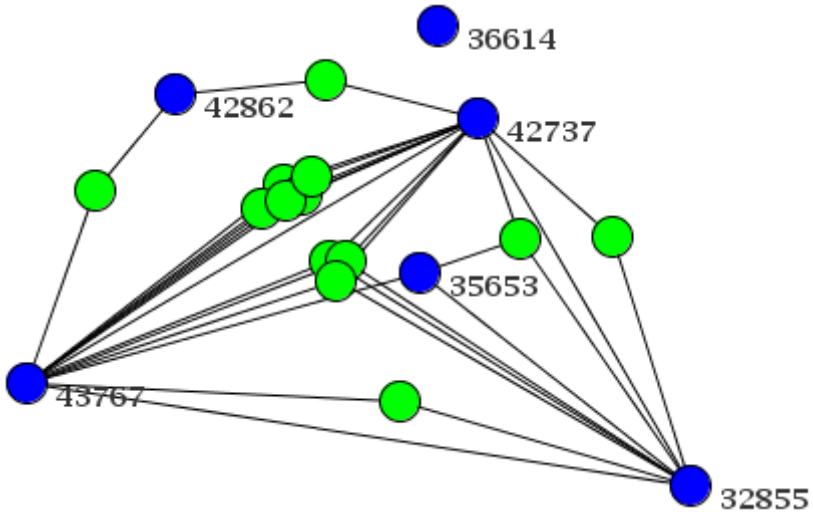


Figure 22:

Interactions between proteins encoded by genes from more than one phenocluster. Depicted is a network with 6 genes from the same phenocluster (blue nodes with Entrez gene IDs) for which encoded proteins are interacting (connected by lines), while other genes with known phenotypes are also shown, encoding for more interacting proteins not found in the same phenocluster (shown as green nodes).

This biologically coherent account of this phenocluster is unsettled by the fact that there are many more interacting proteins with known phenotypes that are not within this phenocluster. A good explanation for this is probably that the hedgehog signaling pathway is highly versatile in *Drosophila melanogaster*. For example, the protein *hedgehog* itself has 78 interactors and 69 GO-annotations according to BioGrid [TheBioGrid, 2009]. Nevertheless, phenoclusters give insight into the structure of biological networks and can be used to gain novel biological insights as it is the case for the only unconnected node in Figure 22 (blue node with Entrez gene ID 36614). This protein named *DEXT1* is part of the Wnt signaling pathway [Bornemann, et al., 2004; Han, et al., 2004]. There is growing evidence that these two signaling pathways are commonly regulated, a fact which is now also evidenced by the membership of *DEXT1* in this phenocluster [Bornemann, et al., 2004; Kalderon, 2002; Nusse, 2003].

3.2.2.2 Genes in phenoclusters have coherent GO-annotations

Another systematic way to determine biological coherence of phenoclusters lies in computing the similarity of the GO-terms assigned to the genes of a group (see section 1.2.5 for more information on GO and section 2.2.3 on the methods for calculation and an interpretation of the following similarity scores). It should be noted here that in PhenomicDB, GO-terms are associated to the gene descriptions and are not part of phenodocs (unless by rare coincidence, i.e. when authors had used GO-terms in the free-text descriptions that may also occur in GO). In the analysis of the 1,000 phenoclusters, 237 groups were found containing 1,957 genes with a mean GO-similarity score ≥ 0.4 (see Figure 23 for the distribution of mean GO-similarity for all phenoclusters). For each distinct group size, 200 control groups were formed from randomly picked genes. Only two control groups reached this threshold by chance. The Pearson correlation coefficient r (see section 2.2.6 for details) calculated between the mean GO-similarity with the mean phenotype similarity of clusters was 0.41, indicating a shared variance in both similarity scores, approximately 16% higher than expected by chance.

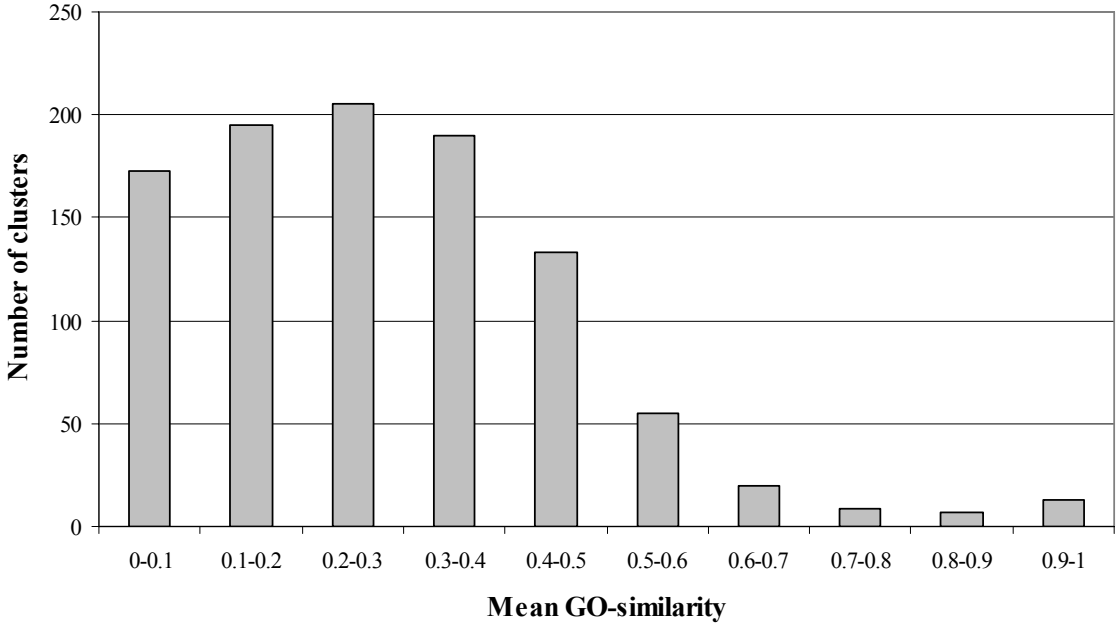


Figure 23:
Distribution of 1,000 clusters with their mean GO-similarity.

This shows that phenotype similarity is indicative for a high probability to share GO-annotations between the associated genes. In Table 5, an exemplary cluster with a GO-similarity score of 0.9 is shown (randomly chosen from one of the 13 the clusters with a GO-similarity score between 0.9 and 1.0). From all of the terms associated with this group, 5 terms are commonly annotated to 14 out of 17 genes (≥ 75% of genes in the group). Due to the homogeneous nature of the annotations, one can hypothesize that the remaining 3 genes (*eif-3.G*, *rrt1*, and *Y37B11A.3*) should play similar biological roles and hence, annotations can be predicted. This idea will be picked up later in the prediction of GO-terms in phenoclusters (see section 3.2.2.4).

Table 5:

Phenocluster with 17 associated genes with a GO-score of 0.9 in the biological process sub-ontology.

From all of the terms associated with this group, 5 terms are commonly annotated to 14 out of 17 genes. Due to the homogeneous nature of the annotations, one can hypothesize that the remaining 3 genes should play similar biological roles and hence, annotations can be predicted.

Entrez ID	Gene symbol	Gene name	# annotated GO-process terms	# terms common to ≥ 50% of genes in group	# terms common to ≥ 75% of genes in group
172805	rps-19	Ribosomal Protein, Small subunit 19	5	5	5
174346	eif-3.G	Eukaryotic Initiation Factor	7	4	4
175501	rpl-3	Ribosomal Protein, Large subunit 3	6	6	5
175538	lrs-1	Leucyl tRNA Synthetase	14	6	5
175584	rps-19	Ribosomal Protein, Small subunit 1	7	6	5
175659	rrt1	aRginyl aa-tRNA syntheTase	8	4	4
175796	rpl-23	Ribosomal Protein, Large subunit 23	8	6	5
175901	rps-13	Ribosomal Protein, Small subunit 13	5	5	5
176007	rpl-36	Ribosomal Protein, Large subunit 36	6	6	5
176011	rps-21	Ribosomal Protein, Small subunit 21	6	6	5
176024	prs-1	Prolyl tRNA Synthetase	9	6	5
176071	rpl-9	Ribosomal Protein, Large subunit 9	7	6	5
176097	rpl-35	Ribosomal Protein, Large subunit 35	5	5	5
176146	rpl-21	Ribosomal Protein, Large subunit 21	5	5	5
177583	rps-21	Ribosomal Protein, Small subunit 2	5	5	5
179063	W02F12.5	W02F12.5	8	5	5
189611	Y37B11A.3	Y37B11A.3	2	1	1

3.2.2.3 Phenocopies co-occur in phenoclusters

If phenoclusters properly reflect phenotype similarity on a biological basis, the genes causing phenocopies (see section 2.1.8 for details) should co-occur within the same clusters. Of the 27 phenocopies induced by 57 genes that were retrieved from literature, 25 phenocopies (55 genes) were found in the data set. In 1,000 phenoclusters, the genes of 13 phenocopies (54.2%) co-occurred in a cluster. In 1,000 random clusters of the same size none of those genes co-occurred in any cluster.

3.2.2.4 Predicting gene function within phenoclusters

The previous results lead to the hypothesis that gene function can be predicted on the basis of association of genes to phenoclusters. If gene groups based on phenoclusters have a coherent GO-annotation, it should be possible to predict similar functions for genes from the same cluster that are not or only partially annotated. The following sections present values for precision and recall of GO-term predictions for different subgroups of genes from the phenodocs. These ‘predictions’ show the percentage of overlaps between the true annota-

tions of a set of genes (test set) and 'predicted' terms which are derived from a training set (according to the Entrez Gene2GO annotations – see sections 2.1.1 and 2.1.4 and 2.2.7 for further details).

In evaluating the correctness of a GO-annotation prediction, one has to consider the structure of GO. Recall that GO-terms form an ontology, and that terms are connected by IS-A and PART-OF relationships. The simplest case would be to consider a prediction as correct only when it appears exactly as it is in the test data. However, this criterion is too strict, since terms which are a little more general or more specific can equally contribute to a biological annotation. In the following, results are therefore given for different definitions of 'correctness' of a prediction. In the most stringent case, a term is considered correctly predicted only if it appears in both test and training set. Thus, predicting a child of a term actually counts negative twice – as a false positive and a false negative. Because this measure is much stricter than that of other studies (see for instance [Kelley, et al., 2003]), it is also shown how the figures change when the criterion for 'equality' of GO-terms is relaxed.

Table 6:

Different criteria for filtering clusters for function prediction.

In order to push the values for precision and recall towards the precision ceiling, filter criteria for selecting appropriate gene groups *a priori* were tested. The following filter criteria were defined.

Filter	Effect
Filter 1	Removes groups with fewer than 3 genes or no GO-terms associated to at least 50% of genes.
Filter 2	Removes groups with a GO-similarity score < 0.4.
Filter 3	Removes groups with a PPi-connectedness < 33%.
Filter 4	Removes all non-single species clusters.
Filter 5	Removes all single-species clusters.

Table 7:

Results for different filters applied to gene groups (k=1,000).

Precision and recall values of function prediction in all clusters and with varying k selected by different combinations of the filters defined in Table 6.

	Filter 1	Filter 1 & Filter 2	Filter 1 & Filter 3	Filter 1 & Filter 4	Filter 1 & Filter 5
# of groups	196	74	53	185	11
# predicted terms	345	159	102	338	16
# of genes	3,213	711	409	2,895	320
Precision	67.91%	62.52%	60.52%	67.73%	64.70%
Recall	22.98%	26.16%	19.78%	23.80%	11.21%

To explore the upper limit of 'predictability' (the so-called 'precision ceiling') of GO-terms based on phenotype clustering with this method, function prediction was performed for all gene groups based on the clustering of the phenodocs. For each group, precision and recall of the predictions were computed (see section 2.2.8 for methods). The groups were filtered to leave the 10% highest-scoring groups sorted by the harmonic mean of recall and precision (so-called F-measure). Thus, clusters were selected *a posteriori* based on their performance in prediction. Of course, this measure cannot be extrapolated to the result of a prediction for unknown groups; however, it gives a good estimate of the maximum performance achievable using this data set and this approach. Function prediction from only the highest scoring groups yielded a mean precision of 81.5% and a mean recall of 61.2%.

Considering this as upper limit, the goal in the following was to find criteria for selecting appropriate gene groups *a priori*.

To this end, five filters were defined for selecting clusters. These filters were based on criteria such as the number of genes they contain, the number of available annotations, and their scores for in-group functional coherence and in-group PPi-connectedness (see Table 6). Precision and recall of function prediction in all clusters selected by various combinations of those filters were then calculated. Results are summarized in Table 7 (refer to section 4.3.2.3 for details of filters and evaluation). Using the least stringent filter (filter 1), but the strict criterion for judging the identity of GO-terms, the initial number of 1,000 clusters was reduced to 856 by filtering all clusters containing fewer than 3 genes and was reduced once more to 295 by filtering all clusters without any descriptive GO-terms (i.e. any biological process terms assigned to at least 50% of cluster members). The prediction gave 345 distinct GO-terms from the biological process subtree at a mean precision of 67.9% and a mean recall of 23.0% over all selected clusters.

Relaxing the criteria for GO-term identity, now allowing for a single deviation towards the root (i.e. a predicted term is considered correct if it exactly matches a removed term or if it matches a parent of the removed term) resulted in a mean precision of 75.6% and a mean recall of 28.7% (191 unique terms for 2,686 genes in 279 groups). Allowing one more step towards the root, 151 unique terms could be predicted with 76.3% mean precision and 30.7% mean recall. The number of correctly predicted terms decreases here as an effect of the collation, e.g. when any of two child terms and a parent term is counted only once (as parent term).

Using function prediction on only those clusters passing filter 1 and showing a mean GO-similarity ≥ 0.4 (filter 2), the mean precision dropped slightly to 62.5% and recall increased to 26.2% (74 groups, 711 genes and 159 predicted distinct GO-process terms). This drop in precision and increase in recall is due to the increasing number of predictions made per gene and group, and is explained in more detail in the following sections. Applying again a less stringent criterion for identity of GO-terms as explained above, mean precision and recall were 75.3% and 31.7% respectively in the first step towards the root (91 unique terms for 612 genes in 80 groups). When only those clusters containing genes from only one species were selected (applying filter 4), the values for precision and recall stayed roughly the same. This was expected as 90% of all clusters met this condition (see the discussion in section 4.3.2.2). The values for precision fell slightly and values for recall dropped quite dramatically when only cross-species clusters were used (filter 5).

Surprisingly, when only those clusters with a PPi-connectivity of at least 33% (filter 3) were used, mean precision and recall dropped (to 60.5% and 19.8% respectively; 53 groups, 409 genes and 102 GO-terms). In a recent study, Schwikowski et al. report that 35% of interactions occur between proteins with no common functional annotation [Schwikowski, et al., 2000]. Lack of common functional annotations in relatively small groups of immediate neighbours in PPi-networks can explain the surprising drop in precision and recall when using only these groups. Nevertheless, both enrichment in pair-wise interactions and common GO-terms show the high biological coherence of phenoclusters. It can be concluded that despite some shortcomings in the data, phenoclusters appear to be a suitable source for functional annotation prediction (see also the discussion in section 4.3.3).

3.2.2.5 Selecting gene groups from PPi-cliques

To compare the prediction method using phenoclusters with another non-random gene selection method, the 13,068 initial genes based on direct pair-wise interaction were grouped, resulting in 2,875 groups in which each gene interacts with each other (i.e. cliques in the PPi graph). Applying filter 1 to this data set, 720 groups were derived resulting in 3,692 predictions with a precision of 56.4% and 32.3% recall. Thus, the precision of this approach (which is similar to the method applied by Spirin and Mirny [Spirin and Mirny, 2003]) was about 10-20% less precise than the method of clustering genes based on phenodoc similarity.

3.2.2.6 Clustering phenotypes with different values of k

As an internal measure for cluster quality (see section 4.3.2.3 for a discussion on cluster quality assessment), different values for k, ranging from 500 to 3,000, were chosen (see Table 8). There are a number of interesting results. Firstly, the mean number of genes per cluster clearly decreases with increasing k. However, the percentage of clusters that comply with filter 1 in Table 6 stays roughly the same. Although in the mean, those clusters contained fewer genes, the number of predicted annotations and affected genes increased considerably with increasing k. This indicates that the top clusters – selected by filter 1 – become more homogeneous (in terms of functional annotation) with increasing k, as more clusters have more terms which are annotated to more than 50% of their members. Partly, this is also a statistical effect of the decreasing cluster sizes in which their phenotypes lead to more homogeneous groups. At the same time, the precision drops slightly with increasing k while recall increases considerably. This means that more predictions come with more errors, but the ratio of errors to the overall number of predictions decreases. Another effect is that in smaller clusters there is usually only a single gene left in the test set. The increasing recall shows that more terms from the test set are descriptive in the training set, but the decreasing precision means that the number of terms associated with a single gene cannot compensate for the number of suggestions derived from the training set.

While the correlation between GO-similarity and phenotype similarity drops significantly for increasing k, the percentage of single-species clusters increases. This is an indication that within clusters mentioned above, the shift from a functional homogeneity to a mere methodical (i.e. a descriptive) homogeneity is due to the fact that similar vocabulary – from the same species – yields less variance than similar function. This is also indicated by the sharp drop in the number of phenocopies found in the same cluster, which are also dependent on functional clustering. This effect may be amplified by a statistical effect that smaller cluster sizes yield less probability that the phenocopies end up in the same cluster.

Thus, k is an important parameter to balance the trade-offs between precision, recall and number of predictions. One can either choose a small k-value, resulting in few high quality predictions, or a larger k-value, resulting in a much larger number of less accurate predictions. Clearly, the choice of the k-value depends on the concrete application. In a large-scale functional prediction approach, it is more desirable to make a few very good predictions than to make many predictions with average quality (especially from the curator's point of view). Therefore, the goal was to achieve the best precision with acceptable recall.

This goal was best reached with k=1,000, even though a large k (k=3,000) results in many small clusters as the best technical solution with an F-measure (calculated as the harmonic mean of recall and precision values) of 0.385 (precision = 60.3% and recall = 28.3%). The choice for k is discussed further in section 4.3.2.3.

Table 8:

The distribution of clusters with their characteristics given different values for k (the number of clusters) from 500 to 3,000. Here, filter 1 has been applied for GO-predictions. For details, see text (the F-measures is calculated as the harmonic mean of recall and precision values).

k	500	1,000	2,000	3,000
# single species cluster	422 (84.4%)	904 (90.4%)	1,897 (94.9%)	2,894 (96.5%)
# of phenocopy pairs (of 25)	25 (100%)	13 (52%)	12 (48%)	8 (32%)
# cluster w/ PT-sim ≥ 0.4	92 (18.4%)	293 (29.3%)	526 (26.3%)	810 (40.5%)
# genes	3,221	5,886	6,379	6,878
# clusters w/ GO-sim ≥ 0.4	51 (10.2%)	206 (20.6%)	522 (26.1%)	921 (46.1%)
Correlation GO-sim vs PT-sim	0.53	0.41	0.37	0.28
# genes	863	1,800	2,392	3,065
# clusters w/ PPi ≥ 75%	21 (4.2%)	60 (6.0%)	174 (8.7%)	305 (10.2%)
# genes	1,497	1,858	2,335	2,702
# cluster w/ PPi ≥ 33%	63 (12.6%)	138 (13.8%)	286 (14.3%)	413 (13.8%)
# genes	3,890	4,322	4,965	4,996
# clusters for GO-predictions	90 (18%)	196 (19.6%)	393 (19.7%)	611 (20.4%)
# genes	2,820	3,213	4,145	4,546
# predicted terms	142	345	730	1,226
Precision	72.55%	67.91%	63.40%	60.31%
Recall	16.73%	22.98%	25.63%	28.32%
F-measure	0.2719	0.3434	0.365	0.3854
Mean genes/cluster	54	29	16	11

3.2.3 Using cross-species phenotype vocabulary for clustering

3.2.3.1 Many ontology terms can be matched to phenodocs

Another highly interesting measure with a great impact on the result of phenotype clustering is the structure of the feature vector of phenodocs. As described in sections 2.2.1 and 2.2.2, the TFIDF measure was used to weight each feature in feature space. In section 2.2.1, the possibility was also mentioned to 'over-weight' certain features, e.g. those features believed to have a special impact on the phenotypes' descriptions, like terms from a specific disease vocabulary or phenotype ontology. Terms and their synonyms from GO, MP and MeSH (see section 2.1.7 for details on these data) were matched to phenotype de-

scriptions (from PhenomicDB version 2.4 – see 2.1.6 for details) by exact matching (see also section 4.4 for other ways to match phenotype descriptions and terms).

224,387 ontology terms and 446,449 synonyms to these terms were matched from the controlled vocabularies GO, MP and MeSH to all 38,656 phenotype descriptions in the set by exact matching (see Table 9 for results). Matching terms and synonyms from all vocabularies (called 'hits') was possible with varying degrees of ontology usage (between 44.66% and 0.55%), where 'ontology usage' is the percentage of 'hits' from the total amount of terms and synonyms in the ontology. Where a synonym was matched in the description, these terms were assigned to their respective term ('Weighted terms').

Table 9:

Result of matching terms from controlled vocabularies to phenotype descriptions.

	MP	MeSH Supplementary Concept Records	MeSH Descriptors	MeSH Qualifiers	GO
Terms	5,606	170,663	24,357	83	23,678
Synonyms	1,980	268,792	152,166	235	23,276
Hits	2,857	2,406	13,242	142	1,792
Ontology usage	37.66%	0.55%	7.50%	44.66%	3.82%
Weighted terms	1,093	1,966	8,412	67	1,642
Mean term length	27.74	27.41	17.57	12.63	38.46

Terms: The number of terms in the vocabulary.

Synonyms: The number of synonyms to keywords in the vocabulary.

Hits: The number of unique keywords or synonyms found in any phenotype description.

Ontology usage: The percentage of hits in all keywords and synonyms.

Weighted terms: The number of unique hits, where hits to a synonym are assigned as hits to their corresponding keyword.

The most terms (in absolute numbers) that could be matched came from the MeSH Descriptors vocabulary. However, since this a very large repository, the coverage of matches to the number of terms and synonyms was very small (only 7.50%). The largest ontology usage was seen with the MeSH Qualifiers vocabulary. This is a very small vocabulary (only 318 terms and synonyms); therefore this high term usage is not surprising. However, as almost 45% of the phenotypes in this set derived from either human or mouse (see section 2.1.6), it was surprising that only slightly more than a third (37.66%) of all terms from

MP could found in at least one phenotype description. The reason for this may be that most annotators commonly use leaf terms for annotations, thus (possibly unintentionally) disregarding many terms in the upper hierarchy of the ontology that cannot be found in this analysis.

In any case, it is noteworthy, that the phenotypes can be generally regarded as fairly well supplemented with vocabulary terms. 5-10% of the terms in almost half of the descriptions of phenotypes (17,351 of 38,656) consist of terms, and no term could be matched in fewer than 5% (1,727 of 38,656) of the phenotype descriptions (see Figure 24). Also, there is no phenotype description consisting of more than 45% keywords, which is not surprising (as this would resemble more a listing of vocabulary than a description).

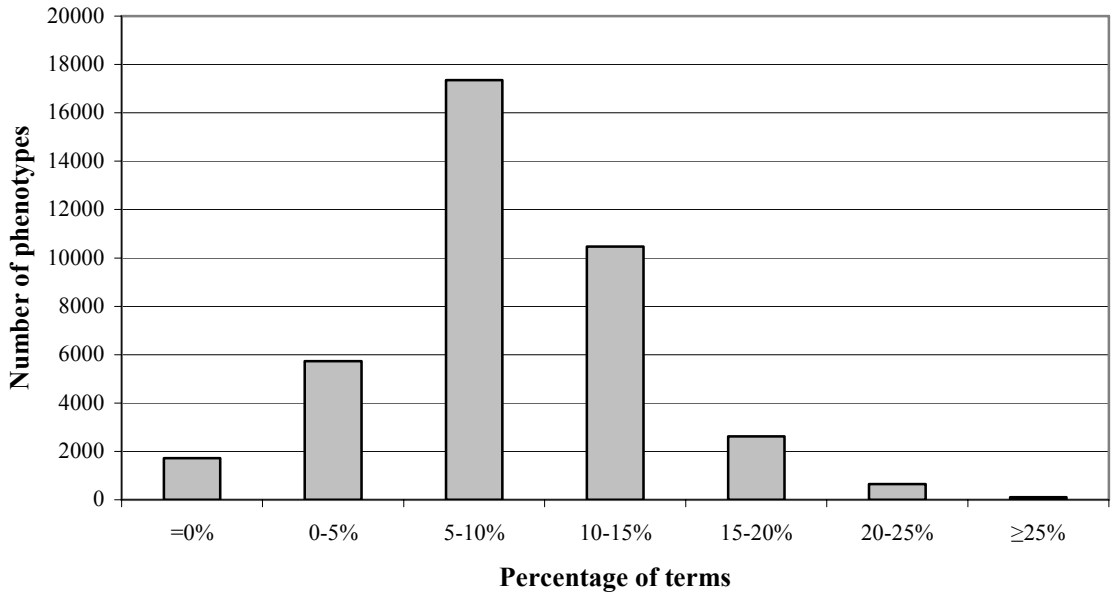


Figure 24:
Distribution of the percentages of keywords from controlled vocabularies in the total number of words in the phenotype description.

These phenotypes were then transformed into phenodocs (see section 2.1.6) and all features that were found in any one of the vocabularies named above were over-weighted 10-fold in comparison to all other features, which were weighted normally according to the TFIDF weighting scheme (see section 2.2.1 for details on both). The phenodocs with

weighted terms were used in another clustering experiment (with k=1,000). The resulting groups of genes were evaluated in respect to the distribution of clusters in species, precision and recall of functional predictions, the number of phenocopies found and the PPi-connectivity (see sections 2.1.1, 2.1.3, 2.2.2, 2.2.7 and 2.2.8 for details). The results can be found in Table 10 and Table 11. The effect of stemming on the clustering results was also evaluated.

3.2.3.2 Weighting ontology terms overcomes species-specific vocabulary

When looking at either Table 10 or Table 11, the first striking feature is the difference between the TF-scores and the TFIDF-scores. The next very obvious result is that the percentage of cluster with mixed species is much higher in the weighted scheme (up to 32.3%) than in the unweighted scheme (~10 to 14%). It seems that the weighting helps to overcome to some extent the boundary set by the species-specific vocabulary (e.g. screening methodology) already discussed in section 3.2.2.6. A closer look at these figures reveals that the fraction of clusters best provided with a controlled vocabulary, i.e. mouse-specific clusters, drops considerably from 16.60% in the unweighted TFIDF-scheme to 12.90% in the weighted TFIDF-scheme (Table 10, stemmed results). This leads to the conclusion that the over-weighting of certain terms helps to overcome the emphasis of specific vocabulary in favour of the less abundant more general (e.g. functional) terminology. As this also holds for other model organisms, it may be a general theme in term over-weighting.

However, a drop in specificity cannot be observed in human phenoclusters. This is unexpected and can only be explained by the fact that human disease descriptions from OMIM are by far the longest free-text descriptions so that even an over-weighting of a few keywords will not compensate for the sheer number of vocabulary-specific clinical descriptions. It is also surprising to see a significant drop in the number of fly-specific clusters (from 44.7% to 35% in the same schemes as above), but not in worm-specific clusters (from 11.6% to 11.1%). The reasons for this is assumed to have something to do with the very specific three-letter controlled vocabulary (e.g. ‘SLO’, ‘EMB’, ‘GRO’, etc.) that is used to describe many *Caenorhabditis elegans* phenotypes (see Figure 17 for an example of a typical *Caenorhabditis elegans* phenotype) but not *Drosophila melanogaster* phenotypes. The assumption thus holds that worm vocabulary is hard to generalize and leads to the conclusions that the most tenacious species-specific vocabulary can be found in phenotypes from *Homo sapiens* and *Caenorhabditis elegans*; likely for different reasons, but with the same effect: they are much harder to be clustered across species.

3.2.3.3 Weighted terms in phenodocs yield higher biological coherence

Regarding the functional predictions, it can be seen that the over-weighting has not only decreased the number of species-specific clusters. It seems indeed that the overweighting has an effect on the functional coherence of the clustering. The F-measures from the over-weighted clustering (between 0.3227 and 0.3417) are higher than those from the unweighted clustering (between 0.2985 and 0.3105). This indicates higher functional (i.e. biological) coherence of the clusters when terms from the controlled vocabularies are over-weighted. Still, it is surprising that the best F-measure of 0.3417 is observed for the over-weighted TF-score (see Table 11), but it is for the smallest number of predictions for only 140 clusters and thus could also be a statistical effect. Notably, the over-weighting of unstemmed terms with a TFIDF-scheme yields a prediction precision of almost 71% (see Table 10), which is an outstanding result and exceeds the precision values observed so far (for k=1,000). It can be concluded that the over-weighting of terms in phenodocs result in a higher biological coherence of the clusters.

3.2.3.4 Phenocopies and PPi are species-specific measures of coherence

This section will show that the prediction of gene functions and the measurement of the resulting precision and recall values are the best available measure of functional coherence that have yet been found. When regarding the number of phenocopies (compare to section 3.2.2.3) and the PPi-connectivity (compare to section 3.2.2.1), it can be seen that the numbers are much better for the unweighted clusters than they are for the over-weighted clusters.

For example, the number of phenocopies found in the same cluster is between 7 and 10 for unweighted clusters (both tables, unweighted results), but only between 2 and 6 for over-weighted clusters (both tables, weighted results). Similarly, the numbers of unweighted clusters with a high PPi-connectivity (over 33% of proteins in a group interact with each other) are between 201 and 255. On the other hand, they range from 202 to 237 for the over-weighted clusters. This difference may not seem large, but the number of TF-scored clusters with very high PPi-connectivity (over 75% of cluster members interact) drops almost by half (from 99 to 53), when comparing weighted clusters derived with TFIDF-score and their weighted counterparts in Table 10.

These observations may appear to contradict the conclusions from the previous sections. However, PPi-connectivity and phenocopies are species-specific measures, since – by

definition – only proteins from the same species interact and phenocopies must also derive from the same species in order to be observed here. Thus, an increase in the number of mixed-species clusters must lead to a decrease for these two measures. This is the case here and it can thus be concluded that PPi-connectivity and phenocopies are species-specific measures of functional coherence.

Now, given the observations from the previous sections, it seems feasible that PPi-connectivity and functional coherence by GO-annotations are to some extent complementary (at least with regard to the cross-species clusters) and thus, the seemingly 'low' number of coherent clusters are somewhat additive and – regarded from this perspective – not so low anymore. This is more evidence that phenoclusters represent valuable units of high functional coherence. Furthermore, it is postulated here that extending the PPi-connectivity measure for proteins across species e.g. by inference from orthologies (which is only feasible in some cases, however), the number of over-weighted clusters with a high PPi-connectivity are likely to increase, possibly even above the observed numbers for the unweighted clusters.

3.2.3.5 Term-weights have a higher impact on coherence than stemming

Looking at the results for phenoclusters from the same weighting scheme for which the terms were stemmed or not stemmed *a priori* (see section 2.2.1 for detailed methods for over-weighting and stemming), there is very little difference in the measures recorded in this experiment. The largest difference that can be observed is between stemmed and weighted clusters and unstemmed and weighted clusters with the TF-score (see Table 11). The number of mixed-species cluster doubles from the unstemmed to the stemmed clustering. This is mostly due to significant drops in mouse- and fly-specific clusters which cannot be observed to this extent in any other pairing of that kind. It is obvious that stemming has an effect and generalizes terms but can only enhance the effect of generality *after* over-weighting of terms.

From these observations, it is concluded that stemming does have an effect on the biological coherence of the phenoclusters, but this effect is much smaller than that of weighting terms. This can already be observed by comparing TF-scoring and TFIDF-scoring; even when comparing across stemmed and unstemmed clusters, the TFIDF-score usually performs slightly better (and in one case only slightly worse) in terms of the percentage of

mixed-species clusters. This underlines the conclusion that an effective weighting scheme is much more important than stemming.

Table 10:

Results of phenodoc clustering using the TFIDF weighting scheme for all phenodocs. Here, the clustering of phenodocs with 10-fold over-weighted vocabulary terms and stemming of words is compared to clustering without stemming and no over-weighting (k = 1,000).

TFIDF		**unweighted & unstemmed**	**unweighted & stemmed**	**weighted & stemmed**	**weighted & unstemmed**
Percentage of species in clusters	Mm	15.8%	16.6%	12.9%	14.3%
	Sc	1.6%	2.0%	1.5%	1.3%
	Dr	1.5%	1.6%	0.5%	1.3%
	Ce	11.2%	11.6%	11.1%	9.3%
	Dm	46.8%	44.7%	35.0%	40.1%
	Hs	4.6%	5.3%	5.7%	6.7%
	Dd	3.7%	3.6%	1.1%	2.0%
	Mixed	14.8%	14.6%	32.2%	25.0%
Prediction	# clusters	234	243	175	201
	# predicted terms	417	431	317	336
	# genes	3,857	3,882	2,743	3,537
	Recall	19.99%	19.92%	22.10%	21.26%
	Precision	65.48%	69.15%	64.14%	70.86%
	F-measure	0.3063	0.3093	0.3287	0.3271
Phenocopies		8 of 25	7 of 27	5 of 25	6 of 25
Intact connectivity clusters	75% or above	97 (1,724 genes)	99 (1,757 genes)	53 (932 genes)	52 (1,423 genes)
	33%-75%	155 (3,338 genes)	156 (3,258 genes)	167 (5,389 genes)	185 (5,095 genes)

Here, phenodocs were clustered applying filter 1 from Table 6 four times with the basic TFIDF weighting scheme to compare different methods of word manipulation: stemming and over-weighting of vocabulary terms. Results are shown for the distribution of clusters in species, functional predictions of GO-terms (counting only unique and exactly matching GO-terms as correctly predicted), the number of phenocopies found, as well as the PPi-connectivity with information taken from IntAct. See text for further details and interpretation of the results shown here. **Ce** = Caenorhabditis elegans; **Dm** = Drosophila melanogaster; **Hs** = Homo sapiens; **Mm** = Mus musculus; **Sc** = Saccharomyces cerevisiae; **Dr** = Danio rerio; **Dd** = Dictyostelium discoideum

Table 11:

Results of phenodoc clustering using the TF weighting scheme. Here, the clustering of phenodocs with 10-fold over-weighted vocabulary terms and stemming of words is compared to clustering without stemming and no over-weighting (k = 1,000).

TF		**unweighted & unstemmed**	**unweighted & stemmed**	**weighted & stemmed**	**weighted & unstemmed**
Percentage of species in clusters	Mm	18.5%	16.9%	13.0%	19.8%
	Sc	1.2%	1.3%	2.5%	1.0%
	Dr	2.4%	2.7%	1.3%	1.7%
	Ce	14.5%	13.7%	12.1%	12.3%
	Dm	41.7%	43.4%	32.6%	41.9%
	Hs	6.6%	5.7%	8.2%	4.9%
	Dd	4.9%	4.9%	1.6%	3.6%
	Mixed	10.2%	11.4%	28.7%	14.8%
Prediction	# clusters	197	211	140	170
	# predicted terms	330	366	226	284
	# genes	3,236	3,747	2,493	2,951
	Recall	20.03%	20.44%	22.83%	21.01%
	Precision	68.99%	66.02%	67.92%	69.52%
	F-measure	0.3105	0.2985	0.3417	0.3227
Phenocopies		7 of 25	10 of 25	2 of 25	3 of 25
Intact connectivity clusters	75% or above	68 (1,846 genes)	67 (1,695 genes)	49 (924 genes)	42 (1,368 genes)
	33%-75%	133 (4,616 genes)	134 (4,544 genes)	153 (6,030 genes)	164 (6,233 genes

Here, phenodocs were clustered applying filter 1 from Table 6 four times with the basic TF weighting scheme to compare different methods of word manipulation: stemming and over-weighting of vocabulary terms. Results are shown for the distribution of clusters in species, functional predictions of GO-terms (counting only unique and exactly matching GO-terms as correctly predicted), the number of phenocopies found, as well as the PPi-connectivity with information taken from IntAct. See text for further details and interpretation of the results shown here. **Ce** = *Caenorhabditis elegans*; **Dm** = *Drosophila melanogaster*; **Hs** = *Homo sapiens*; **Mm** = *Mus musculus*; **Sc** = *Saccharomyces cerevisiae*; **Dr** = *Danio rerio*; **Dd** = *Dictyostelium discoideum*

3.3 PhenoMIX: Genotype/phenotype data for new discoveries

3.3.1 Introduction

PhenoMIX is a prototype system aimed at making the results of this thesis publicly available. It is based on the contents and concepts of PhenomicDB and introduces phenoclusters (see section 3.2), as well as additional pair-wise similarities among genotypes and phenotypes, like sequence similarity (see sections 2.1.1 and 2.2.4), pair-wise PPi (see sections 2.1.3), the similarities of genotypes and phenotypes imposed from Gene Ontology and Mammalian Phenotype ontology annotations, respectively (see sections 2.1.4, 2.1.5 and 2.2.3), as well as the pair-wise cosine similarity measures among phenotype descriptions (see sections 2.1.6 and 2.2.1).

In order to enable easy and systematic access, a new database scheme was implemented, as well as an API programmed in the Java programming language, defining objects for accessing, querying and populating the database and objects, for making calculations on the data, as well as methods appropriate for each class. The implemented classes represent genotypes, phenotypes, annotations and 'networks' thereof (depicted as lists of genes, phenotypes and their similarity values). Such 'networks' contain their members as well as similarities between their members and an evidence code to track the similarity measure and its origin (e.g. similarities of phenotypes calculated from the cosine similarity measure or the similarity of genotypes calculated from sequence similarity). These 'networks' can be combined as multi-evidence networks, for example to emphasize connections between genotypes or phenotypes from several evidences and they can be regarded as groups of genes or phenotypes assembled by applying a similarity threshold. Groups of phenotypes can be used to predict functional annotations for genes, e.g. from the phenoclusters as demonstrated in section 3.2, or from PPi (also shown in section 3.2). Analogously, groups of genes can be used to predict the outcome of a phenotypic experiment, e.g. by predicting phenotype terms according to the PPi-connectivity of their associated genes. Based on the experience with phenoclusters, it is postulated here that it is feasible to infer terms common to most phenotypes grouped from genes in a similarity network. Such terms represent the typical outcome of a phenotype experiment, thus enabling the prediction of plausible phenotypes for genes. This hypothesis is tested in section 3.3.6.

Finally, with the help of a student assistant, the basic functionalities of the database and the API were implemented into a graphical web interface, such that the system can be used in

a productive version until full functionality of PhenoMIX becomes available with PhenomicDB version 3.x. The PhenoMIX prototype has been made freely available at http://www.phenomix.de. This prototype is currently being reimplemented in the development departments of metalife AG for productive service.

3.3.2 Data

As in PhenomicDB, the key feature of PhenoMIX lies in the direct relation between genotypes and phenotypes. By this, it is possible to infer any available information about a genotype or a group of genotypes (grouped by any biologically feasible method) to a group of phenotypes and vice versa. This is a central feature of PhenomicDB, where the orthology of genes (one biologically feasible type of grouping) has enabled showing the phenotypic context of grouped genes (here: orthologous genes). This feature, also available in PhenoMIX, has been extended to include the depiction of related genes by their sequence similarity, their involvement in a pair-wise PPi, and their similarity imposed from annotation of GO.

Table 12:
Data types and derived similarity measures available in PhenoMIX. See Appendix A5 for distributions of values.

Data types	Similarities
Genotypes	
GO-annotations (ontology for genes)	Similarity from GO-annotations
Interactions (IntAct and BioGrid)	Similarity from interaction (binary)
Orthologies	Orthology similarity (binary)
(Nucleotide) Sequences	Sequence similarity
Phenotypes	Similarities by association to phenotypes
Phenotypes	
Descriptions	Cosine similarity from descriptions
MP-annotations	Similarity from MP-annotations
Phenocluster	Membership in a phenocluster

In addition to these extensions, the novel feature is now to show similar phenotypes, either by membership in a phenocluster (see section 3.2) or by phenotype similarity. Similarities of phenotypes are given either by cosine similarity (see section 2.2.1) or by the similarity of the ontology annotations from MP (which is analogous to the pair-wise gene similarity by ontology annotation from GO described in sections 2.2.1 and 2.2.3). Furthermore, the

depiction of a group of genes and phenotypes derived from phenoclusters is possible. Table 12 gives an overview of the available data types and derived similarity measures in PhenoMIX.

3.3.3 Structure

The structure of PhenoMIX very closely resembles the structure of the very first prototype of PhenomicDB (MSP, described in section 3.1.1, see Appendix A1.1 for the database scheme of MSP). It has three central tables, one containing genotypes (see Appendix A2.1 for the 'genotype' tables), the other phenotypes (see Appendix A2.2 for the 'phenotype' tables) and their *genotype_phenotype* relation table enabling both the currently implemented 1:N relationship, but also the envisioned N:M relationship, if 'complex traits' enter the database (see section 1.4.2 for more details). Since some of the genotypes and phenotypes are annotated with GO- or MP-terms, the tables *genotype* and *phenotype* both also have an attribute table *genotype_attr* and *phenotype_attr*, responsible for keeping track of the N:M relationship between these entities and their annotations. As long as these two tables contain only ontology terms, they are both only connected to the tables containing the entire ontology (*go_terms* and *mp_terms* respectively), keeping the ontology's identifier and the term itself (as textual description and – in case of GO – also its source branch in the form of the field *namespace*). Furthermore, within the tables, a count is kept with each term on the number of genotypes or phenotypes that are annotated with it. This count is necessary for efficiently calculating the pair-wise similarities of genes or phenotypes from the group of associated ontology terms (see Definition 6 in section 2.2.3). The other necessary tables for these calculations are named *go_tree* (for genotypes) and *mp_tree* (for phenotypes). In these tables, the hierarchies of the GO and MP ontologies are stored respectively. They contain each term from the ontology and all its parent terms. Further tables that were established keep the pair-wise similarities of GO-terms (named *go_go*) and MP-terms (named *mp_mp*).

In analogy to the table *genotype_phenotype*, there are two tables for storing pair-wise similarities of genotypes and phenotypes, namely *genotype_genotype* and *phenotype_phenotype*. The tables have in common that each entry consists of a unique identifier, two identifiers from either genotype or phenotype, their similarity value (e.g. '1' if the 'similarity' is derived from an interaction or orthology), and the evidence code, which is a

foreign key provided by the table *data_evidence* keeping track of the sources of each type of similarity.

Since some of the genotype-genotype similarity values are calculated from sequences, these sequences are also stored in the database, in the table called *sequences*. Since Entrez gene models its genotypes so that several sequences can be available for one genotype entry (e.g. due to alternative splicing), these two tables are connected by the table *genotype_sequence*, containing as key the two foreign keys pointing to these tables. To reference sequences to their original sources, the table *sequences* contains the RefSeq ID and the source URL.

To keep track of phenoclusters, the tables named *phenocluster*, *phenoclustering, phenotype_cluster* and *phenofeatures* are also part of the database. The table *phenocluster* contains all available clusters, the number of clusters and a member count. The table *phenotype_cluster* contains the identifiers of the phenotypes and the identifiers of the clusters they are member of as well as the identifier of the appropriate clustering information. This clustering information, i.e. the parameter sets with which phenotypes have been clustered, is kept in the table *phenoclustering*. The *phenofeatures* table only keeps each of the feature identifiers from vectors of phenodocs and their associated term from the phenodocs for documentation purposes.

All tables are documented in the database scheme which can be found in Appendix A2.

3.3.4 PhenoMIX API

As described above, PhenoMIX consists of a database and a Java API (see Figure 25 for an overview over the API's classes and Appendix A2.2 for the full documentation). The API's most basic functionality consists of the possibility to extract entities (genes with associated phenotypes or phenotypes with associated genes) from the database by identifier. This, like all requests to the database is performed by the package *database*, where the class *DBHandle* deals with the interaction to the database. The class *DBInterface* wraps pre-formulated queries in the Structured Query Language (SQL) into functions that can be used in the classes. The data from the query is returned from the database via the class *DBReturn* and is then transferred into the appropriate objects found in the package *objects* by the factories from the package *common*. The user can request entities that are similar to the retrieved entity by specified evidence. For example, if all genes should be retrieved that are coding for an interacting protein evidenced in IntAct, then this request is passed on to the database as

explained above, but now, a *Connection* from the package *objects* is created for each pair of interacting proteins ('connectors' represented by their coding genes). In this particular case, the connection receives the connection weight of 1 (interactions only have a Boolean 'similarity' score; there is an interaction or there is no interaction). All of these *Connection* objects of resulting pair-wise interactions are stored as a list and thus form a *Network* (from the same package), which can be returned to the user, indicating what type of *Evidence* (in this example interactions from IntAct) has been used to create the *Network* and giving access to each of its members.

In further steps, these networks can be used for functional prediction or phenotype prediction (depending on the network type); the functionalities are provided in the package *calc.predict* by the classes *FunPred* and *PhenoPred* respectively (see section 2.2.7 and Appendix A2.3.2 for further details on functional predictions). For calculations on pairs of entities, the package *calc.sims* provides the classes *Alignment*, *Cosine* and *Ontology* to calculate the pair-wise sequence similarity (only when entities are of type *Genotype*), cosine similarity (only when entities are of type *Phenotype*) or similarity of ontology annotations (GO for '*Genotype*' and MP for *Phenotype* entities) respectively.

An administrator can use these classes to further populate the database. For example, one could imagine writing a main method that loops through a set of genes (e.g. from another database), calculating their pair-wise sequence similarities and storing them in the database. The class *PhenoCluster* represents objects instantiating a physical clustering. The *PhenoClustering* class from the *objects* package and another class also named *PhenoClustering* in the *calc* package. The former class is designed to store meta-information to a clustering, e.g. the scoring-method (TF or TFIDF), the cutoff for the phenotype lengths (e.g. 200), and the value of k. The latter class has methods to do the actual clustering. This class calls the CLUTO package and is responsible for proper interaction with the vcluster algorithm, collecting its output, populating an instance of the 'PhenoCluster' object and writing the results to the database (in case of success) or an error message (in case of failure).

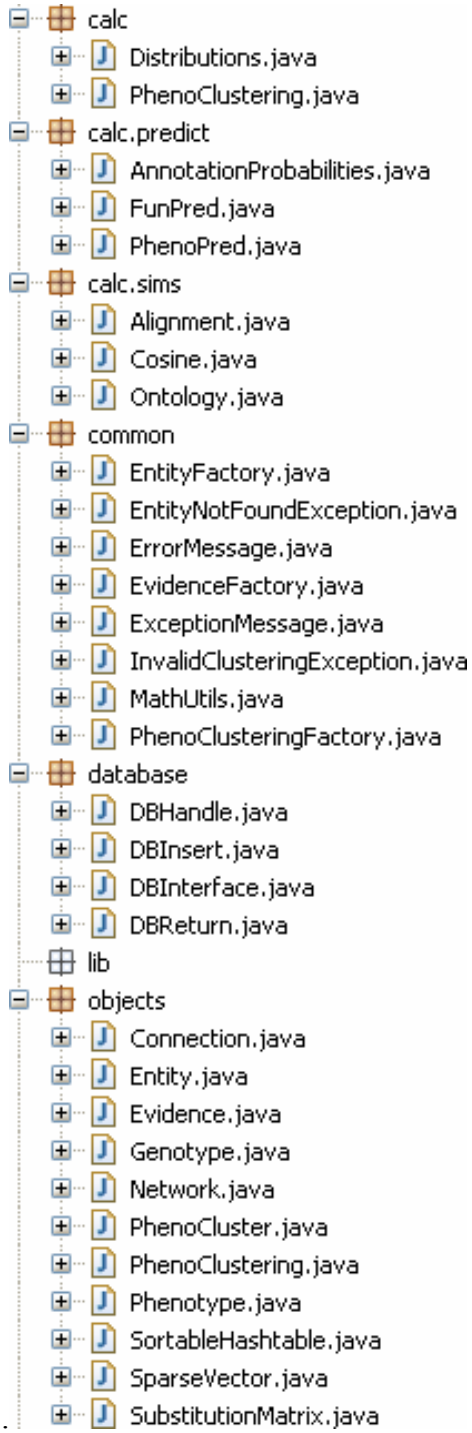


Figure 25:

Java classes forming the API of PhenoMIX. In total, the API consists of 31 classes that are arranged into 4 packages: *calc* with classes for calculations, like the clustering algorithm, *common* with classes common to several packages, like exceptions and factories, *database* with classes for database interactions, like connecting and retrieval and *objects* with classes for all objects, like the phenodoc, vector, genotype, phenotype, but also the network, etc. The *calc* package has two further sub-packages *calc.predict* for making functional and phenotypic predictions and *calc.sims* for calculating similarities between entities. Further details are in the Appendix A2.2.

3.3.5 Interface

The interface of the PhenoMIX prototype does not yet offer all functionalities from the API. It should be considered as a proof-of-concept, providing essential functionality. A full-fledged system is currently under development to be included in PhenomicDB version 3.x in collaboration with Bayer Schering Pharma AG and metalife AG. The new database for PhenomicDB 3.x scheme has already been completed and is shown in Appendix A1.4.

The website of PhenoMIX allows entering an identifier for a genotype or a phenotype (see Figure 26). The genotype identifiers are NCBI Entrez gene IDs, while the phenotype identifiers are taken from PhenomicDB.

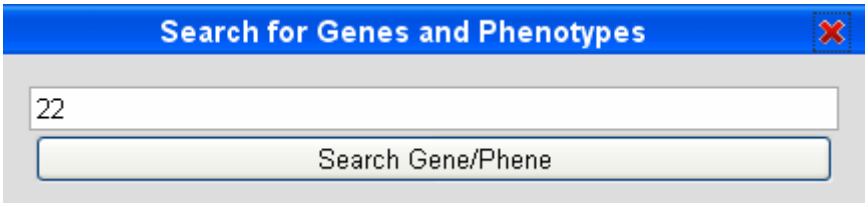


Figure 26:
Entry page for searching for a gene or phenotype identifier in PhenoMIX. Here, the integer '22' is entered as an example, which will retrieve the gene with the Entrez gene ID '22' (which is the human gene *ABCB7*). If a query is ambiguous, all results will be retrieved.

The search results are presented in the main window of the webpage (see Figure 27). To help frequent users of PhenomicDB to cope better with the interface of PhenoMIX, the coloring of genotypes and phenotypes has been imported from the appropriate style sheets. In general, genotypes are shown on a marble background, with all associated phenotypes depicted underneath with green background. The descriptions are shortened for clear arrangement, but can be toggled for a full view. Underneath their descriptions, each entry (genotypes and phenotype alike) has buttons to access a network of the connecting entities which is calculated and presented as a list on the next webpage. Each of the buttons is labeled with the evidence label from which the entity network will be calculated.

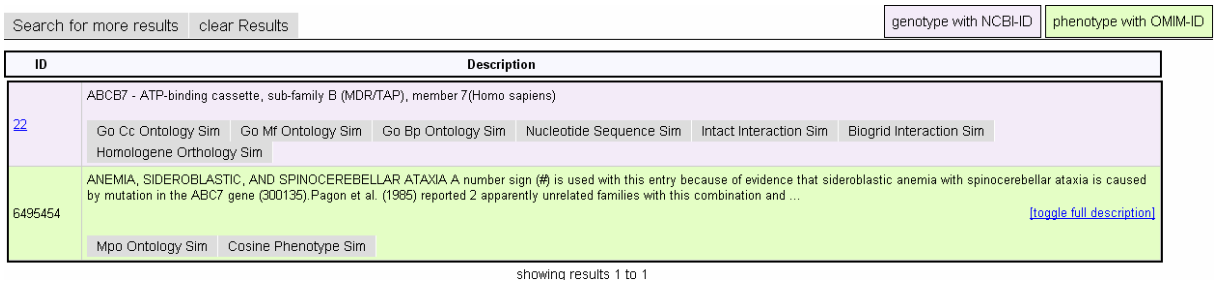


Figure 27:

Result of searching for the identifier '22'. The resulting genotype (the human gene *ABCB7*) is shown with its associated phenotype (sideroblastic anemia).

A click on any of the buttons returns the result of the calculations, i.e. the resulting list of genes and phenotypes with a similarity to the searched entity (see Figure 28). The list depicts genotypes and phenotypes in the same style as all search results are shown. Along with each of the connecting entities from the network, its similarity score to the original query is shown. The results are ordered by descending similarity score. The user can quickly find genes that are most similar to the gene of interest, along with the available descriptions and associated phenotype. It is also possible to find the most similar phenotype to a phenotype of interest and all associated genes, a feature that is not yet available in PhenomicDB.

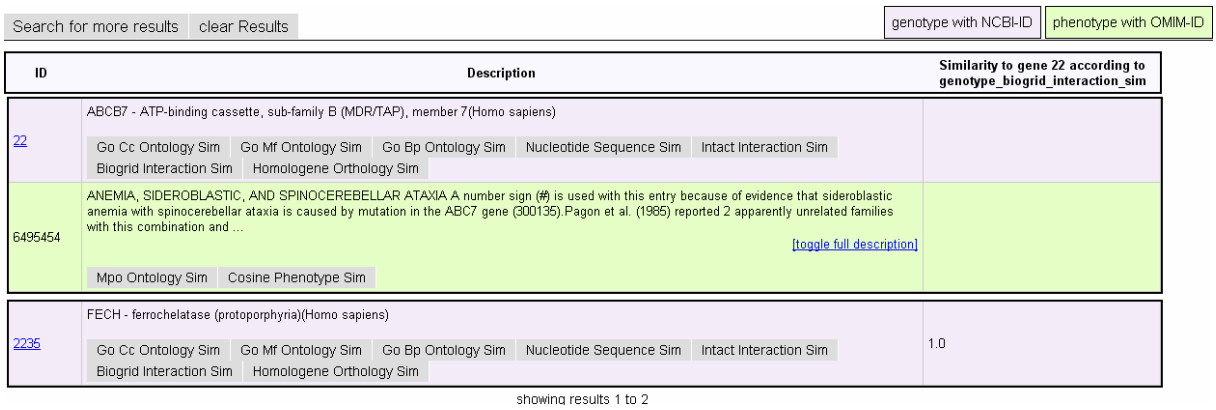


Figure 28:

Display of results when clicking the button 'BioGrid Interaction Sim' on the human gene *ABCB7*. The result is the human gene *FECH*, which is the only gene interacting with *ABCB7* according to BioGrid evidence. The depicted similarity score is '1', since an interaction is a binary 'similarity' (proteins can interact or not). In other cases, a floating point number representing the similarity score is shown.

3.3.6 Making use of PhenoMIX data

3.3.6.1 Building gene groups from similarity data

As shown in section 3.2, groups of genes derived from phenoclusters are biologically coherent and are suited for gene function prediction. As PhenoMIX brings together many more similarity measures for genotypes (see Table 12 in section 3.3.2), the hypothesis that it is possible to predict phenotype terms and GO-terms from such groups shall be tested here. One way to assemble such groups of similar genes is a 'greedy' approach, where genes above a certain similarity threshold are merged into groups (this approach is described in section 2.2.5). However, since it cannot be implied when 'A is similar to B' and 'B is similar to C' that 'A is similar to C', this approach yields many errors, low precision values and the results are therefore omitted.

Instead, all pairwise similarity measures for genotypes associated to phenotypes in PhenoMIX were clustered separately using CLUTO again. This software package also contains an implementation of the k-means algorithm enabling clustering of such large similarity matrices (see section 2.2.2.). With this, clusters of all genotypes in PhenoMIX associated with at least one phenotype and a similarity value greater than or equal to 0.1 were built using each similarity measure (see Appendix A5 for distribution of these values). To ensure comparability with the results from section 3.2 (see Table 7), it was aimed to achieve comparable mean cluster sizes resulting in slight variations of k for each similarity measure (see Table 13). The resulting clusters were filtered accordingly using filter 1 from section 3.2.2.4 (see Table 6 for details on the filter and Table 14 for results of the filtering) and mean precision and recall were recorded for the prediction of GO-terms from the biological process sub-ontology (methods in sections 2.2.7 and 2.2.8). To go beyond the phenotype prediction approaches in single-species networks [Famili, et al., 2003; Lee, et al., 2008] these clusters were then used for a cross-species phenotype term prediction approach which has not yet been tried in literature. By association of genes to phenotypes, phenotype terms (i.e. the treated phenotype descriptions from phenodocs) were considered to be the annotation of the associated gene. By this, phenotype terms could be 'predicted' in the same way as it was done for GO-terms. Furthermore, as an extension of the approach by Gunsalus et al., where three types of similarities for predicting gene function were simultaneously employed in *Caenorhabditis elegans* networks [Gunsalus, et al., 2005], all eight similarity measure were integrated in a linear combination approach, summing the

value of all similarities between two genes and dividing by the number of values used. Here, the similarity of genes in phenoclusters was calculated as a Boolean measure (two genes are either found in the same phenocluster or not).

In analogy to the control groups used in section 3.2, genes from all clusters were randomly assigned to groups of the same sizes and mean precision and recall were recorded subsequently.

Table 13:

The number of genes and pairwise similarities for each of the seven different similarity measures for genes available in PhenoMIX. Only genes associated to at least one phenotype and similarity values ≥ 0.1 were considered. Clusters of sizes comparable to the phenoclusters (see Table 7) were obtained by variations of the parameter k. Bold figures indicate the clusters with the mean size comparable to phenoclusters assembled with k = 1,000. These are evaluated in Figure 29 and in Figure 30. See Appendix A5 for the distributions of values. See Table 14 for the numbers of genes and clusters that remained after filtering.

Similarity measure	# genes	# similarities (threshold ≥ 0.1)	k to obtain mean of 15 genes/cluster	k to obtain mean of 30 genes/cluster	k to obtain mean of 60 genes/cluster
GO biological process	23,406	1,973,214	1,560	**780**	390
GO molecular function	23,565	1,373,660	1,571	**786**	393
GO cellular component	19,288	2,897,516	1,286	**643**	321
Nucleotide	16,069	27,793	1,071	**536**	268
Intact	9,692	44,193	646	**323**	162
Biogrid	11,525	84,927	768	**384**	192
HomoloGene	9,932	7,756	662	**331**	166
Linear combination	32,827	5,286,708	2,188	**1,094**	547

Table 14:

The number of remaining genes (from Table 13) and clusters, as well as mean genes per cluster and standard deviation (sd) of cluster size after application of filter 1 (described in Table 6) for each similarity measure.

Similarity measure	# genes after filtering	k	# clusters (mean # genes/cluster) after filtering	sd of cluster size after filtering
GO biological process	12,864	1,560	944 (14)	9.1287
	11,060	780	423 (26)	13.6837
	9,884	390	199 (50)	21.9711
GO molecular function	5,645	1,571	412 (14)	11.2198
	4,206	786	167 (25)	20.6362
	3,502	393	78 (45)	29.7944
GO cellular component	1,630	1,286	116(14)	23.3822
	1,304	643	36(36)	61.8958
	1,320	321	18 (73)	81.6816
Nucleotide sequence	2,702	1,071	378 (7)	4.067
	1,477	536	127 (12)	6.3867
	523	268	29 (18)	7.6368
Intact PPi	2,128	646	189 (11)	6.4099
	1,197	323	59 (20)	12.4292
	450	162	16 (28)	16.3335
BioGrid PPi	3,149	768	248 (13)	7.1554
	1,756	384	77 (23)	13.4859
	1,021	192	29 (35)	18.4398
Homologene	1,908	662	546 (3)	0.6215
	1,056	331	287 (4)	0.6703
	532	166	141 (4)	0.68
Linear combination	9,354	2,188	675 (14)	11.6594
	8,097	1,094	299 (27)	17.906
	7,251	547	144 (50)	19.7185
Phenocluster	3,213	1,000	196 (16)	21.583

3.3.6.2 Predictions of phenotype- and GO-terms from gene groups

3.3.6.2.1 Predictions of GO-terms from phenoclusters are most successful

When looking only at the comparable clusters with a mean of 30 genes per cluster, the highest observed mean precision is 67.91% for phenoclusters (see Table 8 in section 3.2.2.6 and Figure 29). Moreover, the recall of phenoclusters is the highest (22.98%). Consequently, predicting GO-terms in phenoclusters with an F-measure of 0.343 was most successful.

Clusters from BioGrid and IntAct PPi, as well as clusters from nucleotide sequence similarity performed comparably, with precision values between 54.7% up to 63.8%. Interestingly, the best-performing clusters from sequence similarity have a mean size of 60 genes

per cluster; whereas the clusters from PPi similarity yield the highest precision values at a mean size of 15 genes per cluster (data not shown here; see full data in the Appendix A6). Thus, clusters from PPi similarity (as well as HomoloGene and GO biological process) seem to form small homogeneous 'functional modules', whereas clusters of high sequence similarity are most homogeneous and thus best suited for gene function prediction when assembled in larger groups.

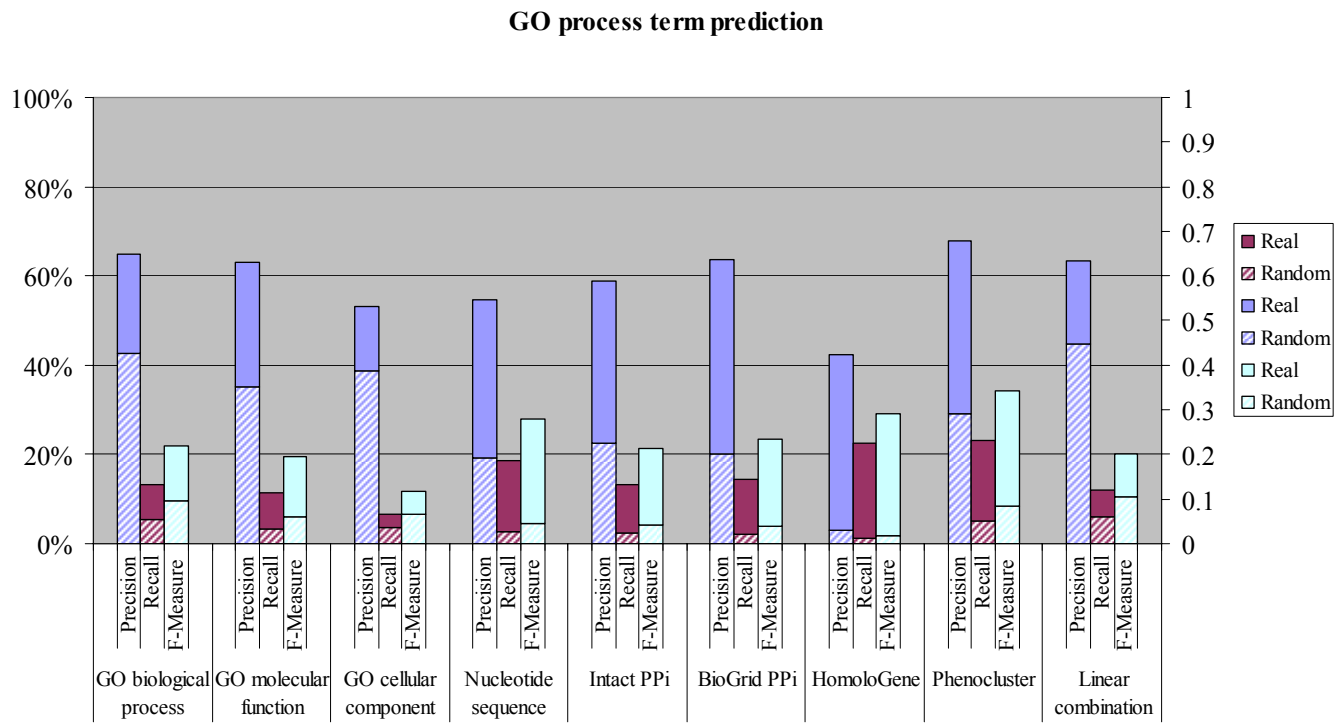


Figure 29:

Precision, recall and F-measures for predicting GO-terms from the biological process sub-ontology. Clusters were assembled with eight different similarity measures (and randomly). Mean size: 30 genes per cluster. More information is shown in Table 13 and Table 14. Full data can be found in Appendix A6.

The attempt to predict GO-terms from clusters that were assembled by similarities in the other sub-ontologies (cellular component and molecular function) resulted in precision values between 53.3% (for cellular component) and 63.0% (for molecular function), with smaller differences to the randomly assembled groups. The only clusters showing even lower precision values were derived from HomoloGene. On the one hand, this is somewhat

surprising, as the resulting clusters should be groups of orthologous genes with similar functions by definition. On the other hand, those genes were clustered aiming for a mean cluster size of 15 genes per cluster. This is a far too large number for orthology groups that commonly consist of orthologs for 10-12 different species, of which only 3-4 have an associated phenotype. Thus, the clustering failed here, as it was not possible to assemble homogeneous clusters.

Predicting GO-terms from the linear combination of all available measures was also very successful with precision values between 60.1% and 63.5%. The reasons why this approach did not yield the best precision values can most likely be found in the averaging of values within the linear combination. Here, a pair of genes that has interactions from BioGrid and IntAct, as well as a low sequence similarity of 0.6 would yield a combined similarity score of (1+1+0.6)/3=0.87, whereas it would yield a higher score in BioGrid and IntAct clusters alone. Still, this measure is better suited than each of the other measures alone due to the abundance of available data points. Furthermore, with this method, up to 850 GO-terms could be correctly predicted (see Appendix A6). This number is only exceeded by clusters derived from the GO-term similarity for biological process and molecular function terms (up to 1,313 correctly predicted terms) which – by definition – are better suited for yielding highly enriched clusters of biological process GO-terms.

3.3.6.2.2 Prediction of phenotype terms from clustered genes

Using the data from PhenoMIX, phenotype terms (from the phenodoc descriptions) were predicted for clusters of genes. Thus, PhenoMIX enables a benchmark of the ability of gene clusters to visibly reflect their functional coherence. Furthermore, this is a feature that may be used for predicting the outcome of phenotype (e.g. knock-down) experiments with genes that are similar to others that already have an associated phenotype. The results from this section can thus also be used to find the most useful similarity measures for this application.

It is interesting to note that groups from protein-protein interactions (apart from phenoclusters) perform best at predicting the terms of their associated phenotypes (see Figure 30). This shows that interaction data is most useful to reflect biological coherence. However, as already stated in section 3.2.2.1, interaction data are not abundantly available for all species. Furthermore, the quality of PPi data in repositories has been criticized [Cusick,

et al., 2009]. It is shown here that other similarity measures are also useful to infer phenotype terms from gene clusters.

Naturally, phenoclusters are best suited to predict phenotype terms. Still, clusters assembled from GO-similarities from the biological process and molecular function sub-ontologies also yield precision values above 60% (see Figure 30). These two ontologies are commonly used to describe the functionality of genes, which in many cases seems to protrude to the surface of the organism.

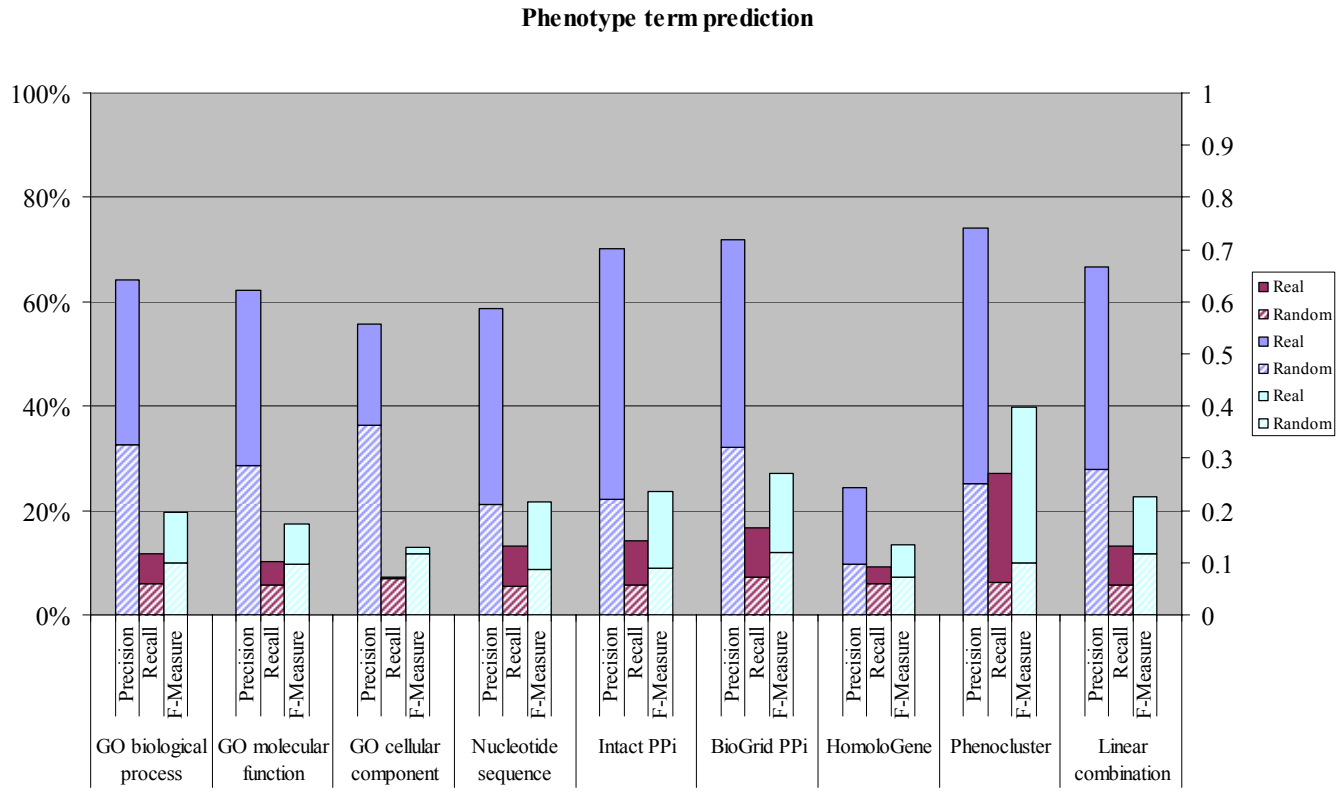


Figure 30: Precision, recall and F-measures for predicting phenotype-terms. Clusters were assembled with eight different similarity measures (and randomly). Mean size: 30 genes per cluster. More information is shown in Table 13 and Table 14. Full data can be found in Appendix A6.

Clusters from GO cellular component and from nucleotide sequence similarity perform less well, with precision values slightly below 60%. In case of sequence similarity, this seems to reflect the property that this measure (even more so on the nucleotide level) can only partially reflect function (also seen in the previous section). However, the precision

rises above 65%, when larger clusters of genes are assembled (with a mean of 60 genes per cluster, see Appendix A5). The higher number for recall may partially be due to the statistical effect that larger groups have a higher chance of 'predicting' a term. For the precision, this indicates that the 'optimum' k for phenoclusters is not the best choice for other similarity measures, as the members of these clusters are part of other biological mechanisms.

As was the case with GO-term predictions, the most correct term predictions could be made with clusters from biological process GO-terms (between 8,200 and 9,219 correctly predicted terms from phenotype descriptions – see Appendix A6). Similar to predicting GO-terms from GO clusters, the most correct phenotype terms could be predicted with phenoclusters (up to 12,896). In predicting phenotype terms, the linear combination did not perform so well, as only 5,019 to 6,397 terms could be predicted correctly, yielding, however, a high precision of up to 69.4%.

These findings show that the genotype-phenotype associations in combination with the similarity measures gathered in PhenoMIX can be used for gene function prediction as well as phenotype prediction with high precision and reasonable recall. Clustering enables the assembly of groups that can be used for this task. As shown in section 3.2.2.1, such groups, e.g. in combination with protein-protein interactions can be used to detect functional modules and supplement biological context. Furthermore, they can be used to predict the outcome of phenotype experiments which has already been shown by Lee et al. [Lee, et al., 2008] on a much smaller set of *Caenorhabditis elegans* genes.

4 Summary & Discussion

4.1 Summary of results and contributions

In this thesis, knowledge discovery and knowledge management combined with bioinformatics methods have been applied to the field of comparative phenomics. PhenomicDB, PhenoClustering and PhenoMIX are the three eye-catching implementations summarizing the results of this work.

After a thorough survey of the field, it must be concluded that the field of comparative phenomics is still in its infancy. Thus, a thorough analysis of phenotype data is a next logical step towards understanding the nature of human disease to help find novel therapeutic approaches. The vast majority of phenotype data generated so far have been gathered in free-text archives of biomedical literature and cannot be found in any structured phenotype database. A consistent approach to automated data extraction and conversion into a single structured repository has not yet been reported. In consequence, many groups have made great efforts to adapt specific parts of these data to their special scientific needs. Hence, large-scale approaches using these data have systematically been hampered by the complex nature of the data and by the difficulties of integration and normalization. Issues that will have to be tackled range from compatibility of data types of the various resources to their systematic comparability. First attempts at systematic storage and comparison have been made, e.g. with PharmGKB and PhenomicDB, but these approaches still suffer from limitations, e.g. the lack of an appropriate and unifying ontology.

This thesis contributes PhenomicDB version 2.x (see section 3.1), which has seen several important functional improvements as well as the largest increases in data content in its four years of existence. More phenotype data have been integrated in a more consistent manner, applying – for the first time in the existence of the Mammalian Phenotype ontology – its terms to cellular phenotypes of non-mammalian origin. Still, more data, especially from whole-genome RNAi screens, are expected to be included in the very near future. It is therefore expected that the percentage of human genes associated with phenotypic data will steadily rise, making it an increasingly valuable resource in biomedical research. In the meantime, the efforts from this thesis and those of others have culminated in the creation of GEN2PHEN, an EU-funded € 12 million effort involving 19 institutions aiming at comparative phenomics.

The steadily growing wealth of information raises the question of how to benefit beyond the mere rearrangement of views and data. In this work, data mining methods have been designed to take advantage of PhenomicDB's data content. Text clustering was applied to its phenotypes, and thus a novel method to create functionally coherent groups of genes was contributed here (see section 3.2). The result is a new method for automated gene function prediction from GO-terms with reasonable recall and precision values competitive to comparable methods. Furthermore, this thesis contributes valuable information about text-mining of phenotypes by identifying the limitations and the possibilities for improvement of the applied methods.

Section 3.3 describes the development of the prototype PhenoMIX. It is a system to access cross-species genotype-phenotype data and to enable comprehensive data retrieval possibilities as a first step to full-fledged data mining. PhenoMIX is a tool to group genes and phenotypes by several types of similarity measure. As shown with another clustering approach, clusters of similar genes help to gain knowledge, by presenting its members in a novel biologically coherent context. Furthermore, these clusters enable the prediction of phenotype terms, and thus the possible outcome of an RNAi experiment. To show this within this thesis, a cross-validation of methods using PhenoMIX data, predicting GO-terms and phenotype terms from gene groups clustered by several similarity measures was presented.

Thus, this work contributes to the field of comparative phenomics, giving insight into its most urgent needs, its promising developments and adding new methods and tools to further advance the field.

4.2 Related Works

4.2.1 Comparative phenomics

Some of the groundwork of comparative phenomics studies has been laid by Piano et al. who used manually curated data sets from one RNAi screen to describe a phenotype as the sum of 45 phenotypic features, each represented by either absence or presence calls. They coined the term 'phenoclusters' to describe groups of such vectorized phenotypes that 'correlate well with sequence-based functional predictions and thus may be useful in predicting functions of uncharacterized genes' [Piano, et al., 2002]. By this method termed 'Pheno-BLAST' [Gunsalus, et al., 2004], phenotypes can be compared within this data set according to the sum of absence or presence of features in the vector (Figure 31). In its version

5.0 from January 2007, PhenoBLAST supports 191 phenotypic features [PhenoBlast, 2007]. Furthermore, the well-structured manually curated *Caenorhabditis elegans* data set of Sonnichsen et al. [Sonnichsen, et al., 2005] was used to create a 'disease map', a graphical display of 45 disease categories – like 'meiotic arrest' – with values characterizing each category – like 'passage through meiosis'. Such categories are ideally taken from phenotype ontologies or other adequate vocabularies/ontologies (e.g. functional classes in GO). This profiling system allows using for example bi-clustering to group genes based on their common phenotypic feature patterns as distance measure, associating genes of unknown function directly with specific disease categories. Other clustering methods based on feature vectors have found broad application in the analysis of post-genomic data and are reviewed elsewhere [Handl, et al., 2005].

The present work goes beyond the notion of Boolean feature vectors in a single species. Here, the use of textual descriptions as the common denominator to build phenoclusters across species and screening methods has enabled function prediction on a broader basis.

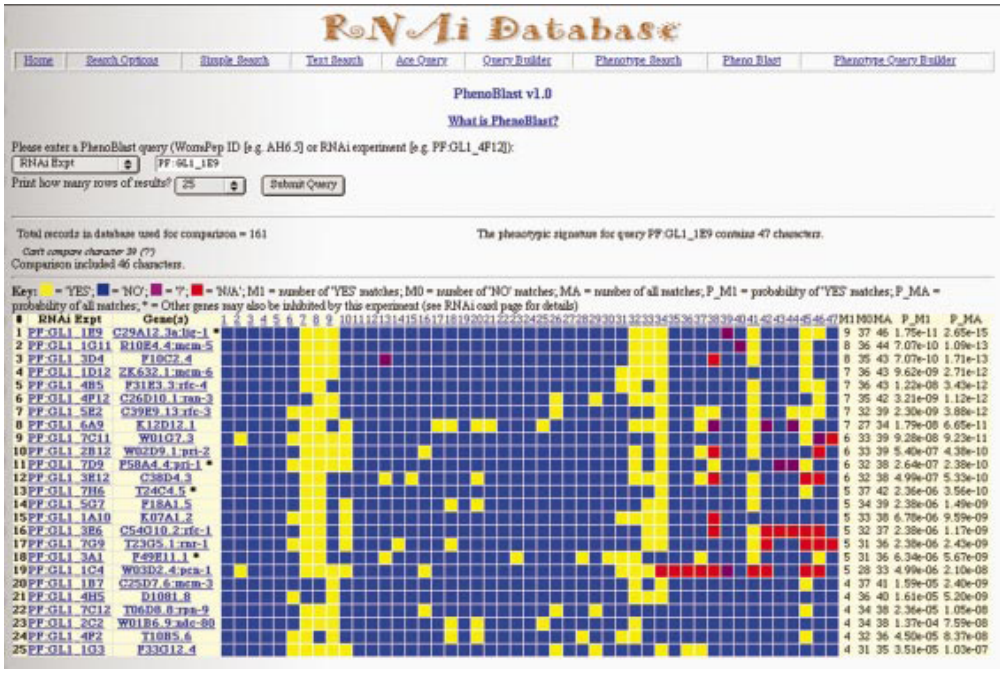


Figure 31:

Result of a PhenoBLAST query for the gene *lig-1*. The query in the top row is followed by the best-matching hits by phenotypic 'fingerprints' from the database, where blue squares represent a wild type and yellow squares an observed change in phenotype. A red square indicates that no information is available. The search is limited to *Caenorhabditis elegans* genes that have been screened and for which these features have been recorded (from: [Gunsalus, et al., 2004]).

In a study by van Driel et al. human phenotype descriptions from OMIM were compared [van Driel, et al., 2006]. They found that grouping such descriptions reflects biological modules of interacting, functionally-related genes. Lage et al. have developed a phenotype similarity score based on text-mining [Lage, et al., 2007]. They show that 90% of their similar phenotypes are equally found to be similar by human curators. They build a 'human phenome-interactome network', integrating interactions of human proteins with this phenotype score for identifying protein complexes ranked as candidates for disease models [Lage, et al., 2007]. These studies are most comparable to the work presented here. However, again, they are limited to one species (i.e. *Homo sapiens*). On the other hand, this enables ranking of candidates for disease models which was not the scope of this thesis. Instead, it was shown in this work that clustering of phenotypes across species and methods is also very successful and thus complements the works of van Driel et al. and Lage et al.

Gaulton et al. have developed a computational system to suggest new genes contributing towards a 'complex trait' (i.e. a phenotype) [Gaulton, et al., 2007]. They use ontologies and entity recognition to extract genes and proteins from phenotype descriptions and rank them to corresponding biological data from online resources. Butte and Kohane clustered keywords from the Unified Medical Language System (UMLS) annotated to gene expression data and interpret the resulting connection between these terms and the associated genes (termed 'phenome-genome network') [Butte and Kohane, 2006]. This thesis goes beyond clustering of gene annotations. Phenotypes from actual experiments are clustered and the existing associations are exploited. However, it was not pursued to find new associations between genes and phenotypes. Still, novel phenotypes and novel gene functions were predicted from actual experimental data. In a next step, this could be used for assigning new associations between genotypes and phenotypes.

Using phenotype data for more than annotation prediction, Eggert et al. compared phenotypes from RNAi as well as chemical genetic screens to find genes responsible for the same cellular phenotype [Eggert, et al., 2004]. Thus, they could identify new members of known pathways as well as small molecules with an effect on the same pathway. Bergholdt et al. have published a study in which they used physical protein interaction data in combination with genetic interactions from genome scan data to identify novel candidate genes for type 1 diabetes [Bergholdt, et al., 2007]. Eggert et al. and Bergholdt et al. have shown the possibilities of comparative phenomics. The present work builds on these studies. Comparison of phenotypes – either experimentally or electronically – and the integration

of different data types enables the discovery of novel biological context. From the experimental side, this was shown in these related works.

Prank et al. have compared methods to determine phenotype-genotype relationships in order to predict genetic alterations that lead to adrenal hyperplasia from complex biochemical data [Prank, et al., 2005]. Using serum level profiles of steroid intermediates from 54 patients with heterozygous 21-hydroxylase (CYP21B) mutations versus healthy controls, they compared traditional clinical methods, traditional linear discriminant analysis, support vector machines and nonlinear methods, i.e. artificial neural networks, and k-nearest neighbour classifiers. They showed that the nonlinear statistical analyses performed with an accuracy of up to 83%, in contrast to prediction accuracy by clinical methods of 39% and of 64% by classical linear analysis.

Generally, in order to classify phenotype data based on vectorization of their phenotypic profile as illustrated in the above examples, various supervised machine learning methods are available. The k-nearest neighbor (kNN) classification maintains a set of training cases in predefined classes (clusters) where each data point is nearest to the mean feature vector of that class. For a test case, the k nearest data points are computed and this new point is allocated to a class, depending on the prior classification of these k points by majority vote [Chaudhuri, et al., 1993]. Artificial neural networks (ANN) are an extension of the standard k-means clustering procedure and take into account a 'neighbourhood ranking' of the nearest vectors. The dynamic neighbourhood ranking takes place during an input-driven adaptation procedure of the reference vectors [Martinetz, et al., 1993]. Support vector machines (SVM) realize pattern recognition between two classes by finding a decision function (hyperplane) determined by selected points from the training data, termed support vectors. In general, this hyperplane corresponds to a linear decision boundary in the input space. While traditional techniques for pattern recognition are based on minimizing the empirical risk (i.e. on the attempt to optimize the performance of the training set), SVMs minimize the structural risk (i.e. the probability of yet-to-be-seen patterns to be classified correctly for an unknown probability distribution of the data) [Vapnik and Chapelle, 2000]. These machine-learning approaches can grasp well the typically nonlinear nature of the underlying complex genetic interactions by learning from a training set. For example, Rodin et al. have applied Bayesian belief networks to phenotype data consisting of plasma apolipoprotein E (apoE) levels from 702 African-Americans and 854 non-Hispanic whites [Rodin, et al., 2005]. From 72 individuals, 20 variable sites in the apoE gene were included

in the belief networks. Three SNPs could be singled out as most likely responsible for plasma apoE levels. This method can be used to reduce the number of candidates in an association study for a phenotype of interest, provided that reliable phenotype data are at hand. A belief network's topology shows a graphical relationship among variables (here SNPs and genes), or nodes, thus showing which variables are dependent on other variables or conditionally independent of them. Edges connecting nodes are therefore undirected and indicate dependence. The edge strength indicates the relative magnitude of the dependency between two variables, given the other interrelationships within the network. It therefore reflects a joint probability distribution among the nodes. Conveniently, an edge between two SNPs also indicates linkage disequilibrium. By employing this approach, Rodin et al. could 'simultaneously take into account linkage disequilibrium while performing genotype–phenotype association analyses' [Rodin, et al., 2005].

Clare and King have applied supervised machine learning methods to the problem of predicting the functional classes of genes in *Saccharomyces cerevisiae* from phenotypic growth data [Clare and King, 2002]. The data are combined from three different sources (TRIPLES [Kumar, et al., 2002], EUROFAN [Dujon, 1998] and MIPS [Mewes, et al., 2008]) and represented as a vectorization of attributes (growth medium) and values (observed sensitivity or resistance of the mutant compared to the wild type). The classes were assigned from the MIPS functional catalogue. The accuracy of the learned rules was then estimated using phenotype data from deletion mutants of genes with known function. Eventually, Clare et al. could predict the function of 83 genes of hitherto unknown function with an estimated precision of at least 80%.

These studies form the methodical basis for discoveries in large-scale biological data. Even though none of the methods presented above were used within the present work, they form a useful repository of statistical and computational methods to which this thesis also contributes.

Troyanskaya et al. developed MAGIC (Multisource Association of Genes by Integration of Clusters) as a general and flexible probabilistic framework to combine heterogeneous data sets for integrated analysis based on Bayesian networks [Troyanskaya, et al., 2003]. To illustrate its utility, clusters of *Saccharomyces cerevisiae* genes were formed using data about genetic and physical interactions, microarray, and transcription factor binding sites with methods like k-means clustering, self-organizing maps, and hierarchical clustering. For these clusters, MAGIC created a posterior belief for whether a gene pair has a func-

tional relationship, identifying a cluster of genes involved in ubiquitin-dependent protein catabolism, which provides 'potential functional annotation for an ORF present in that cluster (YGL004C), and confirms the recently added annotation for YNL311C'. Furthermore, they discovered a gene group involved in protein biosynthesis. In this cluster, 49 genes already annotated as involved in protein biosynthesis were found as well as 10 unknown genes. One could imagine using MAGIC also for integrated analysis of data sets from phenoclusters, making use of the high-precision predictions that have been made within this work.

Another interesting field in phenotype analysis is pathway reconstruction. From only microarray expression profiles ('global transcriptional phenotypes'), groups have successfully used epistasis analysis to reconstruct topologies of pathways in organisms such as *Dictyostelium discoideum* [Van Driessche, et al., 2005] or *Saccharomyces cerevisiae* [van de Peppel, et al., 2005]. To that end, double mutants are generated and the distance of all their expression profiles to each of the profiles of the corresponding single mutants is determined. The single mutant closer to the double mutant is topologically downstream from the other single mutant. However, full reconstruction of pathways with components not transcriptionally regulated is only feasible if additionally external interventions such as RNAi or gene knock-outs are applied and used as 'single-gene phenotypes' as shown by Markowetz et al. [Markowetz, et al., 2005]. The present thesis may contribute to such approaches by providing more abundant and better annotated phenotype data, e.g. from PhenomicDB.

In order to better understand *Caenorhabditis elegans* embryogenesis at a systems level, a large-scale integrative approach has been employed by Gunsalus et al. [Gunsalus, et al., 2005]. Data from protein interactions, gene expression clusters, and phenotypic RNAi profile similarities were incorporated to model one large gene/protein network said to have 'a high predictive value for novel gene functions'. To integrate three different types of functional relationships, graphs were built representing 661 embryogenesis genes as nodes connected by edges suggested by any evidence from the three data sets. Integration was accomplished by finding correlation among pairs of the same nodes in the different graphs. This last high-profile multi source approach gives a first taste of the emerging power of comparative phenomics. Such a system could also work in a cross-species setting where these methods are applied for large-scale functional annotation, e.g. using PhenomicDB. Thus, these approaches form the basis of next steps that are enabled by the framework generated with this thesis. A first taste of such a combined approach has been shown in section

3.3.6, where a linear combination of similarities (using up to eight similarity measures as evidence) was clustered and used for predictions.

4.2.2 Gene similarity based on GO-annotation

Calculating the similarity between two genes based on GO-annotations is actually a question of finding the similarity between two sets of ontology terms. Thus, the first question really is to find a proper similarity measure for pairs of ontology terms (see for example: [Couto, et al., 2007; Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999]). Such measures have been reviewed and benchmarked elsewhere (see the review by Guo et al. [Guo, et al., 2006]). Then, the remaining question is how to measure the similarity of two sets of ontology terms. The frequently cited paper by Lord et al. suggests simply taking the mean of all pair-wise term similarities from the two sets of terms [Lord, et al., 2003]. Other similarity measures for groups of GO-terms have been suggested (see for example: [Tao, et al., 2007; Wang, et al., 2007]) and over 14 approaches have also recently been reviewed in a comprehensive benchmark [Pesquita, et al., 2008].

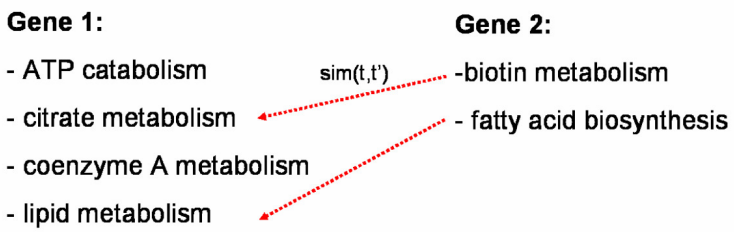


Figure 32:
Assignment of GO-terms from a smaller set of terms to a larger set of terms, where the arrows indicate the highest similarity score for each term from the smaller set (from: [Frohlich, et al., 2007]).

Frohlich et al. have developed a software package in the R programming language comprising some of these measures and an additional, novel one, labelled 'optimal assignment of terms from one gene to those of another one' [Frohlich, et al., 2007]. The idea is that two sets of terms should only be measured by their highest-scoring term pairs. For two sets of terms of unequal size, this means assigning for each term of the smaller set the most similar term from the other set (see Figure 32). For two equally-sized term sets, they suggest application of maximum weighted bipartite matching (see [West, 1999]) such that

each term in either set has exactly one assignment. In either case, the sum of these similarity measures is the similarity score assigned to the gene pair. Corresponding gene similarity calculations from GO-terms have also been suggested in this thesis (see section 2.2.3), except for the application of the maximum weighted bipartite matching for equally-sized term sets.

4.2.3 Prediction of gene function and phenotype terms

4.2.3.1 Gene function prediction from sequence- and GO-Similarity

Two features from Entrez gene can be used to calculate similarity measures reflecting – to some degree – the functional relationship of genes, i.e. the gene sequence and the GO-annotation. Many function prediction approaches are based in part on the use of sequence similarity, e.g. by use of BLAST [Altschul, et al., 1990] or PSI-BLAST [Altschul, et al., 1997; Schaffer, et al., 2001] (see: [Jones and Thornton, 2004; Laskowski, et al., 2003; Whisstock and Lesk, 2003] for some reviews on the use of sequence similarity for function prediction). Among others, approaches using GO for functional prediction are from Guo et al., Lussier et al., Tao et al. and Wu et al. [Guo, et al., 2006; Lussier, et al., 2006; Tao, et al., 2007; Wu, et al., 2005]. Pazos and Sternberg combine both sequence and GO-similarity in PHUNCTIONER, a tool using structural alignments and functional features from GO to assign in a use case correct GO-annotations to 90% of the protein structures, approximately 20% higher than inferring a functional annotation from homology alone [Pazos and Sternberg, 2004].

The availability of many fully sequenced genomes in several species and the increasing number of identified genes and proteins with unknown function has made automated function prediction, i.e. the ability to evaluate and predict gene function annotations (using for example GO), an important field in bioinformatics [Eisenberg, et al., 2000; Enright, et al., 2003; Murali, et al., 2006]. In 2003, only 62% of known eukaryotic genes had either completely unknown or only tentatively known functional annotations [Enright, et al., 2003] and still most of the other 38% were derived from annotation transfer from prokaryotic organisms [Rost, et al., 2003]. Even though this figure may have improved since then, there is still a great need for proper annotation prediction with high precision and recall [Murali, et al., 2006; Pandey, et al., 2006]. Pandey et al. [Pandey, et al., 2006] describe the success of function prediction with the Gene Ontology for many different approaches. The results show precision values for sequence similarity-based methods of 50% when decision trees

were used, a precision of up to 65% when using inductive logic programming and a precision of up to 74% with BLAST searches. Protein structure-based methods showed precision values from 75% up to 90% when predicting 121 GO-terms. With gene proximity on a set of 31 genomes, a precision value of 35% could be reached. Protein-protein interaction networks utilizing neighborhood-based approaches were able to predict annotations for 60 unannotated *Drosophila melanogaster* genes with 65% precision. The same approach reached a precision of 70% for 132 unannotated human genes. Literature-mining approaches using information retrieval on PubMed documents linked to proteins and GO-terms reached a precision of 55%. In another text-mining approach, up to 82% precision was achieved by calculating keyword frequencies in the SWISS-PROT database (see [Pandey, et al., 2006] for more details on the approaches described above).

In general, the most important goal of a function prediction method to generate new and biologically relevant results with high precision and a reasonable recall, as this represents a small rate of errors. The function prediction approaches presented in this work are competitive in the light of the studies shown above. Thus, this thesis contributes another valid method for gene function prediction, i.e. phenoclusters.

4.2.3.2 Gene function prediction from PPi

Groups of interacting proteins have highly interesting biological properties. They can be regarded as networks of genes which have been subject to intensive studies in the past. Jaeger and Leser combine structural and functional conservation sub-graphs in PPi networks of multiple species for the prediction of protein function. Based on proteins from *Mus musculus*, they predict functions for *Homo sapiens* proteins with a precision of 70% [Jaeger and Leser, 2007]. Deng et al. use PPi data *from Saccharomyces cerevisiae* and apply Markov random fields to infer protein functions from the functional annotations of interaction partners. Additionally, each predicted function is assigned a confidence probability [Deng, et al., 2003]. Xiao and Pan make predictions using protein-protein interaction data in combination with clustered gene expression profiles by weighting the evidence of both data sources [Xiao and Pan, 2005].

In this work, PPi have been used to enhance the biological contexts of phenoclusters. Furthermore, genes with PPi information have been clustered and precision and recall of GO-term predictions were measured to receive a benchmark for the same method, using other

data than phenotypes. One could imagine that combinations of PPi and phenoclusters may enable higher prediction accuracy.

4.2.3.3 Prediction of phenotypes

The first to predict phenotypes were Famili et al., conducting an analysis of a genome-scale metabolic network from *Saccharomyces cerevisiae* [Famili, et al., 2003]. Computations of functions within such a network were consistent with observed phenotypic functions for 70-80% of the considered conditions.

In a more comprehensive study, Lee et al. have built networks of genes essential for viability in *Caenorhabditis elegans* for which edges indicate the probability of involvement in the same biological process [Lee, et al., 2008]. To calculate this functional relationship probability, they integrate gene expression profiles, physical or genetic interactions, literature-mined associations, functional associations and co-inherited or operon-related homologs, building an integrated network model from all of these relationships. They then examined 43 different loss-of-function phenotypes from genome-wide RNAi screens and their responsible genes and ranked connected genes within their network by connectivity as candidates for the same phenotype. They have shown that 29 of the 43 phenotypes can be predicted with high accuracy, another 10 with accuracy better than random. The network and data is freely accessible in WormNet (currently version 1), covering 16,113 genes with 384,700 linkages (82% of all protein-coding loci) [WormNet, 2009].

This impressive study is limited to one species, i.e. *Caenorhabditis elegans*. However, even though the presented prototype PhenoMIX does not (yet) enable the prediction of RNAi phenotypes, the phenotype prediction approaches presented in this thesis are comparable to those by Lee et al., yielding similar results. One could imagine implementing the possibilities of WormNet into PhenomicDB in order to enable full-fledged phenotype prediction across species.

4.2.4 Data repositories for genotypes and phenotypes

Almost all databases for phenotypes have in common the lack of vigorous exploitation of the existing orthology information to ease phenotype comparison between species (see Table 15 for an overview over the most common phenotype resources). Besides OMIA [Lenffer, et al., 2006; Nicholas, 2003], an animal equivalent of OMIM on a smaller scale, the first cross-species phenotype database on a larger scope was PhenomicDB [Groth, et

al., 2007; Kahraman, et al., 2005]. The present work contributes to PhenomicDB and thus it remains unique in scope, content and – with PhenoMIX – in functionality. Since then only a few other integrative databases have emerged, e.g. the NCBI's dbGaP [Mailman, et al., 2007], phenotype data in Ensembl [Flicek, et al., 2008], or the recent announcement of 19 research institutes to collaborate in creating a cross-species database of genotypes and phenotypes designated GEN2PHEN (http://www.gen2phen.org) [Brookes, 2008].

Furthermore, these integrative resources are comparatively small in scale. For example the NCBI dbGaP database of genotypes and phenotypes currently stores data from 41 studies of 26 human diseases [Mailman, et al., 2007], the Online Mendelian Inheritance in Man (OMIM) catalog of human genetic disorders currently stores 4,446 diseases mapped to a genetic locus [McKusick, 2007], and only PhenomicDB has a wider scope (see section 2.1.6 for details on the data content of PhenomicDB). There are other large-scale genotype-phenotype repositories, but these are commonly kept to a single species or method.

For mice, the MGI group at Jackson Laboratory has assigned 127,506 phenotypic terms from the Mammalian Phenotype ontology (MP) [Smith, et al., 2005] to 28,986 genotypes from roughly 7,700 unique genes [Bogue and Grubb, 2004; Bult, et al., 2008; Grubb, et al., 2004; MGI, 2008; Smith, et al., 2005]. Furthermore, 853 human diseases with available genotypic mouse models are featured. Most of these data are derived from genetically engineered KO mice or naturally occurring mutants. Recently, a consortium has formed to integrate mouse phenotypes from various resources [Hancock, et al., 2007].

The Rat Genome Database (RGD) aims at integrating its genomics data with phenome data, currently covering information on more than 39,233 rat genes, 1,293 strains, and 1,337 QTLs [de la Cruz, et al., 2005; RGD, 2007; Twigger, et al., 2005]. It also offers a special disease portal presenting 4,443 disease annotations for 1,541 rat genes in their genomic context.

Table 15:

Summary of Phenotype Data Resources

Organism	Name	URL	Reference
cross-species	**PhenomicDB** – Multi-species Phenotype-Genotype Database	http://www.phenomicdb.de	[Groth, et al., 2007]
	OMIA – Online Mendelian Inheritance in Animals	http://omia.angis.org.au	[Lenffer, et al., 2006]
Homo sapiens	**OMIM** – Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/omim	[McKusick, 2007]
	HGMD – Human Gene Mutation Database	http://www.hgmd.cf.ac.uk	[Cooper, et al., 2006]
	PharmGKB – Pharmakogenetics and Pharmakogenomics Knowledge Base	http://www.pharmgkb.org	[Altman, 2007]
	GenAtlas – Gene and Phenotype database	http://www.genatlas.org	[Frezal, 1998]
Rattus norvegicus	**RGD –** Rat Genome Database	http://rgd.mcw.edu	[de la Cruz, et al., 2005]
Mus musculus	**MGD** – The Mouse Genome Database	http://www.informatics.jax.org	[Bult, et al., 2008]
	MPD – The Mouse Phenome Database	http://www.jax.org/phenome	[Bogue, et al., 2007]
Caenorhabditis elegans	**WormBase** – Biology and Genome of *Caenorhabditis elegans*	http://www.wormbase.org	[Rogers, et al., 2008]
	RNAiDB – *Caenorhabditis elegans* RNAi Database	http://nematoda.bio.nyu.edu	[Gunsalus, et al., 2004]
	PhenoBank – *Caenorhabditis elegans* RNAi Screens	http://www.phenobank.org	[Sonnichsen, et al., 2005]
Drosophila melanogaster	**FlyBase** – Database of the *Drosophila* Genome	http://www.flybase.org	[Wilson, et al., 2008]
	FlyMine – Database for *Drosophila* Genomics	http://www.flymine.org	[Lyne, et al., 2007]
	FlyRNAi – *Drosophila* RNAi Screening Center (DRSC)	http://www.flyRNAi.org	[Flockhart, et al., 2006]
Saccharomyces cerevisiae	**PROPHECY** – Profiling of Phenotypic Characteristics in Yeast	http://prophecy.lundberg.gu.se	[Fernandez-Ricaud, et al., 2007]
	SGD – *Saccharomyces* Genome Database	http://www.yeastgenome.org	[Hong, et al., 2008]
	CYGD – Comprehensive Yeast Genome Database	http://mips.gsf.de/genre/proj/yeast	[Guldener, et al., 2005]
Danio rerio	**ZFIN –** The Zebrafish Model Organism Database	http://www.zfin.org	[Sprague, et al., 2008]

WormBase [Rogers, et al., 2008] currently contains 74,602 well-organized RNAi phenotypes from *Caenorhabditis elegans*, gathered from public screens [Fraser, et al., 2000; Kamath and Ahringer, 2003; Kamath, et al., 2003; Piano, et al., 2002; Rual, et al., 2004; Simmer, et al., 2003; Walhout, et al., 2002] and data resources like PhenoBank (24,671 RNAi phenotypes for 20,981 genes) [Gonczy, et al., 2000; Gonczy, et al., 1999; Sonnichsen, et al., 2005], along with data on 1,711 mutant phenotypes and 77,763 genotypes (data counted using WormMart [Schwarz, et al., 2006] in September 2008). By this, the WormBase group has shown impressively the potential of integrating RNAi data from various screens and sources.

Also an integrative resource for RNAi screens for *Caenorhabditis elegans* is RNAiDB [Gunsalus, et al., 2004]. In their current version 5 (January 2007 release [RNAiDB, 2009]), this resource contains 59,991 RNAi phenotypes from several screens ([Fernandez, et al., 2005; Piano, et al., 2000; Piano, et al., 2002; Sonnichsen, et al., 2005]) and WormBase and RNAiDB seem to make an effort to regularly scan supplementary information of RNAi publications.

As mentioned above, large mutant screens based on different methodologies (reviewed by Carroll et al. [Carroll, et al., 2003]) have led to a rich database for *Drosophila melanogaster*. The FlyBase group has associated roughly 150,000 phenotypic statements on 14,029 genes and presents 22,954 mutant aberrations in 27,200 stocks [Tweedie, et al., 2009; Wilson, et al., 2008]. Drysdale and Crosby have given a detailed guide on how to access phenotype data in FlyBase [Drysdale and Crosby, 2005]. In contrast to WormBase, RNAi data from genome-wide screens in *Drosophila melanogaster* are being kept separately in FlyRNAi [Flockhart, et al., 2006] which is run by the Drosophila RNAi Screening Center (DRSC) at Harvard Medical School where in its current version DRSC 2.0, 13,900 genes have been targeted in 37 RNAi knock-down studies [DRSC, 2008].

More than 19,869 genotypes associated with phenotypes from *Danio rerio* (zebrafish), a helpful model organism e.g. for angiogenesis, can be found in the Zebrafish Information Network database [Sprague, et al., 2008; ZFIN, 2008]. DictyBase [Chisholm, et al., 2006] for *Dictyostelium discoideum* (slime mold) contains 1,353 genes with associated phenotype data [dictyBase, 2008].

For *Homo sapiens*, the phenotype data resources are more diverse (Clinical data as a phenotype resource are being omitted here). For a long while, the first address has been OMIM, the Online Mendelian Inheritance in Man database [McKusick, 2007], a rich hand-curated free-text catalogue linking human genes to genetic disorders. OMIM is good at giving excellent textual reviews of the literature but its shallow structure makes the text corpus difficult for automatic parsing into categorized facts. The number of phenotype entries in GenAtlas is similar to that of OMIM (4,039 phenotypes, 21,202 genes) [Frezal, 1998; Roux-Rouquie, et al., 1999]. In contrast to OMIM, GenAtlas is rich in gene details and the phenotype data are listed in the 'Variant and Pathology' section which is further subdivided into several well-structured fields. This substructure is also reflected in the query interface allowing narrow and specific filtering. HGMD, the Human Genome Mutation Database [Cooper, et al., 2006], has been collecting and manually curating disease-causing mutations for over 20 years and now covers 60,036 mutations of 2,232 genes in the public version. The HGMD professional release 2008.2 covers 80,887 mutations of 3,064 genes [HGMD, 2008]. HGMD also features a large list of links to more than 300 specialized locus-specific mutation databases (reviewed by Claustres et al. [Claustres, et al., 2002]). GeneClinics (recently renamed GeneTests) is intended for a medical audience providing over 451 very detailed and peer-reviewed 'GeneReviews' rich in phenotype data of genes with an available genetic test, covering a total of 1,610 diseases [GeneTests, 2008]. Whereas dbSNP [Wheeler, et al., 2008] is the repository for human SNPs, the Human Genome Variation Database (HGVbase) group [Estivill, et al., 2008] has announced they will go one step further in the future, annotating these SNPs with their phenotypic consequences [Patrinos and Brookes, 2005]. HGVbase and dbSNP exchange their SNP content bidirectionally. SNPeffect [Reumers, et al., 2008; Reumers, et al., 2006], in contrast, tries to predict the effects of SNPs on functional or physicochemical properties of the corresponding proteins. The Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) compiles data on how genetic variation contributes to variation in drug response [Altman, 2007; Hodge, et al., 2007]. Currently, PharmGKB holds information on 542 diseases and 546 drugs for 608 genes [PharmGKB, 2008].

There are additional databases on human diseases, genetic variation and phenotypes but they are either too specialized to be mentioned here, or do not connect the phenotypes to the genotypes, or have restricted access. Many of them are reviewed elsewhere [Ayme, 2000; Claustres, et al., 2002; Horaitis and Cotton, 2004].

4.2.5 Data integration approaches

In general, and not specifically for the phenotype data domain, various data integration approaches can be applied, e.g. data warehousing (collating all data in a single large database) [Theodoratos and Sellis, 2000], or federation of independent databases [Busse, et al., 1999]. These approaches have different advantages and disadvantages which are discussed elsewhere [Stein, 2003]. Integration of phenotype data, however, is a notoriously difficult and time-consuming endeavour, especially for cross-platform or cross-species data and has previously been shown by other groups [Hubbard, 2002; Kahraman, et al., 2005; Lutz, et al., 2005; Searls, 2005]. PhenomicDB is contributed as an integrated resource. However, as the integration was mostly done by hand, and 'screen scraping' has been the method of choice for data extraction, the present work does not contribute a novel approach to the field of database integration.

4.3 Discussion

4.3.1 Using phenotype data

For a long time, phenotypes have been regarded solely as indicators for changes in genotypes or diseases. The ability to interfere with the genetic component in a systematic manner, e.g. by gene knock-out or RNA interference [Hannon, 2002; Shi, 2003], has raised the importance of phenotypes as a tool to understand biological processes on the molecular level. Even though whole-genome RNAi screens have created large amounts of publicly available phenotype data, very few attempts have been reported to systematically analyze these data beyond single gene effects. It is noteworthy that even 5 years after the availability of RNAi screens for mammals and many calls for standardization of data types and analysis methods in phenomics [Freimer and Sabatti, 2003; Scriver, 2004], such data are still poorly organized and difficult to access. Only the Eumorphia project has released standard operating procedures for phenotype screening in *Mus musculus* and has created PhenoStat, a tool for visualization and systematic statistical analysis of standardized phenotype data [Brown, et al., 2005; Green, et al., 2005].

One of the reasons for this lack of organization lies in the data heterogeneity. The term 'phenotype' in itself is used for a broad variety of concepts, including the descriptions of clinical diseases, the characterization of naturally occurring mutants or experimentally generated mutants, and RNAi screens or gene knock-out experiments, and sometimes even for large-scale microarray gene expression data, which makes an integrated analysis of

phenotypes from different experiments and laboratories particularly hard (see section 1.3.1 and [Groth and Weiss, 2006a]). Another issue was that until very recently, no comprehensive set of phenotypes with associated genes was available. This issue has been partly addressed by the creation of PhenomicDB and the extensions made within this thesis (see section 3.1). PhenomicDB helps to integrate more data in a more consistent manner. Especially the large amounts of data from high-throughput screening methods that are being generated can thus be integrated in a way that is more likely to keep their value beyond the single genotype-phenotype relationship and its single screen of origin.

4.3.2 Cross-species phenotype clustering

4.3.2.1 Background

As stated above, integrating these data in PhenomicDB was a necessary step in order to use them beyond a single gene-phenotype relation. Almost all approaches described in section 4.2.1 have in common that they work either with very little phenotype data (usually only one data set from one screen) or with a large but very unspecific set of 'phenotypes' (such as a selected subset of Medline abstracts or ontology terms). Still, to go beyond these limitations is the next logical step in comparative phenomics. This works presents one feasible approach, i.e. to use the common denominator of these phenotypes – their textual descriptions – and cluster them (see section 3.2).

In the attempt to find the best-matching phenotypes (and in section 3.3 also genotypes for different similarity measures), clustering is a promising method. It was also attempted (but not shown) to group genes by similarity threshold ('guilty-by-association'-approach). However, such an approach does not yield precision and recall values above random thresholds. In contrast, the utilized clustering method is comparable to the search for quasi-cliques in protein interaction networks that has successfully been employed, e.g. by Jaeger et al. [Jaeger and Leser, 2007].

4.3.2.2 Cross-species clustering

One of the hurdles of clustering phenotypes is the difference in the lengths of their descriptions. Here, a drastic approach has been taken, i.e. to disregard all phenotype descriptions shorter than 200 characters (see section 2.1.6). Due to the large number of extremely short phenotype descriptions like 'embryonic lethality' created e.g. in high-throughput RNAi screens, this lead to a data reduction of 90%, probably leaving out a lot of the valuable

RNAi data tediously gathered before. However, the feasibility of clustering textual descriptions with hundreds of words to descriptions with only one or two words is questionable. This data reduction can be seen as a sacrifice to feasibility and is planned to be tackled in the release of PhenomicDB version 3.x (see section 4.4). However, this limitation of clustering can even then only be overcome by application of a complementary method for short feature vectors. Piano et al. (see section 4.2.1 and [Piano, et al., 2002]) have show that such very short vectors can effectively be transformed into Boolean absence or presence calls and then 'quasi-clustered', using e.g. 'PhenoBLAST'. Thus, a combination of approaches, phenoclustering long phenodocs, PhenoBLASTing short phenodocs, and finally merging the results in the end seems most feasible for future applications.

The next issue to tackle with cross-species phenotype clustering is the species-specific vocabulary, reflected not only by the terminology used to describe certain characteristics, but also by the descriptions of methodology of examination, which is clearly different in each of the species. As a result of this, shown in section 3.2, more than 90% of clusters contain only genes from a single species. However, this can be partially overcome by over-weighting of terms from controlled phenotype vocabularies like MP or MeSH, believed to be more general and thus applicable across species and methods. By application of a ten-fold over-weight to terms found with simple exact matching within the phenotype descriptions, it was possible to push the portion of mixed-species clusters to almost one third. However, many clusters still remain within their species, which is even more prominent for clusters from *Homo sapiens* and *Caenorhabditis elegans*, than it is for *Mus musculus* or *Drosophila melanogaster*. This tendency of genes to fall into species-specific clusters shows that the terminology used to describe a phenotype depends partly on the species, and thus on the community of researchers studying it. Such a separation of vocabulary is only partly justified, as many phenotype-effects are highly similar across species. However, until now, no common terminology for describing phenotypes in different species has been established. Such a unified ontology would open the door to more powerful ways of analyzing phenotypes, in the same manner as the establishment of GO has opened the door for many new approaches to analyzing biological knowledge. Boehm has shown that the automated build-up of such an ontology can only be partially successful and will therefore be mainly dependent on a community effort [Boehm, 2008]. However, it seems clear that the approach by Boehm to extract phenotype terms from descriptions with more elaborate

methods than simple exact matching in combination with a proper weighting of these terms prior to the clustering will result in a very feasible workaround for the time being.

4.3.2.3 Assessing Cluster Quality

The applied method for clustering here, namely k-means, is a clustering method that requires the *a priori* determination of the number of clusters (k). Typically, to assess cluster quality, internal and external measures are evaluated [Hur, et al., 2002; SanJuan and Ibekwe-SanJuan, 2006]. External measures, however (e.g. a comparison with a gold standard), could not be applied due to the lack of a gold standard for clustered phenodocs.

It is evident that in order to achieve reasonable clustering results, the parameter k has to be chosen well. The basic assumption of clustering is the presence of natural clusters within the data. However, such natural clusters are not immediately evident, especially in a high-dimensional space; otherwise the clustering task would be trivial and unnecessary.

Thus, the clustering result is most dependent on a rigorous analysis of all possibilities for k, or a more educated *a priori* selection of k and a thorough *a posteriori* validation, possibly followed by another clustering with an adjusted k, which has been done here, see section 3.2.2.6. Immediately, the main disadvantage of the *a priori* selection of k becomes evident. The existence of natural groups within the data is required but not enforced and furthermore can not be determined with certainty. Thus, in order to be able to interpret the clustering, a thorough validation is necessary. For this, many strategies have been suggested (and reviewed by Handl et al. [Handl, et al., 2005]), e.g. assessment of stability (= cluster reproducibility), or coherence in the similarity of the cluster members. There are also a number of statistical measures like the gap statistic [Tibshirani, et al., 2001], its refinement as weighted gap statistic [Yan and Ye, 2007] or the elbow criterion (controversially discussed by Bezdek and Pal [Bezdek and Pal, 1998]), both implemented in the software package SenseCluster [Pedersen, 2006].

This list of (controversial) measures for finding the ‘optimum’ k raises the question of whether that is actually possible. It is certainly possible to define a dense group of samples surrounded by a lower density boundary in a feature space as a natural cluster. The statistical measures define such density boundaries. However, it can be argued that in a high-dimensional space with many samples of similar or even overlapping origin or content, such boundaries are no longer evident. In such a case it is likely that there are several sensible values for k, dependent on the granularity of the data representation. Let us assume

that proteins should be clustered into groups according to their functional domains. Thus, the similarity of any two proteins depends on the number of domains they have in common. It is now feasible to select a k that clusters all proteins with a transmembrane domain. A decrease in k would add further proteins to this cluster – similar but without a transmembrane domain. Increasing k would result in the exemption of proteins from the cluster, even though they have a transmembrane domain. Increasing k even further could result in a cluster of all proteins with 7 transmembrane domains. Another slight increase of k now results in a cluster of only those proteins that are in fact G-protein coupled receptors. Thus, different choices for k yield different clusterings, several of which make sense and none of which is an 'optimum' in terms of an unspecific statistical boundary. In consequence, the choice for k must be dependent on the goal of the clustering. If the aim is to cluster proteins into their respective families, a good choice for k could be either the number of expected families in the data set or the number of clusters that would divide the proteins into groups of the expected size of a protein family. Hopefully, if any of these assumptions are correct, the result will be a division of proteins into families. Here, the focus lies on the *biological coherence* of the clustering result, especially in the light of different choices for k and several selection filters for samples.

Thus, in order to determine the success of a clustering in this study, biological coherence was measured by examining the relatedness of the genes in a cluster using several independent measures, i.e. protein-protein-interaction (PPi) of encoded proteins (see section 3.2.2.1), functional annotations from GO (see section 3.2.2.2), and the co-occurrence of pairs of genes known to be responsible for identical phenotypes (so-called 'phenocopies') (see sections 2.1.8 and 3.2.2.3). By these measures, it has been shown in anectodic examples that the phenoclusters presented in this work are biologically coherent and can be used to generate and extend biological contexts.

4.3.3 Prediction of annotations

As outlined in section 3.3, one of the key features of PhenoMIX is to enable new discoveries. Showing data in an as yet unseen but biologically interesting context (see section 3.3.6.2), discoveries are achieved through integrating many similarity evidences and thus enabling to look up much of the available information for a genotype or phenotype of interest.

Besides aggregating known information in a novel context, it is even more interesting to discover new knowledge, e.g. a gene found to be a member of a certain pathway. To gain insights of this sort, scientists like Eggert et al. [Eggert, et al., 2004] have developed methods, which are, however, usually accompanied by very tedious laboratory work. For such a scientist, the attractiveness of predicting candidates lies in the possibility to focus quickly on the few promising targets and compounds from a large heap, which can then be biologically verified. The possibility to predict functional annotations (i.e. GO-terms) for genotypes and the actual descriptive terms of phenotypes is a central theme throughout this thesis. PhenoMIX is designed to gather these groups, accumulated by different similarity scores and to present them to enable predictions.

The predictions that have been made within this thesis are realistic, as the cross-validation for predictions of GO-terms from phenoclusters (in section 3.2.2) and the benchmark of predicting GO-terms from other gene similarity measures, as well as the cross-validation for phenotype term predictions from groups of similar genes (in section 3.3.6.2) have shown. A linear combination of scores is a method yielding very good F-measures while combining most data points and thus enabling a large number of correct predictions for more genes than from any single score. However, linear combination can in some cases reduce the score between edges, where complementary evidences should actually increase it. This situation could be improved with a formula that adds the score of supporting edges, or applies a correlation measure as it has been done by Gunsalus et al. (see section 4.2.1 and [Gunsalus, et al., 2005]). Also, the integration of other methods, e.g. kNN (see section 4.2.1), may help improving the situation in the future, especially in the light of the development of PhenomicDB version 3.x. Variations in the application of the clustering algorithm, e.g. tuning the parameter k more towards biological coherence than towards similar cluster sizes is a possibility to improve results even without application of further methods. This was not attempted in section 3.3 because there, it was aimed to compare the possibilities of gene function prediction from phenoclusters with other similarity measures and consequently, the same methods were applied to clusters of similar sizes.

The benchmarks for the phenotype term predictions have shown that the system can feasibly be used for predicting the outcome of phenotype experiments. Still, there are other approaches, e.g. by Lee et al. (see section 4.2.3.3) that are more tuned to predicting actual RNAi phenotypes. Another possibility to make phenotype predictions is predicting terms from phenotype ontologies. However, the available ontology annotation for phenotypes is

not yet abundant enough for such an approach. This can be seen, e.g. in the similarity measures for MP-term annotations to phenotypes (in Appendix A5.2), where most MP-terms are equal. Such a limited variability within the annotation will lead to inconclusive predictions. This lack in unified annotation of phenotypes needs to be tackled by a community effort to build a cross-species phenotype ontology. This is the logical next step in the field of comparative phenomics and can be seen as one of the most urgent tasks that will have to be done in order to get better and more consistent results from phenotype data (see also section 4.4 and [Groth and Weiss, 2006a]).

4.3.4 Biological insights from phenoclusters

4.3.4.1 Overview

Evaluation of any of the presented prediction methods has the fundamental drawback that only included annotations are considered as correct. Apart from the more general criticism of GO as annotation system [Smith, et al., 2003], it is well-known that GO-annotations are highly incomplete for virtually all species. This is even more the case with annotations from MP. Thus, there is a considerable chance for false predictions that actually may be the most interesting ones from a biological point of view. False positive predictions potentially represent new functional insights, e.g. when a gene not yet annotated with a particular function is found in a cluster with a strong consensus annotation for that function. In the following, the biological nature of exemplary phenoclusters is discussed to show their biological significance beyond pure precision values (see also section 3.2.2.1 for further biological examples of phenoclusters in combination with PPi data).

4.3.4.2 Example 1: Odorant Receptors from Drosophila melanogaster

One of the clusters consisting of 25 genes shows a high consensus annotation for the three GO-terms *G-protein coupled receptor protein signaling pathway* (GO:0007186), *sensory perception of smell* (GO:0007608) and *cell-cell signaling* (GO:0007267). This group contains 24 genes from the *Drosophila melanogaster* odorant receptor (*DOR*) group and one other gene. This other gene is called *myospheroid* (*mys*). It is in several ways a very interesting group:

Firstly, all 24 genes are antennal *DOR* genes, a physiological region of *Drosophila melanogaster* in which a total of 32 *DOR* genes are located [Vosshall, et al., 2000]. The other 8 antennal *DOR* genes not found in this group (*Or13a*, *Or22b*, *Or33a*, *Or42b*, *Or56a*,

Or69a, *Or69b* and *Or83c*) are not in the initial list of 15,426 genes, likely due to a lack of substantial phenotype description. 13 of the 24 genes are annotated with all three GO-terms from the consensus annotation in that cluster, the other 11 genes are only annotated with GO-term *sensory perception of smell* (GO:0007608). The other interesting aspect is the occurrence of *mys*, which is neither a gene from the DOR group, nor annotated with any of the three GO-terms, but instead with 19 other terms, among them *signal transduction*, *axon guidance*, *calcium-dependent cell-cell adhesion*, *calcium-dependent cell-matrix adhesion*, *cell adhesion*, *cell migration* and *cell-matrix adhesion*. Even though it has been previously suggested that *mys* is a candidate gene for *Drosophila* olfactory associative learning [Roman and Davis, 2001], a link between the *Drosophila* olfactory system and *mys* has been reported only recently, in a publication not yet included in PhenomicDB. Bhandari et al. have shown that expression of *mys*-RNAi transgenes in the antennae, antennal lobes, and mushroom bodies disrupted olfactory behaviour, confirming that *mys* is important for the development and function of the *Drosophila* olfactory system [Bhandari, et al., 2006].

It is not clear why some of the antennal receptors are not annotated with the two GO-terms *G-protein coupled receptor protein signaling pathway* and *cell-cell signaling*. After all, all are clearly odorant receptor proteins consisting of seven transmembrane domains, and transduce odour recognition into neural activation through G-protein-coupled second messenger signaling pathways [Keller and Vosshall, 2003]. In the analysis of GO-annotations, these genes represent false false positive results, i.e. the annotation that has been predicted is in fact true but missing, thus reducing the precision of the prediction.

Another important lesson that can be learned from this example is the dependence of phenotypic similarity on complete and well-structured phenotypic data. Even though one study reports on both antennal receptors *Or22a* and *Or22b* (which are co-expressed in the ab3A antennal neuron and share 78% amino acid identity, separated by an intergenic region of only 650 base pairs [Dobritsa, et al., 2003]), the gene group only includes *OR22a*, simply due to missing phenotypic information about the other gene.

4.3.4.3 Example 2: 8 Drosophila melanogaster genes

In another group, yielding very high mean pair-wise phenotype- and GO-similarities, most genes are associated with the two GO-terms *cellularization* (GO:0007349) and *pole cell formation* (GO:0007279). This group consists of 8 *Drosophila melanogaster* genes, including 6 genes from the *mat(2)syn* (maternal effect syncytial blastoderm arrest) family, as well

as the *indirect flight muscle gene RU2* (*ifm(2)RU2*) and a gene called *presto*. *Presto* and the genes from the *mat(2)syn* group are both associated with deficiencies on the second *Drosophila melanogaster* chromosome and are part of closely related maternal effect loci that cause defects before cellularization of the blastoderm embryo [Schupbach and Wieschaus, 1989]. During these stages of nuclear division, the embryo is called a syncytial blastoderm, meaning that all the cleavage nuclei are contained within a common cytoplasm, including cleavages that form the two cell types of early development (pole cell and blastoderm cell) [Bate and Martinez-Arias, 1993].

Seven genes are annotated with both GO-terms, while *ifm(2)RU2* is not annotated with any of those terms. Instead, it is annotated with the term *muscle development* (GO:0007517). Still, all phenotypes in the cluster show a high similarity and this may indicate that the genes are commonly regulated or that they are part of one developmental process. The development of the indirect flight muscle has been closely associated with the *myosin heavy chain* gene (*MHC*) [Cripps, et al., 1994]. Since *ifm(2)RU2* and *MHC* are found on adjacent loci, they have been studied together in an ethyl methanesulfonate mutagenesis screen [Nongthomba and Ramachandra, 1999], where mutated *MHC* and *ifm(2)RU2* have been found to act together, enhancing muscle disorganization compared to their respective heterozygous phenotypes.

Another gene, the *95F unconventional myosin* gene (*95F MHC*) is shown to be required for proper organization of the *Drosophila melanogaster* syncytial blastoderm [Mermall and Miller, 1995]. Compared to *MHC*, this gene shows a high degree of conservation in the ATP-binding and actin-binding regions, and SH2, one of the two reactive thiols (SH1 and SH2) found in many muscle *MHC* heads is also present in *95F MHC*. The amino-terminal two thirds of the protein comprise a head domain that is 29-33% identical (60-65% similar) to other myosin heads, and contains ATP-binding, actin-binding and calmodulin/myosin light chain-binding motifs [Kellerman and Miller, 1992].

When looking only at GO-annotations, this gene is a true false positive result, i.e. its available annotation is in line with the biology but not part of the consensus annotation for the rest of the genes from this cluster, so the prediction is classed as wrong. Even though there is not yet proof for an immediate interaction between *MHC* and *95F MHC*, these genes are very similar to one another. A relationship between *MHC* and *ifm(2)RU2* has already been suggested and here, further indications have been found that those genes are responsible for similar phenotypes (the term ‘myosin’ does not occur in any of the phenotype descrip-

tions). These are reasons to believe that there are some undisclosed functional links between *95F MHC*, *MHC*, *ifm(2)RU2* and at least some *mat(2)syn* gene family members. These very interesting findings are provided here to be tested biologically.

4.4 Conclusion and Outlook

It was shown that a great deal of the heterogeneous nature of phenotype data can be overcome by application of knowledge management, integration, text-mining and prediction. Within the framework of PhenomicDB, textual descriptions of clinical diseases, naturally occurring mutants, RNAi screens, gene knock-out experiments, and many others were systematically integrated. Using clustering, a reasonable fraction of genes associated to phenotypes from PhenomicDB can be grouped into biologically meaningful categories. This works even better when the phenotype vectors are created with methods tuned to the nature of their descriptions, i.e. over-weighting of terms specific to the phenotype domain and yet general enough and thus valid across species. Grouping genes based on certain properties is a powerful tool that has often been applied for function prediction before, using criteria such as participation in the same pathway [Huynen, et al., 2003; Jaeger and Leser, 2007], participation in PPi cliques [Spirin and Mirny, 2003], or mentioning in the same Medline abstracts [Raychaudhuri, et al., 2002] – but not on phenotypes. This is an important new approach, as phenotypes (in contrast to interaction data for example) yield more information on the high diversity of biological meaning that is innate to any gene. It is, in fact, the intrinsic nature of phenotypes to visibly reflect genetic activity. Thus, phenotype data has the potential to be more useful for functional studies than most other types of data. Furthermore, the integration of genotype and phenotype data enables large-scale discoveries across species that are highly useful when real data is sparse and can help e.g. to model human diseases.

A much larger fraction of clusters are more homogeneous with respect to pair-wise interactions, GO-annotations and re-occurrence of phenocopies than expected by chance. The remaining part of the phenodocs do also cluster, probably not driven by biology but by spurious effects of the data set itself and the clustering methods; this is an effect typical of high-throughput methods (e.g. gene expression). It was shown that phenoclusters can be used to infer gene functions for poorly annotated genes with high precision and reasonable recall. In a recent survey, Pandey et al. [Pandey, et al., 2006] have collected success rates for different approaches, like sequence clustering, to protein function prediction using GO.

These approaches show precision values between 50% and 74%, i.e. from slightly below to slightly above the values presented here (note that some of the methods applied in these works are not directly comparable to the approach of this thesis). In the light of this survey, large-scale phenotype clustering as carried out here should be considered as a novel tool to predict gene functions with highly competitive results. As a generalization, it can be assumed that the prediction methods are applicable to all functionally coherent groups of biological entities and will return appropriate predictions with comparable recall and precision values.

It is planned to advance the development of the prototypes and proof-of-concepts presented here to productive versions within the framework of PhenomicDB. More specifically, the groups of similar genes and phenotypes in PhenoMIX will be used to enable automated predictions and a scoring system for the confidence of prediction results. As an example, it should be possible for an extracted group of phenotypes, their phenotypic features and their associated TFIDF-score, to multiply this score with the weight of the similarity connections (e.g. when evidence from several sources supports the connection of two entities). This would result in a score of importance for each phenotype term and can then be used to assign the most important terms as the most likely phenotypic outcome of an experiment for those genes of the group. Another desirable functionality is the prediction of ontology annotations for both genotypes and phenotypes including a probability measure for the prediction confidence.

It has been shown that rigorous amplification and application of ontologies is necessary to successfully apply innovative or classical tools to foster a more widespread use of phenotype data. Automated annotation of phenotypes will thus be the next hot topic in comparative phenomics, with the ultimate goal of developing a more general ontology. However, to retain this ontology in high quality would require the collective work of experts from all fields of phenotyping. This can possibly be achieved in combination with other promising public biomedical ontologies including more specialized phenotype ontologies (e.g. for cell type, plant trait or *Drosophila* development) which can be found in the open biomedical ontologies (OBO) repository at http://obo.sf.net. Consequently, an international consortium of phenotype research groups should be established (see also [Groth and Weiss, 2006a]), in order to drive forward the development of a universal cross-species ontology for phenotypes.

These ontologies will eventually lead to a more homogeneous body of phenotype descriptions, enabling a more refined application of term-weights than 'simple' TFIDF and an over-weighting of vocabulary terms extracted by exact matching. Adding to this, more sophisticated extraction methods, such as shown by Boehm [Boehm, 2008], as well as phenotype annotation software like Phenote (see section 1.3.2) will further advance the field.

In a final step, the potential of PhenomicDB to enable discoveries in a large cross-species scope should be combined with the methods to develop integrated functional gene networks from Lee et al. (see section 4.2.3.3 and [Lee, et al., 2008]) on more species and data and thus could leverage comparative phenomics to systemic discoveries leading – eventually – to novel therapeutic approaches.

Bibliography

Altman, R. B. (2007): PharmGKB: a logical home for knowledge relating genotype to drug response phenotype, Nat Genet (vol. 39), No. 4, p. 426.

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. and Lipman, D. J. (1990): Basic local alignment search tool, Journal of Molecular Biology (vol. 215), No. 3, pp. 403-10.

Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. and Lipman, D. J. (1997): Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research (vol. 25), No. 17, pp. 3389-402.

Apionishev, S.; Katanayeva, N. M.; Marks, S. A.; Kalderon, D. and Tomlinson, A. (2005): Drosophila Smoothened phosphorylation sites essential for Hedgehog signal transduction, Nat Cell Biol (vol. 7), No. 1, pp. 86-92.

Ascano, M., Jr. and Robbins, D. J. (2004): An intramolecular association between two domains of the protein kinase Fused is necessary for Hedgehog signaling, Mol Cell Biol (vol. 24), No. 23, pp. 10397-405.

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. and Sherlock, G. (2000): Gene ontology: tool for the unification of biology., Nature Genetics (vol. 25), No. 1, pp. 25-9.

Ayme, S. (2000): Bridging the gap between molecular genetics and metabolic medicine: access to genetic information, Eur J Pediatr (vol. 159 Suppl 3), pp. S183-5.

Backhaus; Erichson; Plinke and Weiber (2000): Multivariate Analysemethoden, Springer, Berlin, ISBN: 3540004912.

Bate, M. and Martinez-Arias, A. (1993): The Development of Drosophila melanogaster, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Bergholdt, R.; Størling, Z.M.; Lage, K.; Karlberg, E.O.; Olason, P.I.; Aalund, M.; Nerup, J.; Brunak, S.; Workman, C.T. and Pociot, F. (2007): Integrative analysis for finding genes and networks involved in diabetes and other complex diseases., Genome Biology (vol. 8), No. 11, p. R253.

Bergmann, A.; Yang, A. Y. and Srivastava, M. (2003): Regulators of IAP function: coming to grips with the grim reaper, Curr Opin Cell Biol (vol. 15), No. 6, pp. 717-24.

Bezdek, J. C. and Pal, N. R. (1998): Some new indexes of cluster validity, IEEE Trans Syst Man Cybern B Cybern (vol. 28), No. 3, pp. 301-15.

Bhandari, P.; Gargano, J. W.; Goddeeris, M. M. and Grotewiel, M. S. (2006): Behavioral responses to odorants in drosophila require nervous system expression of the beta integrin gene myospheroid, Chemical Senses (vol. 31), No. 7, pp. 627-39.

Bijlsma, M. F.; Spek, C. A. and Peppelenbosch, M. P. (2004): Hedgehog: an unusual signal transducer, Bioessays (vol. 26), No. 4, pp. 387-94.

Boehm, Christoph (2008): Ontology Construction from Phenotype Data, Diploma Thesis, Knowledge Management in Bioinformatics, Humboldt University, Berlin.

Bogue, M. (2003): Mouse Phenome Project: understanding human biology through mouse genetics and genomics, J Appl Physiol (vol. 95), No. 4, pp. 1335-7.

Bogue, M. A. and Grubb, S. C. (2004): The Mouse Phenome Project, Genetica (vol. 122), No. 1, pp. 71-4.

Bogue, M. A.; Grubb, S. C.; Maddatu, T. P. and Bult, C. J. (2007): Mouse Phenome Database (MPD), Nucleic Acids Res (vol. 35), No. Database issue, pp. D643-9.

Bonaccorsi, S.; Mottier, V.; Giansanti, M. G.; Bolkan, B. J.; Williams, B.; Goldberg, M. L. and Gatti, M. (2007): The Drosophila Lkb1 kinase is required for spindle formation and asymmetric neuroblast division, Development (vol. 134), No. 11, pp. 2183-93.

Bornemann, D. J.; Duncan, J. E.; Staatz, W.; Selleck, S. and Warrior, R. (2004): Abrogation of heparan sulfate synthesis in Drosophila disrupts the Wingless, Hedgehog and Decapentaplegic signaling pathways, Development (vol. 131), No. 9, pp. 1927-38.

Botstein, D. and Risch, N. (2003): Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease, Nat Genet (vol. 33 Suppl), pp. 228-37.

Bowler, P. J. (1996): Charles Darwin, Reissue edition. ed., Knight, D., Ed, Cambridge Science Biographies, Cambridge University Press, Cambridge, ISBN: 0521562228.

Breitkreutz, B. J.; Stark, C.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Livstone, M.; Oughtred, R.; Lackner, D. H.; Bahler, J.; Wood, V.; Dolinski, K. and Tyers, M. (2008): The BioGRID Interaction Database: 2008 update, Nucleic Acids Res (vol. 36), No. Database issue, pp. D637-40.

Brookes, A. J. (2008): Major European Project to create new Knowledgebase of Gene-Disease Relationships, The GEN2PHEN Consortium, University of Leicester, UK, accessed: April 2008, http://www.gen2phen.org/docs/PressRelease_FINAL.pdf

Brown, S. D.; Chambon, P. and de Angelis, M. H. (2005): EMPReSS: standardized phenotype screens for functional annotation of the mouse genome, Nat Genet (vol. 37), No. 11, p. 1155.

Brudno, M.; Malde, S.; Poliakov, A.; Do, C. B.; Couronne, O.; Dubchak, I. and Batzoglou, S. (2003): Glocal alignment: finding rearrangements during alignment, Bioinformatics (vol. 19 Suppl 1), pp. i54-62.

Brummelkamp, T. R.; Bernards, R. and Agami, R. (2002): A system for stable expression of short interfering RNAs in mammalian cells, Science (vol. 296), No. 5567, pp. 550-3.

Bult, C. J.; Eppig, J. T.; Kadin, J. A.; Richardson, J. E. and Blake, J. A. (2008): The Mouse Genome Database (MGD): mouse biology and model systems, Nucleic Acids Res (vol. 36), No. Database issue, pp. D724-8.

Busse, S.; Kutsche, R. D.; Leser, U. and Weber, H. (1999): Federated Information Systems: Concepts, Terminology and Architectures, Forschungsbericht des Fachbereichs Informatik, Technical University of Berlin, Germany, 1-40

Butte, A. J. and Kohane, I. S. (2006): Creation and implications of a phenome-genome network, Nat Biotechnol (vol. 24), No. 1, pp. 55-62.

Campbell, N.A. and Reece, J.B. (2008): Biology, 8th. ed., Pearson Education, Harlow, Essex, ISBN: 0-321-54424-2

Carroll, P. M.; Dougherty, B.; Ross-Macdonald, P.; Browman, K. and FitzGerald, K. (2003): Model systems in drug discovery: chemical genetics meets genomics, Pharmacol Ther (vol. 99), No. 2, pp. 183-220.

Carthew, R. W. and Rubin, G. M. (1990): seven in absentia, a gene required for specification of R7 cell fate in the Drosophila eye, Cell (vol. 63), No. 3, pp. 561-77.

Cenci, G.; Rawson, R. B.; Belloni, G.; Castrillon, D. H.; Tudor, M.; Petrucci, R.; Goldberg, M. L.; Wasserman, S. A. and Gatti, M. (1997): UbcD1, a Drosophila ubiquitin-conjugating enzyme required for proper telomere behavior, Genes Dev (vol. 11), No. 7, pp. 863-75.

Chaudhuri, B. B.; Sarkar, N. and Kundu, P. (1993): Improved fractal geometry based texture segmentation technique., IEE Proc. E (vol. 140), No. 5, pp. 233-241.

Chervitz, S. A.; Hester, E. T.; Ball, C. A.; Dolinski, K.; Dwight, S. S.; Harris, M. A.; Juvik, G.; Malekian, A.; Roberts, S.; Roe, T.; Scafe, C.; Schroeder, M.; Sherlock, G.; Weng, S.; Zhu, Y.; Cherry, J. M. and Botstein, D. (1999): Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure, Nucleic Acids Res (vol. 27), No. 1, pp. 74-8.

Chisholm, R. L.; Gaudet, P.; Just, E. M.; Pilcher, K. E.; Fey, P.; Merchant, S. N. and Kibbe, W. A. (2006): dictyBase, the model organism database for Dictyostelium discoideum, Nucleic Acids Res (vol. 34), No. Database issue, pp. D423-7.

Cirelli, C.; Bushey, D.; Hill, S.; Huber, R.; Kreber, R.; Ganetzky, B. and Tononi, G. (2005): Reduced sleep in Drosophila Shaker mutants, Nature (vol. 434), No. 7037, pp. 1087-92.

Clare, A. and King, R. D. (2002): Machine learning of functional class from phenotype data, Bioinformatics (vol. 18), No. 1, pp. 160-6.

Claustres, M.; Horaitis, O.; Vanevski, M. and Cotton, R. G. (2002): Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases, Genome Res (vol. 12), No. 5, pp. 680-8.

Cohen, M. M., Jr. (2003): The hedgehog signaling network, Am J Med Genet A (vol. 123A), No. 1, pp. 5-28.

Cooper, D. N.; Stenson, P. D. and Chuzhanova, N. A. (2006): The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms, Curr Protoc Bioinformatics (vol. Chapter 1), p. Unit 1 13.

Cooper, S. E. (2007): In vivo function of a novel Siah protein in Drosophila, Mech Dev (vol. 124), No. 7-8, pp. 584-91.

Cooper, S. E.; Murawsky, C. M.; Lowe, N. and Travers, A. A. (2008): Two modes of degradation of the tramtrack transcription factors by Siah homologues, J Biol Chem (vol. 283), No. 2, pp. 1076-83.

Couto, F. M.; Silva, M. J. and Coutinho, P. M. (2007): Measuring semantic similarity between Gene Ontology terms, Data & Knowledge Engineering (vol. 61), No. 1, pp. 137-152.

Couzin, J. (2006): Nobel Prize Awarded for RNAi, ScienceNow Magazine, Accessed: April 2008, http://sciencenow.sciencemag.org/cgi/content/full/2006/1002/1

Crawford, D. C. and Nickerson, D. A. (2005): Definition and clinical importance of haplotypes, Annu Rev Med (vol. 56), pp. 303-20.

Crick, F. (1970): Central dogma of molecular biology, Nature (vol. 227), No. 5258, pp. 561-3.

Cripps, R. M.; Ball, E.; Stark, M.; Lawn, A. and Sparrow, J. C. (1994): Recovery of dominant, autosomal flightless mutants of Drosophila melanogaster and identification of a new gene required for normal muscle structure and function, Genetics (vol. 137), No. 1, pp. 151-64.

Cui, X.; Vinar, T.; Brejova, B.; Shasha, D. and Li, M. (2007): Homology search for genes, Bioinformatics (vol. 23), No. 13, pp. i97-103.

Cummins, R. and O'Riordan, C. (2005): Evolving General Term-Weighting Schemes for Information Retrieval: Tests on Larger Collections Artificial Intelligence Review (vol. 24), No. 3-4, pp. 277-299.

Cusick, M. E.; Yu, H.; Smolyar, A.; Venkatesan, K.; Carvunis, A. R.; Simonis, N.; Rual, J. F.; Borick, H.; Braun, P.; Dreze, M.; Vandenhaute, J.; Galli, M.; Yazaki, J.; Hill, D. E.; Ecker, J. R.; Roth, F. P. and Vidal, M. (2009): Literature-curated protein interaction datasets, Nat Methods (vol. 6), No. 1, pp. 39-46.

Daraselia, N.; Yuryev, A.; Egorov, S.; Mazo, I. and Ispolatov, I. (2007): Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks., BMC Bioinformatics (vol. 8), No. 1, p. 243.

de la Cruz, N.; Bromberg, S.; Pasko, D.; Shimoyama, M.; Twigger, S.; Chen, J.; Chen, C. F.; Fan, C.; Foote, C.; Gopinath, G. R.; Harris, G.; Hughes, A.; Ji, Y.; Jin, W.; Li, D.; Mathis, J.; Nenasheva, N.; Nie, J.; Nigam, R.; Petri, V.; Reilly, D.; Wang, W.; Wu, W.; Zuniga-Meyer, A.; Zhao, L.; Kwitek, A.; Tonellato, P. and Jacob, H. (2005): The Rat Genome Database (RGD): developments towards a phenome database, Nucleic Acids Res (vol. 33), No. Database issue, pp. D485-91.

Dean, M.; Rzhetsky, A. and Allikmets, R. (2001): The human ATP-binding cassette (ABC) transporter superfamily, Genome Res (vol. 11), No. 7, pp. 1156-66.

Delvey, J. and Barbara, R. H. (2005): Hippocrates, Virtual Museum San Jose State University (vol. accessed April 4, 2005). URL: http://www2.sjsu.edu/depts/Museum/hippoc.html

Deng, M.; Zhang, K.; Mehta, S.; Chen, T. and Sun, F. (2003): Prediction of protein function using protein-protein interaction data, J Comput Biol (vol. 10), No. 6, pp. 947-60.

DeVries, H. (1900): Sur la loi de disjonction des Hybrides, Comptes Rendus de 1'Académie des Sciences (vol. 130), pp. 845-847.

dictyBase (2008): An Online Informatics Resource for Dictyostelium, Northwestern University, 2008, accessed: September 2008, http://dictybase.org/Downloads/all-curated-mutants.html

DiSilvestre, D.; Koch, R. and Groffen, J. (1991): Different clinical manifestations of hyperphenylalaninemia in three siblings with identical phenylalanine hydroxylase genes, Am J Hum Genet (vol. 48), No. 5, pp. 1014-6.

Dobritsa, A. A.; van der Goes van Naters, W.; Warr, C. G.; Steinbrecht, R. A. and Carlson, J. R. (2003): Integrating the molecular and cellular basis of odor coding in the Drosophila antenna, Neuron (vol. 37), No. 5, pp. 827-41.

Downward, J. (2004): Use of RNA interference libraries to investigate oncogenic signalling in mammalian cells, Oncogene (vol. 23), No. 51, pp. 8376-83.

DRSC (2008): DRSC: Completed Screens by Topic, Harvard Medical School, 2008, accessed: September 2008, http://flyrnai.org/DRSC-PTO.html

Drysdale, R. A. and Crosby, M. A. (2005): FlyBase: genes and gene models, Nucleic Acids Res (vol. 33), No. Database issue, pp. D390-5.

Dujon, B. (1998): European Functional Analysis Network (EUROFAN) and the functional analysis of the Saccharomyces cerevisiae genome, Electrophoresis (vol. 19), No. 4, pp. 617-24.

Eggert, U. S.; Kiger, A. A.; Richter, C.; Perlman, Z. E.; Perrimon, N.; Mitchison, T. J. and Field, C. M. (2004): Parallel chemical genetic and genome-wide RNAi screens identify cytokinesis inhibitors and targets, PLoS Biol (vol. 2), No. 12, p. e379.

Eisenberg, D.; Marcotte, E. M.; Xenarios, I. and Yeates, T. O. (2000): Protein function in the post-genomic era, Nature (vol. 405), No. 6788, pp. 823-6.

Enright, A. J.; Kunin, V. and Ouzounis, C. A. (2003): Protein families and TRIBES in genome sequence space, Nucleic Acids Res (vol. 31), No. 15, pp. 4632-8.

Estivill, X.; Cox, N. J.; Chanock, S. J.; Kwok, P. Y.; Scherer, S. W. and Brookes, A. J. (2008): SNPs meet CNVs in genome-wide association studies: HGV2007 meeting report, PLoS Genet (vol. 4), No. 4, p. e1000068.

Famili, I.; Forster, J.; Nielsen, J. and Palsson, B. O. (2003): Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network, Proc Natl Acad Sci U S A (vol. 100), No. 23, pp. 13134-9.

Farzan, S. F.; Ascano, M., Jr.; Ogden, S. K.; Sanial, M.; Brigui, A.; Plessis, A. and Robbins, D. J. (2008): Costal2 functions as a kinesin-like protein in the hedgehog signal transduction pathway, Curr Biol (vol. 18), No. 16, pp. 1215-20.

Felsenfeld, G. and Groudine, M. (2003): Controlling the double helix, Nature (vol. 421), No. 6921, pp. 448-53.

Fernandez-Ricaud, L.; Warringer, J.; Ericson, E.; Glaab, K.; Davidsson, P.; Nilsson, F.; Kemp, G. J.; Nerman, O. and Blomberg, A. (2007): PROPHECY--a yeast phenome database, update 2006, Nucleic Acids Res (vol. 35), No. Database issue, pp. D463-7.

Fernandez, A. G.; Gunsalus, K. C.; Huang, J.; Chuang, L. S.; Ying, N.; Liang, H. L.; Tang, C.; Schetter, A. J.; Zegar, C.; Rual, J. F.; Hill, D. E.; Reinke, V.; Vidal, M. and Piano, F. (2005): New genes with roles in the C. elegans embryo revealed using RNAi of ovary-enriched ORFeome clones, Genome Res (vol. 15), No. 2, pp. 250-9.

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.; Graf, S.; Haider, S.; Hammond, M.; Holland, R.; Howe, K. L.; Howe, K.; Johnson, N.; Jenkinson, A.; Kahari, 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.; Kasprzyk, A.; Proctor, G.; Smith, J.; Ureta-Vidal, A. and Searle, S. (2008): Ensembl 2008, Nucleic Acids Research (vol. 36), No. Database issue, pp. D707-14.

Flockhart, I.; Booker, M.; Kiger, A.; Boutros, M.; Armknecht, S.; Ramadan, N.; Richardson, K.; Xu, A.; Perrimon, N. and Mathey-Prevot, B. (2006): FlyRNAi: the Drosophila RNAi screening center database, Nucleic Acids Res (vol. 34), No. Database issue, pp. D489-94.

Foley, E.; O'Farrell, P. H. and Sprenger, F. (1999): Rux is a cyclin-dependent kinase inhibitor (CKI) specific for mitotic cyclin-Cdk complexes, Curr Biol (vol. 9), No. 23, pp. 1392-402.

Frankel, W. N.; Goldowitz, D.; Takahash, J. S. and Yates, C. J. (2008): Neuromice, Evanston, The Jackson Laboratory, The Tennessee Mouse Genome Consortium and The Neurogenomics Project at Northwestern University, accessed: September 2008, http://www.neuromice.org

Fraser, A. G.; Kamath, R. S.; Zipperlen, P.; Martinez-Campos, M.; Sohrmann, M. and Ahringer, J. (2000): Functional genomic analysis of C. elegans chromosome I by systematic RNA interference, Nature (vol. 408), No. 6810, pp. 325-30.

Freimer, N. and Sabatti, C. (2003): The human phenome project, Nat Genet (vol. 34), No. 1, pp. 15-21.

Frezal, J. (1998): Genatlas database, genes and development defects, C R Acad Sci III (vol. 321), No. 10, pp. 805-17.

Friedberg, I. (2006): Automated protein function prediction--the genomic challenge, Briefings in Bioinformatics (vol. 7), No. 3, pp. 225-42.

Friedman, A. and Perrimon, N. (2004): Genome-wide high-throughput screens in functional genomics, Curr Opin Genet Dev (vol. 14), No. 5, pp. 470-6.

Frohlich, H.; Speer, N.; Poustka, A. and Beissbarth, T. (2007): GOSim--an R-package for computation of information theoretic GO similarities between terms and gene products, BMC Bioinformatics (vol. 8), p. 166.

Furey, T. S. (2006): Comparison of human (and other) genome browsers, Hum Genomics (vol. 2), No. 4, pp. 266-70.

Gaulton, K. J.; Mohlke, K. L. and Vision, T. J. (2007): A computational system to select candidate genes for complex human traits, Bioinformatics (vol. 23), No. 9, pp. 1132-40.

GeneTests (2008): GeneTests Homepage, University of Washington, 2008, accessed: September 2008, http://www.geneclinics.org/

Gerstein, M. B.; Bruce, C.; Rozowsky, J. S.; Zheng, D.; Du, J.; Korbel, J. O.; Emanuelsson, O.; Zhang, Z. D.; Weissman, S. and Snyder, M. (2007): What is a gene, post-ENCODE? History and updated definition, Genome Res (vol. 17), No. 6, pp. 669-81.

Ginalski, K. (2006): Comparative modeling for protein structure prediction, Current Opinion in Structural Biology (vol. 16), No. 2, pp. 172-7.

Giunta, K. L.; Jang, J. K.; Manheim, E. A.; Subramanian, G. and McKim, K. S. (2002): subito encodes a kinesin-like protein required for meiotic spindle pole formation in Drosophila melanogaster, Genetics (vol. 160), No. 4, pp. 1489-501.

Goldowitz, D.; Frankel, W. N.; Takahashi, J. S.; Holtz-Vitaterna, M.; Bult, C.; Kibbe, W. A.; Snoddy, J.; Li, Y.; Pretel, S.; Yates, J. and Swanson, D. J. (2004): Large-scale mutagenesis of the mouse to understand the genetic bases of nervous system structure and function, Brain Res Mol Brain Res (vol. 132), No. 2, pp. 105-15.

Goldstein, L. S. and Gunawardena, S. (2000): Flying through the drosophila cytoskeletal genome, J Cell Biol (vol. 150), No. 2, pp. F63-8.

Gonczy, P.; Echeverri, C.; Oegema, K.; Coulson, A.; Jones, S. J.; Copley, R. R.; Duperon, J.; Oegema, J.; Brehm, M.; Cassin, E.; Hannak, E.; Kirkham, M.; Pichler, S.; Flohrs, K.; Goessen, A.; Leidel, S.; Alleaume, A. M.; Martin, C.; Ozlu, N.; Bork, P. and Hyman, A. A. (2000): Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III, Nature (vol. 408), No. 6810, pp. 331-6.

Gonczy, P.; Schnabel, H.; Kaletta, T.; Amores, A. D.; Hyman, T. and Schnabel, R. (1999): Dissection of cell division processes in the one cell stage Caenorhabditis elegans embryo by mutational analysis, J Cell Biol (vol. 144), No. 5, pp. 927-46.

Green, E. C.; Gkoutos, G. V.; Lad, H. V.; Blake, A.; Weekes, J. and Hancock, J. M. (2005): EMPReSS: European mouse phenotyping resource for standardized screens, Bioinformatics (vol. 21), No. 12, pp. 2930-1.

Griffiths, P. E. and Stotz, K. (2006): Genes in the postgenomic era, Theor Med Bioeth (vol. 27), No. 6, pp. 499-521.

Groth, P.; Pavlova, N.; Kalev, I.; Tonov, S.; Georgiev, G.; Pohlenz, H. D. and Weiss, B. (2007): PhenomicDB: a new cross-species genotype/phenotype resource, Nucleic Acids Res (vol. 35), No. Database issue, pp. D696-9.

Groth, P. and Weiss, B. (2006a): Phenotype Data: A Neglected Resource in Biomedical Research?, Current Bioinformatics (vol. 1), No. 3, pp. 347-358.

Groth, P. and Weiss, B. (2006b): Species-übergreifende Ressource für RNAi-Daten und Phänotypen, Laborwelt (vol. 7), No. 6, pp. 14-16.

Groth, P.; Weiss, B.; Pohlenz, H.D. and Leser, U. (2008): Mining phenotypes for gene function prediction., BMC Bioinformatics (vol. 9), No. 1, pp. 136-150.

Grubb, S. C.; Churchill, G. A. and Bogue, M. A. (2004): A collaborative database of inbred mouse strain characteristics, Bioinformatics (vol. 20), No. 16, pp. 2857-9.

Guldener, U.; Munsterkotter, M.; Kastenmuller, G.; Strack, N.; van Helden, J.; Lemer, C.; Richelles, J.; Wodak, S. J.; Garcia-Martinez, J.; Perez-Ortin, J. E.; Michael, H.; Kaps, A.; Talla, E.; Dujon, B.; Andre, B.; Souciet, J. L.; De Montigny, J.; Bon, E.; Gaillardin, C. and Mewes, H. W. (2005): CYGD: the Comprehensive Yeast Genome Database, Nucleic Acids Res (vol. 33), No. Database issue, pp. D364-8.

Gunsalus, K. C.; Ge, H.; Schetter, A. J.; Goldberg, D. S.; Han, J. D.; Hao, T.; Berriz, G. F.; Bertin, N.; Huang, J.; Chuang, L. S.; Li, N.; Mani, R.; Hyman, A. A.; Sonnichsen, B.; Echeverri, C. J.; Roth, F. P.; Vidal, M. and Piano, F. (2005): Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis, Nature (vol. 436), No. 7052, pp. 861-5.

Gunsalus, K. C.; Yueh, W. C.; MacMenamin, P. and Piano, F. (2004): RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects, Nucleic Acids Res (vol. 32), No. Database issue, pp. D406-10.

Guo, X.; Liu, R.; Shriver, C. D.; Hu, H. and Liebman, M. N. (2006): Assessing semantic similarity measures for the characterization of human regulatory pathways, Bioinformatics (vol. 22), No. 8, pp. 967-73.

Han, C.; Belenkaya, T. Y.; Khodoun, M.; Tauchi, M. and Lin, X. (2004): Distinct and collaborative roles of Drosophila EXT family proteins in morphogen signalling and gradient formation, Development (vol. 131), No. 7, pp. 1563-75.

Hancock, J. M.; Adams, N. C.; Aidinis, V.; Blake, A.; Bogue, M.; Brown, S. D.; Chesler, E. J.; Davidson, D.; Duran, C.; Eppig, J. T.; Gailus-Durner, V.; Gates, H.; Gkoutos, G. V.; Greenaway, S.; Hrabe de Angelis, M.; Kollias, G.; Leblanc, S.; Lee, K.; Lengger, C.; Maier, H.; Mallon, A. M.; Masuya, H.; Melvin, D. G.; Muller, W.; Parkinson, H.; Proctor, G.; Reuveni, E.; Schofield, P.; Shukla, A.; Smith, C.; Toyoda, T.; Vasseur, L.; Wakana, S.; Walling, A.; White, J.; Wood, J. and Zouberakis, M. (2007): Mouse Phenotype Database Integration Consortium: integration [corrected] of mouse phenome data resources, Mamm Genome (vol. 18), No. 3, pp. 157-63.

Handl, J.; Knowles, J. and Kell, D. B. (2005): Computational cluster validation in post-genomic data analysis, Bioinformatics (vol. 21), No. 15, pp. 3201-12.

Hannon, G. J. (2002): RNA interference, Nature (vol. 418), No. 6894, pp. 244-51.

HapMapConsortium (2003): The International HapMap Project, Nature (vol. 426), No. 6968, pp. 789-96.

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. and White, R. (2004): The Gene Ontology (GO) database and informatics resource, Nucleic Acids Res (vol. 32), No. Database issue, pp. D258-61.

HGMD (2008): Number of entries in HGMD by type, Cardiff University, 2008, accessed: September 2008, http://www.hgmd.cf.ac.uk/ac/hahaha.php

Hodge, A. E.; Altman, R. B. and Klein, T. E. (2007): The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge, Clin Pharmacol Ther (vol. 81), No. 1, pp. 21-4.

Hong, E. L.; Balakrishnan, R.; Dong, Q.; Christie, K. R.; Park, J.; Binkley, G.; Costanzo, M. C.; Dwight, S. S.; Engel, S. R.; Fisk, D. G.; Hirschman, J. E.; Hitz, B. C.; Krieger, C. J.; Livstone, M. S.; Miyasato, S. R.; Nash, R. S.; Oughtred, R.; Skrzypek, M. S.; Weng, S.; Wong, E. D.; Zhu, K. K.; Dolinski, K.; Botstein, D. and Cherry, J. M. (2008): Gene Ontology annotations at SGD: new data sources and annotation methods, Nucleic Acids Res (vol. 36), No. Database issue, pp. D577-81.

Horaitis, O. and Cotton, R. G. (2004): The challenge of documenting mutation across the genome: the human genome variation society approach, Hum Mutat (vol. 23), No. 5, pp. 447-52.

Hubbard, T. (2002): Biological information: making it accessible and integrated (and trying to make sense of it), Bioinformatics (vol. 18 Suppl 2), p. S140.

Hur, B.; Elisseeff, A. and Guyon, I. (2002): A stability-based method for discovering structure in clustered data, Pacific Symposium on Biocomputing.

Huynen, M. A.; Snel, B.; von Mering, C. and Bork, P. (2003): Function prediction and protein networks, Current Opinion in Cell Biology (vol. 15), No. 2, pp. 191-8.

Ivy, J. M. (1981): Mutations that disrupt meiosis in males of Drosophila melanogaster, Ph.D. Thesis, University of California, San Diego.

Jaeger, S. and Leser, U. (2007): High-Precision Function Prediction using Conserved Interactions, German Conference on Bioinformatics (GCB), Potsdam, Germany

Jang, J. K.; Rahman, T. and McKim, K. S. (2005): The kinesinlike protein Subito contributes to central spindle assembly and organization of the meiotic spindle in Drosophila oocytes, Mol Biol Cell (vol. 16), No. 10, pp. 4684-94.

Ji, J. Y.; Haghnia, M.; Trusty, C.; Goldstein, L. S. and Schubiger, G. (2002): A genetic screen for suppressors and enhancers of the Drosophila cdk1-cyclin B identifies maternal factors that regulate microtubule and microfilament stability, Genetics (vol. 162), No. 3, pp. 1179-95.

Jiang, J.J. and Conrath, D.W. (1997): Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, CoRR (vol. cmp-lg/9709008).

Jizba, R. (2000): Measuring Search Effectiveness, Omaha, Nebraska, USA, Creighton University Health Sciences Library and Learning Resources Center, Accessed: April 2008, http://newadonis.creighton.edu/HSL/searching/Recall-Precision.html

Johannsen, Wilhelm (1911): The Genotype Conception of Heredity, The American Naturalist (vol. Vol. xlv), No. 531, p. 531.

Johnston, L. A. and Gallant, P. (2002): Control of growth and organ size in Drosophila, Bioessays (vol. 24), No. 1, pp. 54-64.

Jones, S. and Thornton, J. M. (2004): Searching for functional sites in protein structures, Current Opinion in Chemical Biology (vol. 8), No. 1, pp. 3-7.

Kahraman, A.; Avramov, A.; Nashev, L. G.; Popov, D.; Ternes, R.; Pohlenz, H. D. and Weiss, B. (2005): PhenomicDB: a multi-species genotype/phenotype database for comparative phenomics, Bioinformatics (vol. 21), No. 3, pp. 418-20.

Kalderon, D. (2002): Similarities between the Hedgehog and Wnt signaling pathways, Trends Cell Biol (vol. 12), No. 11, pp. 523-31.

Kamath, R. S. and Ahringer, J. (2003): Genome-wide RNAi screening in Caenorhabditis elegans, Methods (vol. 30), No. 4, pp. 313-21.

Kamath, R. S.; Fraser, A. G.; Dong, Y.; Poulin, G.; Durbin, R.; Gotta, M.; Kanapin, A.; Le Bot, N.; Moreno, S.; Sohrmann, M.; Welchman, D. P.; Zipperlen, P. and Ahringer, J. (2003): Systematic functional analysis of the Caenorhabditis elegans genome using RNAi, Nature (vol. 421), No. 6920, pp. 231-7.

Karaflou, M.; Lambrinoudaki, I. and Christodoulakos, G. (2008): Apoptosis in atherosclerosis: a mini-review, Mini Rev Med Chem (vol. 8), No. 9, pp. 912-8.

Keller, A. and Vosshall, L. B. (2003): Decoding olfaction in Drosophila, Current Opinion in Neurobiology (vol. 13), No. 1, pp. 103-10.

Kellerman, K. A. and Miller, K. G. (1992): An unconventional myosin heavy chain gene from Drosophila melanogaster, Journal of Cell Biology (vol. 119), No. 4, pp. 823-34.

Kelley, B. P.; Sharan, R.; Karp, R. M.; Sittler, T.; Root, D. E.; Stockwell, B. R. and Ideker, T. (2003): Conserved pathways within bacteria and yeast as revealed by global protein network alignment, Proceedings of the National Academy of Sciences of the United States of America (vol. 100), No. 20, pp. 11394-9.

Kemmer, D.; Huang, Y.; Shah, S. P.; Lim, J.; Brumm, J.; Yuen, M. M.; Ling, J.; Xu, T.; Wasserman, W. W. and Ouellette, B. F. (2005): Ulysses - an application for the projection of molecular interactions across species, Genome Biology (vol. 6), No. 12, p. R106.

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. and Hermjakob, H. (2007): IntAct--open source resource for molecular interaction data, Nucleic Acids Res (vol. 35), No. Database issue, pp. D561-5.

Klare, R. (1997): Gregor Mendel: Father of Genetics, Great Minds of Science, Enslow Publishers, ISBN: 0894907891.

Kolfschoten, I. G.; van Leeuwen, B.; Berns, K.; Mullenders, J.; Beijersbergen, R. L.; Bernards, R.; Voorhoeve, P. M. and Agami, R. (2005): A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity, Cell (vol. 121), No. 6, pp. 849-58.

Kumar, A.; Cheung, K. H.; Tosches, N.; Masiar, P.; Liu, Y.; Miller, P. and Snyder, M. (2002): The TRIPLES database: a community resource for yeast molecular biology, Nucleic Acids Res (vol. 30), No. 1, pp. 73-5.

Kuttenkeuler, D. and Boutros, M. (2004): Genome-wide RNAi as a route to gene function in Drosophila, Brief Funct Genomic Proteomic (vol. 3), No. 2, pp. 168-76.

Lage, K.; Karlberg, E. O.; Storling, Z. M.; Olason, P. I.; Pedersen, A. G.; Rigina, O.; Hinsby, A. M.; Tumer, Z.; Pociot, F.; Tommerup, N.; Moreau, Y. and Brunak, S. (2007): A human phenome-interactome network of protein complexes implicated in genetic disorders, Nature Biotechnology (vol. 25), No. 3, pp. 309-16.

Laskowski, R. A.; Watson, J. D. and Thornton, J. M. (2003): From protein structure to biochemical function?, Journal of Structural & Functional Genomics (vol. 4), No. 2-3, pp. 167-77.

Lee, I.; Lehner, B.; Crombie, C.; Wong, W.; Fraser, A. G. and Marcotte, E. M. (2008): A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans, Nat Genet (vol. 40), No. 2, pp. 181-8.

Lenffer, J.; Nicholas, F. W.; Castle, K.; Rao, A.; Gregory, S.; Poidinger, M.; Mailman, M. D. and Ranganathan, S. (2006): OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI, Nucleic Acids Res (vol. 34), No. Database issue, pp. D599-601.

Lin, D. (1998): An information-theoretic definition of similarity, Proc. 15th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 296-304

Lord, P. W.; Stevens, R. D.; Brass, A. and Goble, C. A. (2003): Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, Bioinformatics (vol. 19), No. 10, pp. 1275-83.

Lussier, Y.; Borlawsky, T.; Rappaport, D.; Liu, Y. and Friedman, C. (2006): PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing, Pacific Symposium on Biocomputing, pp. 64-75.

Lutz, M.W.; Warren, P.V.; Gill, R.W. and Searls, D.B. (2005): Managing genomic and proteomic knowledge, Drug Discovery Today: Technologies (vol. 2), No. 3, pp. 197-204.

Lyne, R.; Smith, R.; Rutherford, K.; Wakeling, M.; Varley, A.; Guillier, F.; Janssens, H.; Ji, W.; McLaren, P.; North, P.; Rana, D.; Riley, T.; Sullivan, J.; Watkins, X.; Woodbridge, M.; Lilley, K.; Russell, S.; Ashburner, M.; Mizuguchi, K. and Micklem, G. (2007): FlyMine: an integrated database for Drosophila and Anopheles genomics, Genome Biol (vol. 8), No. 7, p. R129.

MacKay, D. J. C. (2003): Information Theory, Inference, and Learning Algorithms, Cambridge University Press, Cambridge, ISBN: 0521642981.

Maglott, D.; Ostell, J.; Pruitt, K. D. and Tatusova, T. (2007): Entrez Gene: gene-centered information at NCBI, Nucleic Acids Res (vol. 35), No. Database issue, pp. D26-31.

Mailman, M. D.; Feolo, M.; Jin, Y.; Kimura, M.; Tryka, K.; Bagoutdinov, R.; Hao, L.; Kiang, A.; Paschall, J.; Phan, L.; Popova, N.; Pretel, S.; Ziyabari, L.; Lee, M.; Shao, Y.; Wang, Z. Y.; Sirotkin, K.; Ward, M.; Kholodov, M.; Zbicz, K.; Beck, J.; Kimelman, M.; Shevelev, S.; Preuss, D.; Yaschenko, E.; Graeff, A.; Ostell, J. and Sherry, S. T. (2007): The NCBI dbGaP database of genotypes and phenotypes, Nature Genetics (vol. 39), No. 10, pp. 1181-6.

Manning, B. D. and Cantley, L. C. (2003): Rheb fills a GAP between TSC and TOR, Trends Biochem Sci (vol. 28), No. 11, pp. 573-6.

Manning, C. D.; Raghavan, P. and Schuetze, H. (2008): Introduction to Information Retrieval, Cambridge University Press, Cambridge.

Manske, Magnus (1995): Karyogram of a human female, Human_karyogram.png. University of Washington, Department of Pathology, Cytogenetics Gallery, Accessed: April 2008, http://www.biocrawler.com/encyclopedia/Image:Human_karyogram.png

Markowetz, F.; Bloch, J. and Spang, R. (2005): Non-transcriptional pathway features reconstructed from secondary effects of RNA interference, Bioinformatics (vol. 21), No. 21, pp. 4026-32.

Martinetz, T. M.; Berkovich, S. G. and Schulten, K. J. (1993): 'Neural-gas' network for vector quantization and its application to timeseries prediction., IEEE Transactions on Neural Networks (vol. 4), pp. 558–569.

Mathivanan, S.; Periaswamy, B.; Gandhi, T. K.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y. L. and Pandey, A. (2006): An evaluation of human protein-protein interaction data in the public domain, BMC Bioinformatics (vol. 7 Suppl 5), p. S19.

Mattick, J. S. (2003): Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms, Bioessays (vol. 25), No. 10, pp. 930-9.

Maydanovych, O. and Beal, P. A. (2006): Breaking the central dogma by RNA editing, Chem Rev (vol. 106), No. 8, pp. 3397-411.

McKusick, V. A. (2007): Mendelian Inheritance in Man and its online version, OMIM, Am J Hum Genet (vol. 80), No. 4, pp. 588-604.

Mermall, V. and Miller, K. G. (1995): The 95F unconventional myosin is required for proper organization of the Drosophila syncytial blastoderm, Journal of Cell Biology (vol. 129), No. 6, pp. 1575-88.

Mewes, H. W.; Dietmann, S.; Frishman, D.; Gregory, R.; Mannhaupt, G.; Mayer, K. F.; Munsterkotter, M.; Ruepp, A.; Spannagl, M.; Stumpflen, V. and Rattei, T. (2008): MIPS: analysis and annotation of genome information in 2007, Nucleic Acids Res (vol. 36), No. Database issue, pp. D196-201.

MGI (2008): MGI-Statistics for the Mouse Genome Informatics database resource, The Jackson Laboratory, 2008, accessed: September 2008, http://www.informatics.jax.org/mgihome/homepages/stats/all_stats.shtml

Morales-Mulia, S. and Scholey, J. M. (2005): Spindle pole organization in Drosophila S2 cells by dynein, abnormal spindle protein (Asp), and KLP10A, Mol Biol Cell (vol. 16), No. 7, pp. 3176-86.

Morgan, T.H.; Sturtevant, A.H.; Muller, H.J. and Bridges, C.B. (1915): The Mechanism of Mendelian Heredity, Robbins, Robert. Henry Holt and Company, On-line Electronic Edition by Electronic Scholarly Publishing, Accessed February 2008, http://www.esp.org/books/morgan/mechanism/facsimile

Murali, T. M.; Wu, C. J. and Kasif, S. (2006): The art of gene function prediction, Nat Biotechnol (vol. 24), No. 12, pp. 1474-5; author reply 1475-6.

Needleman, S. B. and Wunsch, C. D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins, J Mol Biol (vol. 48), No. 3, pp. 443-53.

Nelson, S. J.; Schopen, M.; Savage, A. G.; Schulman, J. L. and Arluk, N. (2004): The MeSH Translation Maintenance System: Structure, Interface Design, and Implementation., Fieschi, M., 11th World Congress on Medical Informatics, San Francisco, CA, Amsterdam: IOS Press, 67-69

Neumuller, R. A.; Betschinger, J.; Fischer, A.; Bushati, N.; Poernbacher, I.; Mechtler, K.; Cohen, S. M. and Knoblich, J. A. (2008): Mei-P26 regulates microRNAs and cell growth in the Drosophila ovarian stem cell lineage, Nature (vol. 454), No. 7201, pp. 241-5.

Nicholas, F. W. (2003): Online Mendelian Inheritance in Animals (OMIA): a comparative knowledgebase of genetic disorders and other familial traits in non-laboratory animals, Nucleic Acids Res (vol. 31), No. 1, pp. 275-7.

Nijhout, H. F. (2003): Development and evolution of adaptive polyphenisms, Evol Dev (vol. 5), No. 1, pp. 9-18.

Nongthomba, U. and Ramachandra, N. B. (1999): A direct screen identifies new flight muscle mutants on the Drosophila second chromosome, Genetics (vol. 153), No. 1, pp. 261-74.

Nusse, R. (2003): Wnts and Hedgehogs: lipid-modified proteins and similarities in signaling mechanisms at the cell surface, Development (vol. 130), No. 22, pp. 5297-305.

Nusslein-Volhard, C. and Wieschaus, E. (1980): Mutations affecting segment number and polarity in Drosophila, Nature (vol. 287), No. 5785, pp. 795-801.

O'Brien, T. P. and Frankel, W. N. (2004): Moving forward with chemical mutagenesis in the mouse, J Physiol (vol. 554), No. Pt 1, pp. 13-21.

Oladipo, A.; Cowan, A. and Rodionov, V. (2007): Microtubule motor Ncd induces sliding of microtubules in vivo, Mol Biol Cell (vol. 18), No. 9, pp. 3601-6.

Orr-Weaver, T. L. (1995): Meiosis in Drosophila: seeing is believing, Proc Natl Acad Sci U S A (vol. 92), No. 23, pp. 10443-9.

Page, D. R. and Grossniklaus, U. (2002): The art and design of genetic screens: Arabidopsis thaliana, Nat Rev Genet (vol. 3), No. 2, pp. 124-36.

Page, S. L.; McKim, K. S.; Deneen, B.; Van Hook, T. L. and Hawley, R. S. (2000): Genetic studies of mei-P26 reveal a link between the processes that control germ cell proliferation in both sexes and those that control meiotic exchange in Drosophila, Genetics (vol. 155), No. 4, pp. 1757-72.

Pandey, Gaurav; Kumar, Vipin and Steinbach, Michael (2006): Computational Approaches for Protein Function Prediction: A Survey, Technical Report no. TR 06-028, Minneapolis, MN, Department of Computer Science and Engineering, University of Minnesota, TR 06-028, October 31, 2006

Patrinos, G. P. and Brookes, A. J. (2005): DNA, diseases and databases: disastrously deficient, Trends Genet (vol. 21), No. 6, pp. 333-8.

Pazos, F. and Sternberg, M. J. (2004): Automated prediction of protein function and detection of functional sites from structure, Proceedings of the National Academy of Sciences of the United States of America (vol. 101), No. 41, pp. 14754-9.

Pedersen, T. and A. Kulkarni (2006): Automatic Cluster Stopping with Criterion Functions and the Gap Statistic, Human Language Technology Conference of the NAACL, New York City, USA, Association for Computational Linguistics, 276-279

Pesquita, C.; Faria, D.; Bastos, H.; Ferreira, A. E.; Falcao, A. O. and Couto, F. M. (2008): Metrics for GO based protein semantic similarity: a systematic evaluation, BMC Bioinformatics (vol. 9 Suppl 5), p. S4.

Peters, J. L.; Cnudde, F. and Gerats, T. (2003): Forward genetics and map-based cloning approaches, Trends Plant Sci (vol. 8), No. 10, pp. 484-91.

PharmGKB (2008): The Pharmacogenetics and Pharmacogenomics Knowledge Base, Stanford University, 2008, accessed: September 2008, http://www.pharmgkb.org/

PhenoBlast (2007): RNAiDB and PhenoBlast, Center for Comparative Functional Genomics, New York University, 2008, accessed: September 2008, http://nematoda.bio.nyu.edu:8001/cgi-bin/index.cgi

Piano, F.; Schetter, A. J.; Mangone, M.; Stein, L. and Kemphues, K. J. (2000): RNAi analysis of genes expressed in the ovary of Caenorhabditis elegans, Curr Biol (vol. 10), No. 24, pp. 1619-22.

Piano, F.; Schetter, A. J.; Morton, D. G.; Gunsalus, K. C.; Reinke, V.; Kim, S. K. and Kemphues, K. J. (2002): Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans, Curr Biol (vol. 12), No. 22, pp. 1959-64.

Porter, M.F. (1980): An algorithm for suffix stripping, Program (vol. 14), No. 3, p. 130−137.

Potter, C. J.; Pedraza, L. G. and Xu, T. (2002): Akt regulates growth by directly phosphorylating Tsc2, Nat Cell Biol (vol. 4), No. 9, pp. 658-65.

Prank, K.; Schulze, E.; Eckert, O.; Nattkemper, T. W.; Bettendorf, M.; Maser-Gluth, C.; Sejnowski, T. J.; Grote, A.; Penner, E.; von Zur Muhlen, A. and Brabant, G. (2005): Machine learning approaches for phenotype-genotype mapping: predicting heterozygous mutations in the CYP21B gene from steroid profiles, Eur J Endocrinol (vol. 153), No. 2, pp. 301-305.

Pruitt, K. D.; Tatusova, T.; Klimke, W. and Maglott, D. R. (2009): NCBI Reference Sequences: current status, policy and new initiatives, Nucleic Acids Res (vol. 37), No. Database issue, pp. D32-6.

Punta, M.; Forrest, L. R.; Bigelow, H.; Kernytsky, A.; Liu, J. and Rost, B. (2007): Membrane protein prediction methods, Methods (Duluth) (vol. 41), No. 4, pp. 460-74.

Rankinen, T.; Bray, M. S.; Hagberg, J. M.; Perusse, L.; Roth, S. M.; Wolfarth, B. and Bouchard, C. (2006): The human gene map for performance and health-related fitness phenotypes: the 2005 update, Medicine & Science in Sports & Exercise (vol. 38), No. 11, pp. 1863-88.

Raychaudhuri, S.; Chang, J. T.; Sutphin, P. D. and Altman, R. B. (2002): Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature, Genome Research (vol. 12), No. 1, pp. 203-14.

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 (vol. 11), pp. 95-130.

Reumers, J.; Conde, L.; Medina, I.; Maurer-Stroh, S.; Van Durme, J.; Dopazo, J.; Rousseau, F. and Schymkowitz, J. (2008): Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases, Nucleic Acids Res (vol. 36), No. Database issue, pp. D825-9.

Reumers, J.; Maurer-Stroh, S.; Schymkowitz, J. and Rousseau, F. (2006): SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs, Bioinformatics (vol. 22), No. 17, pp. 2183-5.

RGD (2007): Rat Genome Database Yearly Report, Bioinformatics Program, HMGC at the Medical College of Wisconsin, 2008, accessed: September 2008, http://labs.rgd.mcw.edu/files/Rat%20Genome%20Database%20Yearly%20Report%20Sept%202007.doc

Rheinberger, H.J. (1995): When did Carl Correns read Gregor Mendel's paper?, Isis (vol. 86), pp. 612-616.

Riley, R.; Lee, C.; Sabatti, C. and Eisenberg, D. (2005): Inferring protein domain interactions from databases of interacting proteins, Genome Biology (vol. 6), No. 10, p. R89.

Riparbelli, M. G.; Massarelli, C.; Robbins, L. G. and Callaini, G. (2004): The abnormal spindle protein is required for germ cell mitosis and oocyte differentiation during Drosophila oogenesis, Exp Cell Res (vol. 298), No. 1, pp. 96-106.

Rison, S. C.; Hodgman, T. C. and Thornton, J. M. (2000): Comparison of functional annotation schemes for genomes, Functional & Integrative Genomics (vol. 1), No. 1, pp. 56-69.

RNAiDB (2009): RNAi database for C. elegans, Center for Comparative Functional Genomics, New York University, 2009, accessed: February 2009, http://www.rnai.org

Robinson, P. N.; Kohler, S.; Bauer, S.; Seelow, D.; Horn, D. and Mundlos, S. (2008): The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease, Am J Hum Genet (vol. 83), No. 5, pp. 610-5.

Rodin, A.; Mosley, T. H., Jr.; Clark, A. G.; Sing, C. F. and Boerwinkle, E. (2005): Mining genetic epidemiology data with Bayesian networks application to APOE gene variation and plasma lipid levels, J Comput Biol (vol. 12), No. 1, pp. 1-11.

Rogers, A.; Antoshechkin, I.; Bieri, T.; Blasiar, D.; Bastiani, C.; Canaran, P.; Chan, J.; Chen, W. J.; Davis, P.; Fernandes, J.; Fiedler, T. J.; Han, M.; Harris, T. W.; Kishore, R.; Lee, R.; McKay, S.; Muller, H. M.; Nakamura, C.; Ozersky, P.; Petcherski, A.; Schindelman, G.; Schwarz, E. M.; Spooner, W.; Tuli, M. A.;

Van Auken, K.; Wang, D.; Wang, X.; Williams, G.; Yook, K.; Durbin, R.; Stein, L. D.; Spieth, J. and Sternberg, P. W. (2008): WormBase 2007, Nucleic Acids Res (vol. 36), No. Database issue, pp. D612-7.

Roman, G. and Davis, R. L. (2001): Molecular biology and anatomy of Drosophila olfactory associative learning, Bioessays (vol. 23), No. 7, pp. 571-81.

Rost, B.; Liu, J.; Nair, R.; Wrzeszczynski, K. O. and Ofran, Y. (2003): Automatic prediction of protein function, Cell Mol Life Sci (vol. 60), No. 12, pp. 2637-50.

Roux-Rouquie, M.; Chauvet, M. L.; Munnich, A. and Frezal, J. (1999): Human genes involved in chromatin remodeling in transcription initiation, and associated diseases: An overview using the GENATLAS database, Mol Genet Metab (vol. 67), No. 4, pp. 261-77.

Rual, J. F.; Ceron, J.; Koreth, J.; Hao, T.; Nicot, A. S.; Hirozane-Kishikawa, T.; Vandenhaute, J.; Orkin, S. H.; Hill, D. E.; van den Heuvel, S. and Vidal, M. (2004): Toward improving Caenorhabditis elegans phenome mapping with an ORFeome-based RNAi library, Genome Res (vol. 14), No. 10B, pp. 2162-8.

Rubin, G. M.; Yandell, M. D.; Wortman, J. R.; Gabor Miklos, G. L.; Nelson, C. R.; Hariharan, I. K.; Fortini, M. E.; Li, P. W.; Apweiler, R.; Fleischmann, W.; Cherry, J. M.; Henikoff, S.; Skupski, M. P.; Misra, S.; Ashburner, M.; Birney, E.; Boguski, M. S.; Brody, T.; Brokstein, P.; Celniker, S. E.; Chervitz, S. A.; Coates, D.; Cravchik, A.; Gabrielian, A.; Galle, R. F.; Gelbart, W. M.; George, R. A.; Goldstein, L. S.; Gong, F.; Guan, P.; Harris, N. L.; Hay, B. A.; Hoskins, R. A.; Li, J.; Li, Z.; Hynes, R. O.; Jones, S. J.; Kuehl, P. M.; Lemaitre, B.; Littleton, J. T.; Morrison, D. K.; Mungall, C.; O'Farrell, P. H.; Pickeral, O. K.; Shue, C.; Vosshall, L. B.; Zhang, J.; Zhao, Q.; Zheng, X. H. and Lewis, S. (2000): Comparative genomics of the eukaryotes, Science (vol. 287), No. 5461, pp. 2204-15.

Ruel, L.; Gallet, A.; Raisin, S.; Truchi, A.; Staccini-Lavenant, L.; Cervantes, A. and Therond, P. P. (2007): Phosphorylation of the atypical kinesin Costal2 by the kinase Fused induces the partial disassembly of the Smoothened-Fused-Costal2-Cubitus interruptus complex in Hedgehog signalling, Development (vol. 134), No. 20, pp. 3677-89.

Rusan, N. M. and Peifer, M. (2007): A role for a novel centrosome cycle in asymmetric cell division, J Cell Biol (vol. 177), No. 1, pp. 13-20.

Ryoo, H. D.; Bergmann, A.; Gonen, H.; Ciechanover, A. and Steller, H. (2002): Regulation of Drosophila IAP1 degradation and apoptosis by reaper and ubcD1, Nat Cell Biol (vol. 4), No. 6, pp. 432-8.

SanJuan, E. and Ibekwe-SanJuan, F. (2006): Text mining without document context, Information Processing & Management (vol. 42), No. 6, pp. 1532-52

Schaffer, A. A.; Aravind, L.; Madden, T. L.; Shavirin, S.; Spouge, J. L.; Wolf, Y. I.; Koonin, E. V. and Altschul, S. F. (2001): Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic Acids Research (vol. 29), No. 14, pp. 2994-3005.

Schoof, H.; Ernst, R.; Nazarov, V.; Pfeifer, L.; Mewes, H. W. and Mayer, K. F. (2004): MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics, Nucleic Acids Res (vol. 32), No. Database issue, pp. D373-6.

Schupbach, T. and Wieschaus, E. (1989): Female sterile mutations on the second chromosome of Drosophila melanogaster. I. Maternal effect mutations, Genetics (vol. 121), No. 1, pp. 101-17.

Schuster-Bockler, B. and Bateman, A. (2007): Reuse of structural domain-domain interactions in protein networks, BMC Bioinformatics (vol. 8), p. 259.

Schwarz, E. M.; Antoshechkin, I.; Bastiani, C.; Bieri, T.; Blasiar, D.; Canaran, P.; Chan, J.; Chen, N.; Chen, W. J.; Davis, P.; Fiedler, T. J.; Girard, L.; Harris, T. W.; Kenny, E. E.; Kishore, R.; Lawson, D.; Lee, R.; Muller, H. M.; Nakamura, C.; Ozersky, P.; Petcherski, A.; Rogers, A.; Spooner, W.; Tuli, M. A.; Van Auken, K.; Wang, D.; Durbin, R.; Spieth, J.; Stein, L. D. and Sternberg, P. W. (2006): WormBase: better software, richer content, Nucleic Acids Res (vol. 34), No. Database issue, pp. D475-8.

Schwikowski, B.; Uetz, P. and Fields, S. (2000): A network of protein-protein interactions in yeast., Nature Biotechnology (vol. 18), No. 12, pp. 1257-61.

Scriver, C. R. (2004): After the genome--the phenome?, Journal of Inherited Metabolic Disease (vol. 27), No. 3, pp. 305-17.

Scriver, C. R.; Eisensmith, R. C.; Woo, S. L. and Kaufman, S. (1994): The hyperphenylalaninemias of man and mouse, Annu Rev Genet (vol. 28), pp. 141-65.

Scriver, C. R.; Hurtubise, M.; Konecki, D.; Phommarinh, M.; Prevost, L.; Erlandsen, H.; Stevens, R.; Waters, P. J.; Ryan, S.; McDonald, D. and Sarkissian, C. (2003): PAHdb 2003: what a locus-specific knowledge-base can do, Hum Mutat (vol. 21), No. 4, pp. 333-44.

Scriver, C. R. and Waters, P. J. (1999): Monogenic traits are not simple: lessons from phenylketonuria, Trends Genet (vol. 15), No. 7, pp. 267-72.

Scriver, C. R.; Waters, P. J.; Sarkissian, C.; Ryan, S.; Prevost, L.; Cote, D.; Novak, J.; Teebi, S. and Nowacki, P. M. (2000): PAHdb: a locus-specific knowledgebase, Hum Mutat (vol. 15), No. 1, pp. 99-104.

Searls, D. B. (2005): Data integration: challenges for drug discovery, Nat Rev Drug Discov (vol. 4), No. 1, pp. 45-58.

Sezen, Tug (2000): Homologous sequences, orthologs3.gif. Folding@Home Educational Project, Accessed: April 2008, http://www.stanford.edu/group/pandegroup/folding/education/orthologs3.gif

Shastry, B. S. (2003): SNPs and haplotypes: genetic markers for disease and drug response (review), Int J Mol Med (vol. 11), No. 3, pp. 379-82.

Shi, T. L.; Li, Y. X.; Cai, Y. D. and Chou, K. C. (2005): Computational methods for protein-protein interaction and their application, Current Protein & Peptide Science (vol. 6), No. 5, pp. 443-9.

Shi, Y. (2003): Mammalian RNAi for the masses, Trends Genet (vol. 19), No. 1, pp. 9-12.

Simmer, F.; Moorman, C.; van der Linden, A. M.; Kuijk, E.; van den Berghe, P. V.; Kamath, R. S.; Fraser, A. G.; Ahringer, J. and Plasterk, R. H. (2003): Genome-wide RNAi of C. elegans using the hypersensitive rrf-3 strain reveals novel gene functions, PLoS Biol (vol. 1), No. 1, p. E12.

Skold, H. N.; Komma, D. J. and Endow, S. A. (2005): Assembly pathway of the anastral Drosophila oocyte meiosis I spindle, J Cell Sci (vol. 118), No. Pt 8, pp. 1745-55.

Smith, B.; Williams, J. and Schulze-Kremer, S. (2003): The ontology of the gene ontology, AMIA Annu Symp Proc, pp. 609-13.

Smith, C. L.; Goldsmith, C. A. and Eppig, J. T. (2005): The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information, Genome Biol (vol. 6), No. 1, p. R7.

Smith, T. F. and Waterman, M. S. (1981): Identification of common molecular subsequences, J Mol Biol (vol. 147), No. 1, pp. 195-7.

Sonnichsen, B.; Koski, L. B.; Walsh, A.; Marschall, P.; Neumann, B.; Brehm, M.; Alleaume, A. M.; Artelt, J.; Bettencourt, P.; Cassin, E.; Hewitson, M.; Holz, C.; Khan, M.; Lazik, S.; Martin, C.; Nitzsche, B.; Ruer, M.; Stamford, J.; Winzi, M.; Heinkel, R.; Roder, M.; Finell, J.; Hantsch, H.; Jones, S. J.; Jones, M.; Piano, F.; Gunsalus, K. C.; Oegema, K.; Gonczy, P.; Coulson, A.; Hyman, A. A. and Echeverri, C. J. (2005): Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans, Nature (vol. 434), No. 7032, pp. 462-9.

Spirin, V. and Mirny, L. A. (2003): Protein complexes and functional modules in molecular networks, Proceedings of the National Academy of Sciences of the United States of America (vol. 100), No. 21, pp. 12123-8.

Sprague, J.; Bayraktaroglu, L.; Bradford, Y.; Conlin, T.; Dunn, N.; Fashena, D.; Frazer, K.; Haendel, M.; Howe, D. G.; Knight, J.; Mani, P.; Moxon, S. A.; Pich, C.; Ramachandran, S.; Schaper, K.; Segerdell, E.; Shao, X.; Singer, A.; Song, P.; Sprunger, B.; Van Slyke, C. E. and Westerfield, M. (2008): The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes, Nucleic Acids Res (vol. 36), No. Database issue, pp. D768-72.

Stark, C.; Breitkreutz, B. J.; Reguly, T.; Boucher, L.; Breitkreutz, A. and Tyers, M. (2006): BioGRID: a general repository for interaction datasets, Nucleic Acids Research (vol. 34), No. Database issue, pp. D535-9.

Stegman, M. A.; Vallance, J. E.; Elangovan, G.; Sosinski, J.; Cheng, Y. and Robbins, D. J. (2000): Identification of a tetrameric hedgehog signaling complex, J Biol Chem (vol. 275), No. 29, pp. 21809-12.

Stein, L. D. (2003): Integrating biological databases, Nat Rev Genet (vol. 4), No. 5, pp. 337-45.

Steinbach, M.; Karypis, G. and Kumar, V. (2000): A Comparison of Document Clustering Techniques, KDD Workshop on Text Mining, http://rakaposhi.eas.asu.edu/cse494/notes/clustering-doccluster.pdf

Stephens, P.; Edkins, S.; Davies, H.; Greenman, C.; Cox, C.; Hunter, C.; Bignell, G.; Teague, J.; Smith, R.; Stevens, C.; O'Meara, S.; Parker, A.; Tarpey, P.; Avis, T.; Barthorpe, A.; Brackenbury, L.; Buck, G.; Butler, A.; Clements, J.; Cole, J.; Dicks, E.; Edwards, K.; Forbes, S.; Gorton, M.; Gray, K.; Halliday, K.; Harrison, R.; Hills, K.; Hinton, J.; Jones, D.; Kosmidou, V.; Laman, R.; Lugg, R.; Menzies, A.; Perry, J.; Petty, R.; Raine, K.; Shepherd, R.; Small, A.; Solomon, H.; Stephens, Y.; Tofts, C.; Varian, J.; Webb, A.; West, S.; Widaa, S.; Yates, A.; Brasseur, F.; Cooper, C. S.; Flanagan, A. M.; Green, A.; Knowles, M.; Leung, S. Y.; Looijenga, L. H.; Malkowicz, B.; Pierotti, M. A.; Teh, B.; Yuen, S. T.; Nicholson, A. G.; Lakhani, S.; Easton, D. F.; Weber, B. L.; Stratton, M. R.; Futreal, P. A. and Wooster, R. (2005): A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer, Nat Genet (vol. 37), No. 6, pp. 590-2.

Swan, A. and Schupbach, T. (2007): The Cdc20 (Fzy)/Cdh1-related protein, Cort, cooperates with Fzy in cyclin destruction and anaphase progression in meiosis I and II in Drosophila, Development (vol. 134), No. 5, pp. 891-9.

Tagarelli, A. and Karypis, G. (2008): A Segment-based Approach To Clustering Multi-Topic Documents., Text Mining Workshop, SIAM Datamining Conference

Tao, Y.; Sam, L.; Li, J.; Friedman, C. and Lussier, Y. A. (2007): Information theory applied to the sparse gene ontology annotation network to predict novel gene function, Bioinformatics (vol. 23), No. 13, pp. i529-38.

Tapon, N.; Ito, N.; Dickson, B. J.; Treisman, J. E. and Hariharan, I. K. (2001): The Drosophila tuberous sclerosis complex gene homologs restrict cell growth and cell proliferation, Cell (vol. 105), No. 3, pp. 345-55.

Terada, Y.; Uetake, Y. and Kuriyama, R. (2003): Interaction of Aurora-A and centrosomin at the microtubule-nucleating site in Drosophila and mammalian cells, J Cell Biol (vol. 162), No. 5, pp. 757-63.

TheBioGrid (2009): General repository for Interaction datasets, Toronto, Tyers Lab, The Samuel Lunenfeld Research Institute, Mount Sinai Hospital, 2009, accessed: February 2009, http://www.thebiogrid.org/SearchResults/summary/67682

Theodoratos, D. and Sellis, T. (2000): Incremental Design of a Data Warehouse, Journal of intelligent information systems (vol. 15), pp. 7-28.

Tibshirani, R.; Walther, G. and Hastie, T. (2001): Estimating the number of clusters in a dataset via the Gap statistic, Journal of the Royal Statistics Society (vol. Series B), pp. 411–23.

Treacy, E.; Pitt, J. J.; Seller, K.; Thompson, G. N.; Ramus, S. and Cotton, R. G. (1996): In vivo disposal of phenylalanine in phenylketonuria: a study of two siblings, J Inherit Metab Dis (vol. 19), No. 5, pp. 595-602.

Troyanskaya, O. G.; Dolinski, K.; Owen, A. B.; Altman, R. B. and Botstein, D. (2003): A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae), Proc Natl Acad Sci U S A (vol. 100), No. 14, pp. 8348-53.

Tschermak, E. (1900): Ueber kuenstliche Kreuzung bei Pisum sativum, Berichte der deutschen Botanischen Gesellschaft (vol. XVIII), No. 6, pp. 232-39.

Tuschl, T. (2003): Functional genomics: RNA sets the standard, Nature (vol. 421), No. 6920, pp. 220-1.

Tuschl, T. and Borkhardt, A. (2002): Small interfering RNAs: a revolutionary tool for the analysis of gene function and gene therapy, Mol Interv (vol. 2), No. 3, pp. 158-67.

Tweedie, S.; Ashburner, M.; Falls, K.; Leyland, P.; McQuilton, P.; Marygold, S.; Millburn, G.; Osumi-Sutherland, D.; Schroeder, A.; Seal, R. and Zhang, H. (2009): FlyBase: enhancing Drosophila Gene Ontology annotations, Nucleic Acids Res (vol. 37), No. Database issue, pp. D555-9.

Twigger, S. N.; Pasko, D.; Nie, J.; Shimoyama, M.; Bromberg, S.; Campbell, D.; Chen, J.; dela Cruz, N.; Fan, C.; Foote, C.; Harris, G.; Hickmann, B.; Ji, Y.; Jin, W.; Li, D.; Mathis, J.; Nenasheva, N.; Nigam, R.; Petri, V.; Reilly, D.; Ruotti, V.; Schauberger, E.; Seiler, K.; Slyper, R.; Smith, J.; Wang, W.; Wu, W.; Zhao, L.; Zuniga-Meyer, A.; Tonellato, P. J.; Kwitek, A. E. and Jacob, H. J. (2005): Tools and strategies for physiological genomics: the Rat Genome Database, Physiol Genomics (vol. 23), No. 2, pp. 246-56.

UniProtConsortium (2009): The Universal Protein Resource (UniProt) 2009, Nucleic Acids Res (vol. 37), No. Database issue, pp. D169-74.

van de Peppel, J.; Kettelarij, N.; van Bakel, H.; Kockelkorn, T. T.; van Leenen, D. and Holstege, F. C. (2005): Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets, Mol Cell (vol. 19), No. 4, pp. 511-22.

van Driel, M. A.; Bruggeman, J.; Vriend, G.; Brunner, H. G. and Leunissen, J. A. (2006): A text-mining analysis of the human phenome, Eur J Hum Genet (vol. 14), No. 5, pp. 535-42.

Van Driessche, N.; Demsar, J.; Booth, E. O.; Hill, P.; Juvan, P.; Zupan, B.; Kuspa, A. and Shaulsky, G. (2005): Epistasis analysis with global transcriptional phenotypes, Nat Genet (vol. 37), No. 5, pp. 471-7.

Vapnik, V. and Chapelle, O. (2000): Bounds on error expectation for support vector machines, Neural Comput (vol. 12), No. 9, pp. 2013-36.

Varmark, H.; Llamazares, S.; Rebollo, E.; Lange, B.; Reina, J.; Schwarz, H. and Gonzalez, C. (2007): Asterless is a centriolar protein required for centrosome function and embryo development in Drosophila, Curr Biol (vol. 17), No. 20, pp. 1735-45.

von Mering, C.; Jensen, L. J.; Kuhn, M.; Chaffron, S.; Doerks, T.; Kruger, B.; Snel, B. and Bork, P. (2007): STRING 7--recent developments in the integration and prediction of protein interactions, Nucleic Acids Research (vol. 35), No. Database issue, pp. D358-62.

Vosshall, L. B.; Wong, A. M. and Axel, R. (2000): An olfactory sensory map in the fly brain, Cell (vol. 102), No. 2, pp. 147-59.

Wain, H. M.; Bruford, E. A.; Lovering, R. C.; Lush, M. J.; Wright, M. W. and Povey, S. (2002): Guidelines for human gene nomenclature, Genomics (vol. 79), No. 4, pp. 464-70.

Walhout, A. J.; Reboul, J.; Shtanko, O.; Bertin, N.; Vaglio, P.; Ge, H.; Lee, H.; Doucette-Stamm, L.; Gunsalus, K. C.; Schetter, A. J.; Morton, D. G.; Kemphues, K. J.; Reinke, V.; Kim, S. K.; Piano, F. and Vidal, M. (2002): Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline, Curr Biol (vol. 12), No. 22, pp. 1952-8.

Wang, J.Z.; Du, Z.; Payattakool, R.; Yu, P.S. and Chen, C.F. (2007): A new method to measure the semantic similarity of GO terms., Bioinformatics (vol. 23), No. 10, pp. 1274-81.

Wang, Z. and Lin, H. (2005): The division of Drosophila germline stem cells and their precursors requires a specific cyclin, Curr Biol (vol. 15), No. 4, pp. 328-33.

Waters, P. J. (2003): How PAH gene mutations cause hyper-phenylalaninemia and why mechanism matters: insights from in vitro expression, Hum Mutat (vol. 21), No. 4, pp. 357-69.

Watson, J. D. and Crick, F. H. (1953): Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid, Nature (vol. 171), No. 4356, pp. 737-8.

Weatherall, D. (1999): From genotype to phenotype: genetics and medical practice in the new millennium, Philos Trans R Soc Lond B Biol Sci (vol. 354), No. 1392, pp. 1995-2010.

Werner, E. (2005): Genome semantics, in silico multicellular systems and the Central Dogma, FEBS Lett (vol. 579), No. 8, pp. 1779-82.

West, D. B. (1999): Introduction to Graph Theory, 2nd. ed., Prentice Hall, ISBN: 0130144002.

Westbrook, T. F.; Martin, E. S.; Schlabach, M. R.; Leng, Y.; Liang, A. C.; Feng, B.; Zhao, J. J.; Roberts, T. M.; Mandel, G.; Hannon, G. J.; Depinho, R. A.; Chin, L. and Elledge, S. J. (2005): A genetic screen for candidate tumor suppressors identifies REST, Cell (vol. 121), No. 6, pp. 837-48.

Wheeler, D. B.; Carpenter, A. E. and Sabatini, D. M. (2005): Cell microarrays and RNA interference chip away at gene function, Nat Genet (vol. 37 Suppl), pp. S25-30.

Wheeler, D. L.; Barrett, T.; Benson, D. A.; Bryant, S. H.; Canese, K.; Chetvernin, V.; Church, D. M.; Dicuccio, M.; Edgar, R.; Federhen, S.; Feolo, M.; Geer, L. Y.; Helmberg, W.; Kapustin, Y.; Khovayko, O.; Landsman, D.; Lipman, D. J.; Madden, T. L.; Maglott, D. R.; Miller, V.; Ostell, J.; Pruitt, K. D.; Schuler, G. D.; Shumway, M.; Sequeira, E.; Sherry, S. T.; Sirotkin, K.; Souvorov, A.; Starchenko, G.; Tatusov, R. L.; Tatusova, T. A.; Wagner, L. and Yaschenko, E. (2008): Database resources of the National Center for Biotechnology Information, Nucleic Acids Res (vol. 36), No. Database issue, pp. D13-21.

Whisstock, J. C. and Lesk, A. M. (2003): Prediction of protein function from protein sequence and structure, Quarterly Reviews of Biophysics (vol. 36), No. 3, pp. 307-40.

Wilson, R. J.; Goodman, J. L. and Strelets, V. B. (2008): FlyBase: integration and improvements to query tools, Nucleic Acids Res (vol. 36), No. Database issue, pp. D588-93.

WormNet (2009): Probabilistic functional gene network of C. elegans, Austin, Institute for Cellular and Molecular Biology, University of Texas, 2009, accessed: February 2009, http://www.functionalnet.org/wormnet/

Wu, H.; Su, Z.; Mao, F.; Olman, V. and Xu, Y. (2005): Prediction of functional modules based on comparative genome analysis and Gene Ontology application, Nucleic Acids Res (vol. 33), No. 9, pp. 2822-37.

Xiao, G. and Pan, W. (2005): Gene function prediction by a combined analysis of gene expression data and protein-protein interaction data, J Bioinform Comput Biol (vol. 3), No. 6, pp. 1371-89.

Yan, M. and Ye, K. (2007): Determining the number of clusters using the weighted gap statistic, Biometrics (vol. 63), No. 4, pp. 1031-7.

ZFIN (2008): The Zebrafish Information Network Data Download, University of Oregon, 2008, accessed: September 2008, http://zfin.org/zf_info/downloads.html#phenotype

Zhao, Y. and Karypis, G. (2002): Criterion Functions for Document Clustering, University of Minnesota, Department of Computer Science / Army HPC Research Center, Minneapolis, Technical Report no. 01-40

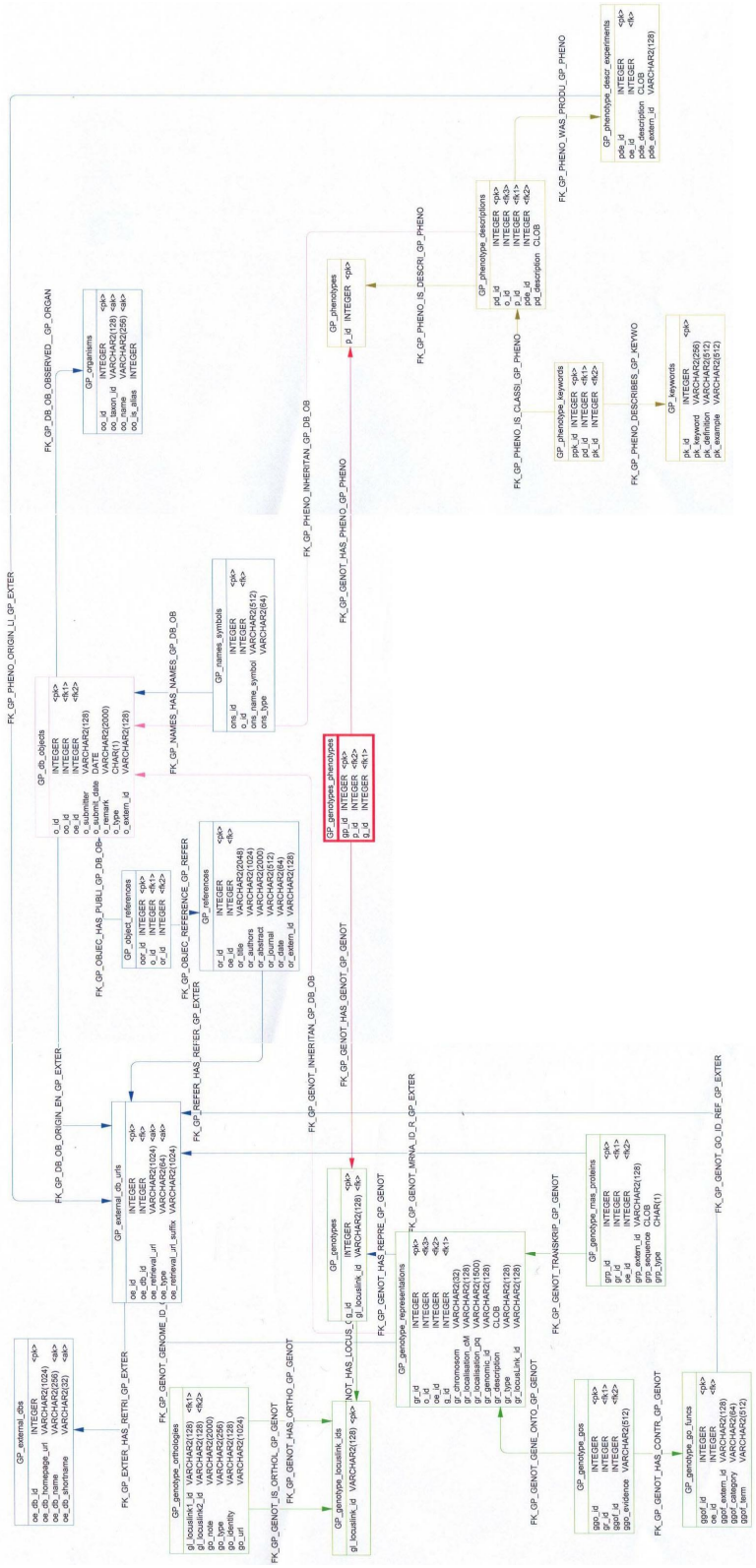
Zhao, Y. and Karypis, G. (2003): Clustering in life sciences, Methods Mol Biol (vol. 224), pp. 183-218.

Zhao, Y. and Karypis, G. (2005): Data clustering in life sciences, Mol Biotechnol (vol. 31), No. 1, pp. 55-80.

Appendix

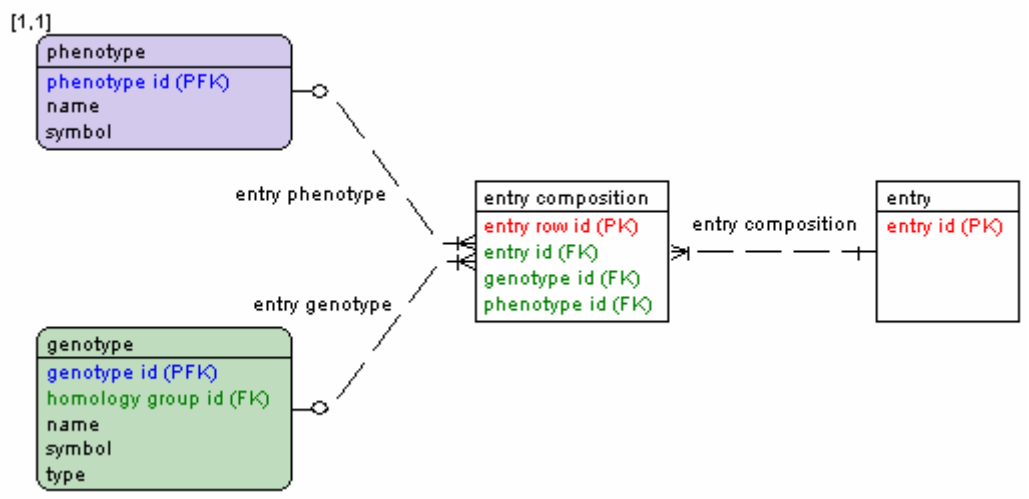
A1 PhenomicDB: Database schemes

A1.1 Database scheme of MSP

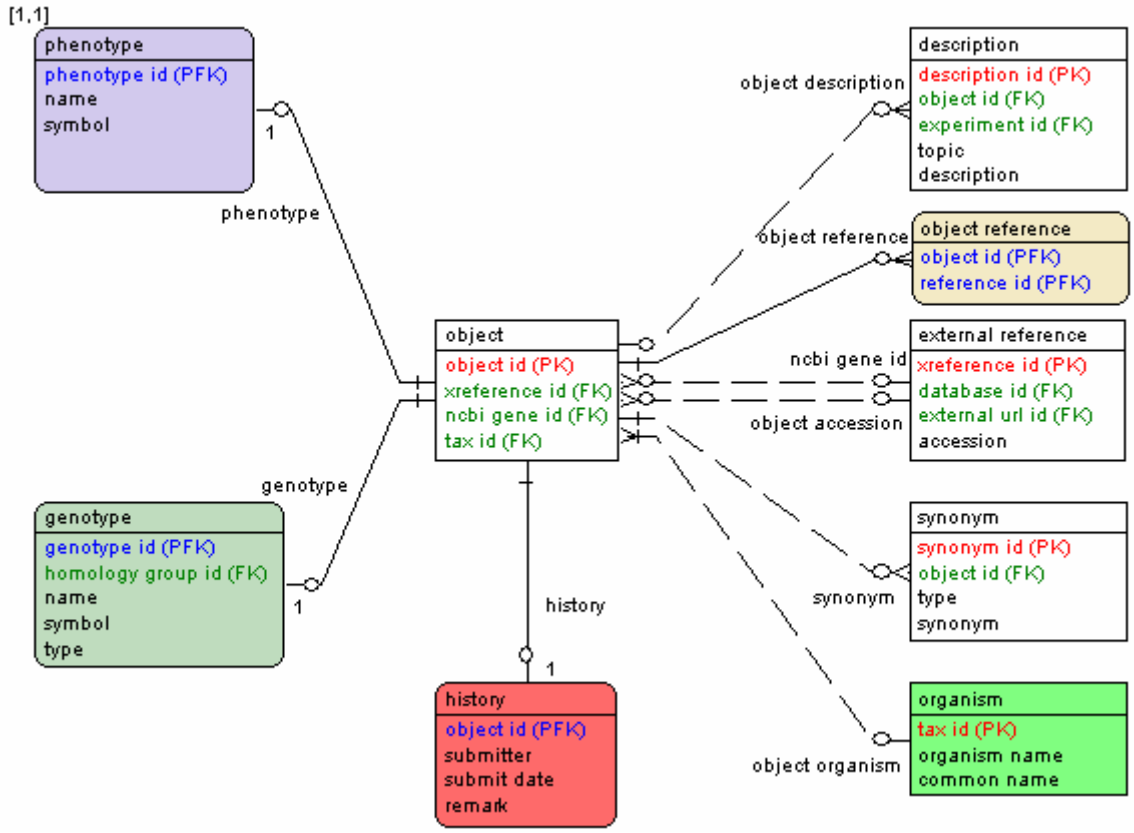


A1.2 Database schemes of PhenomicDB version 1.x

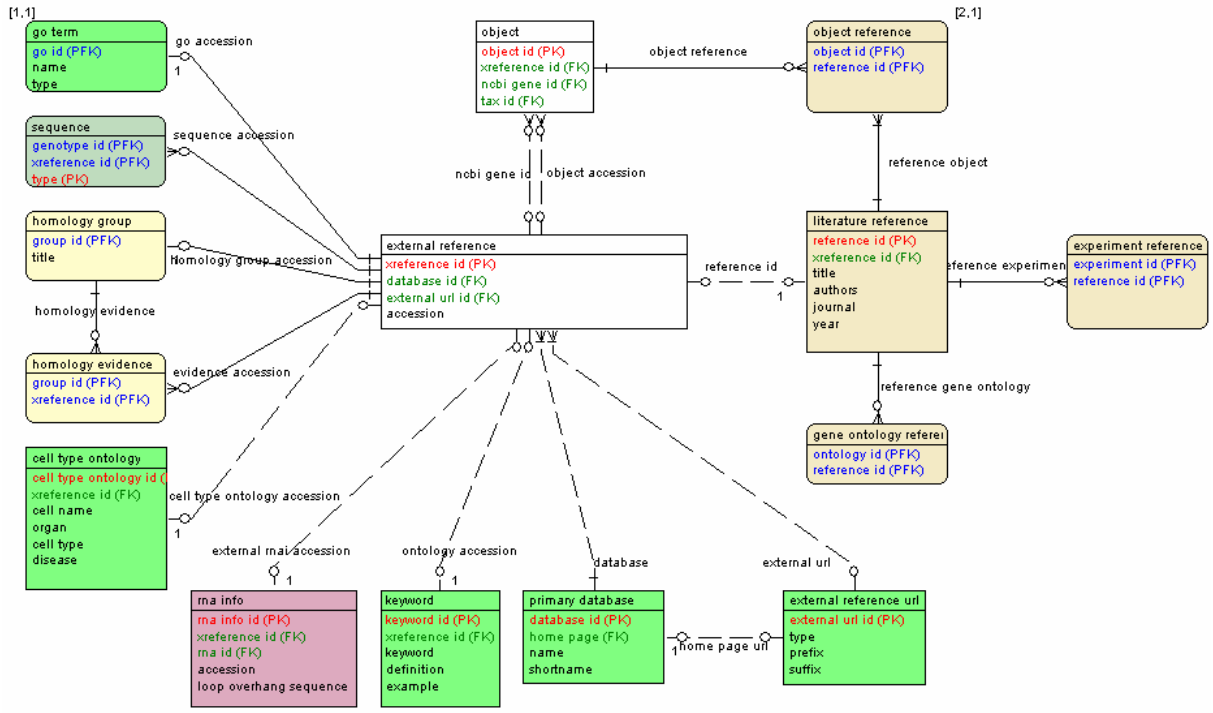
A1.2.1 Database scheme 'Entry'



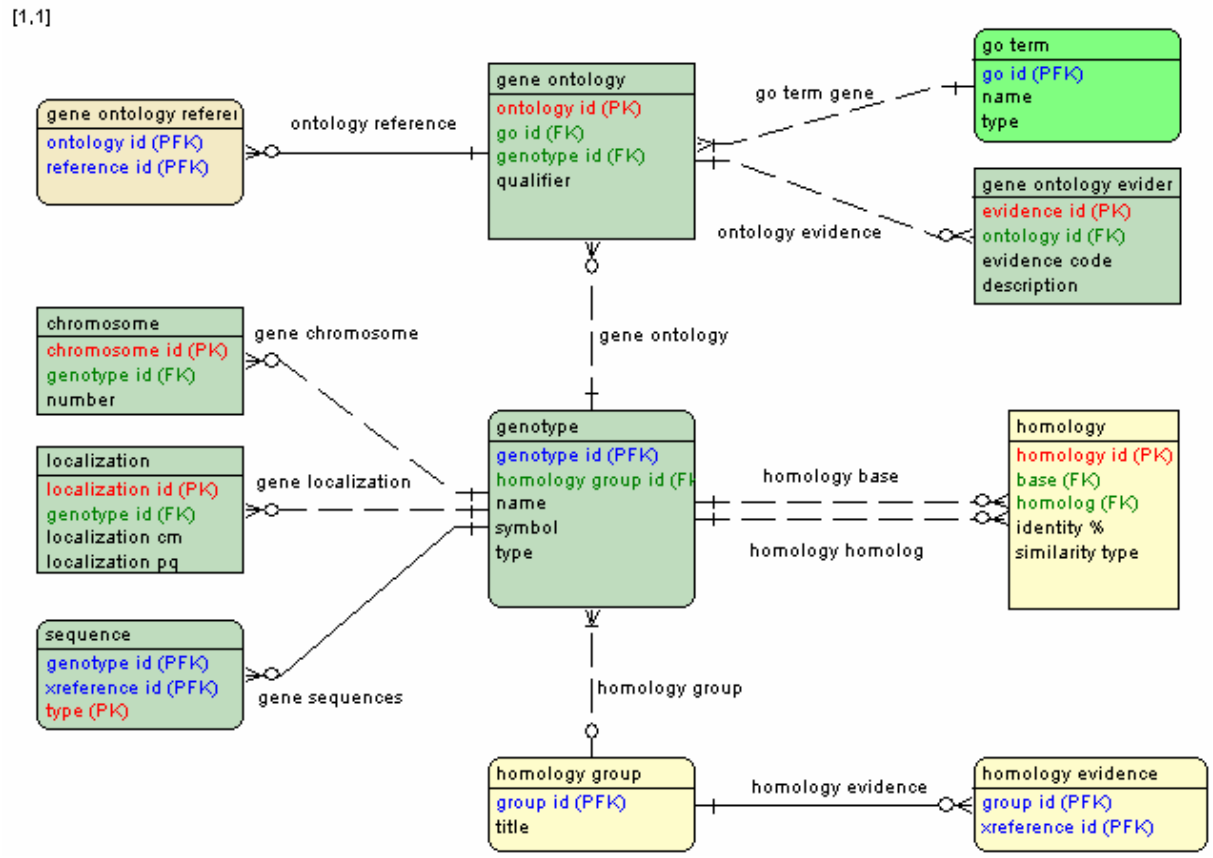
A1.2.2 Database scheme 'Common'



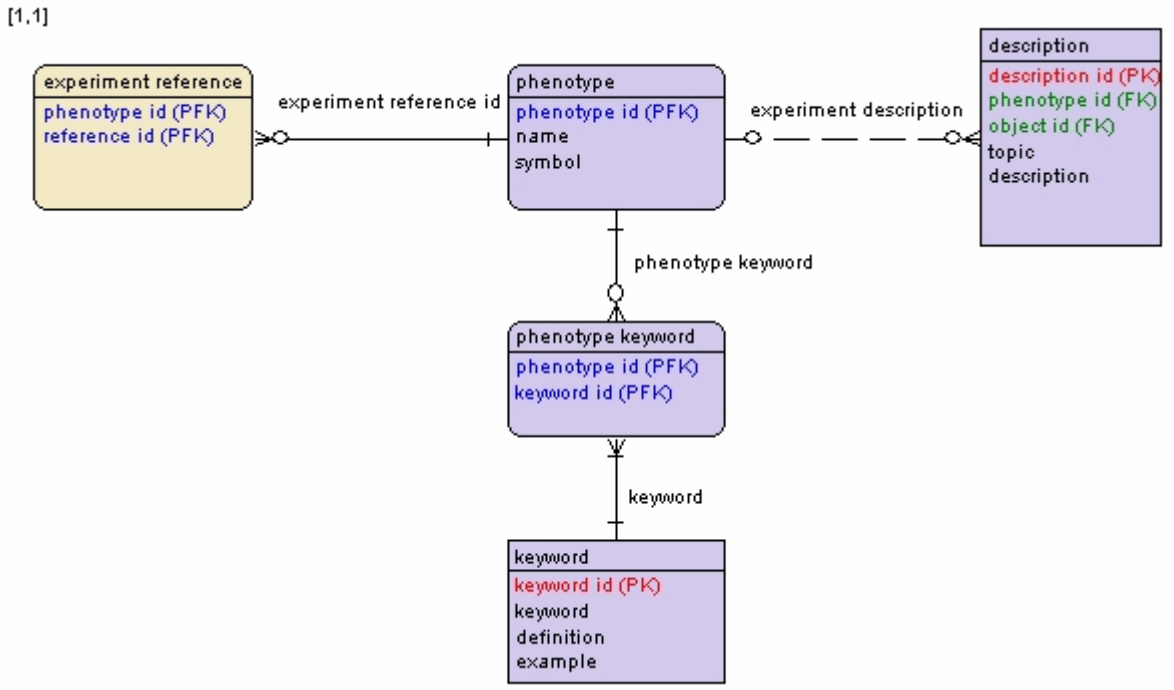
A1.2.3 Database scheme 'External reference'



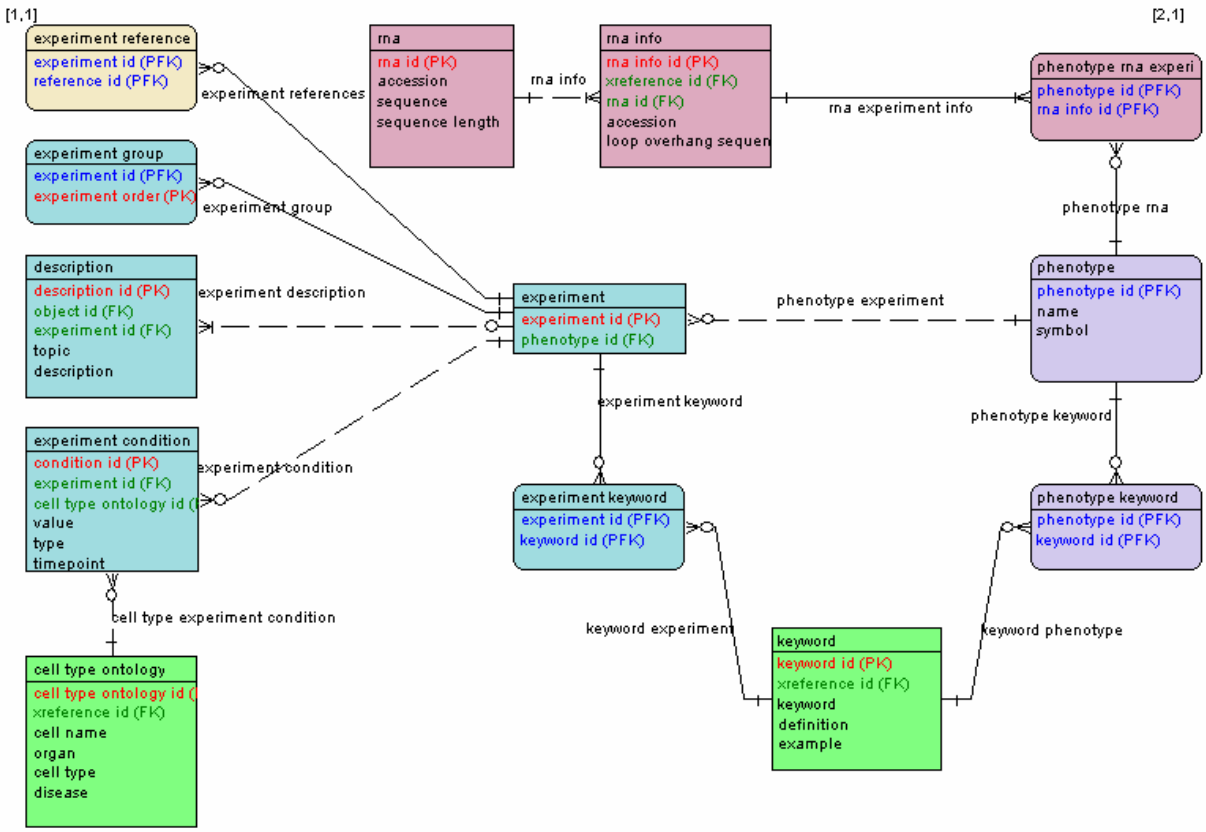
A1.2.4 Database scheme 'Genotype'



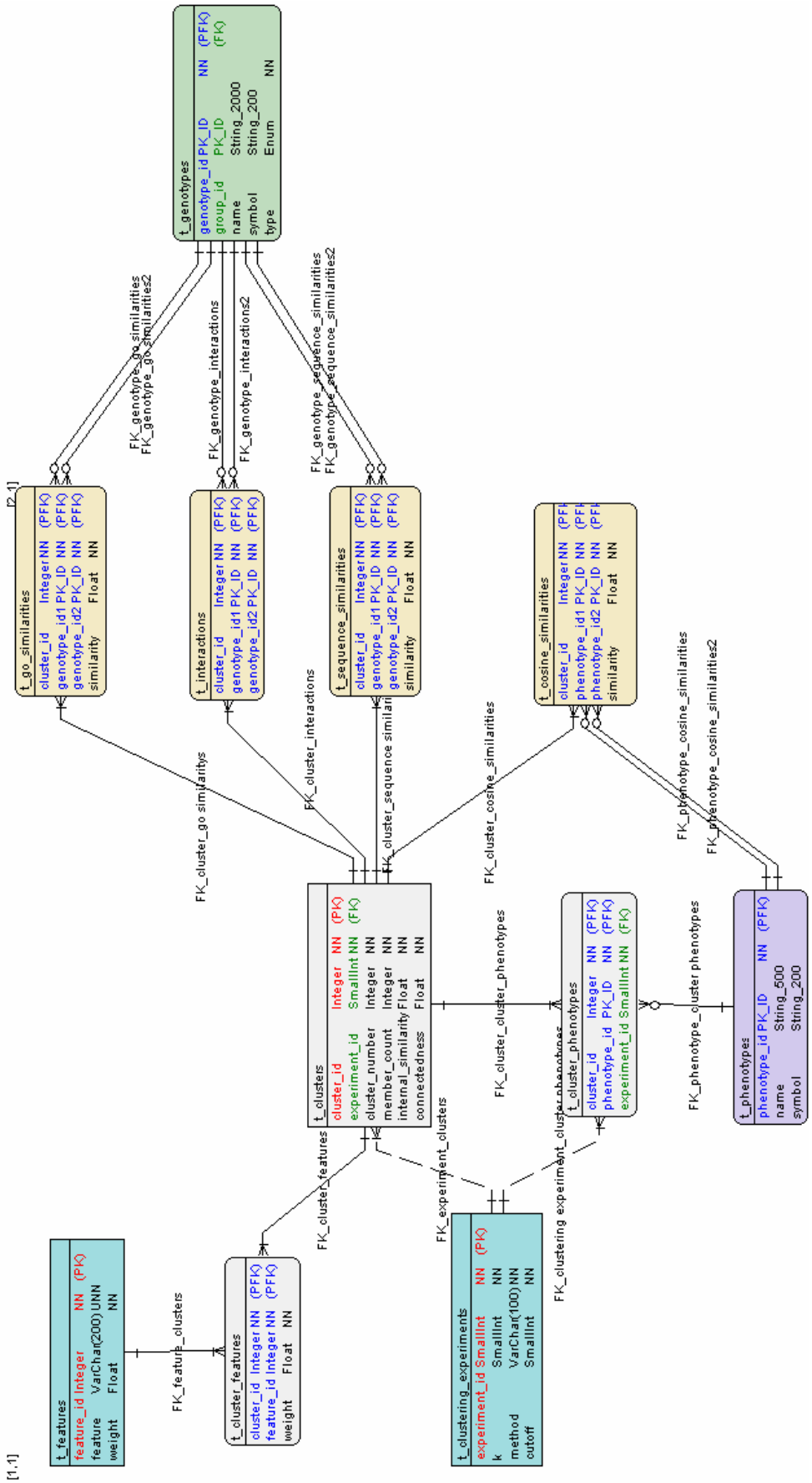
A1.2.5 Database scheme 'Phenotype'



A1.3 PhenomicDB version 2.x database scheme 'Phenotype'

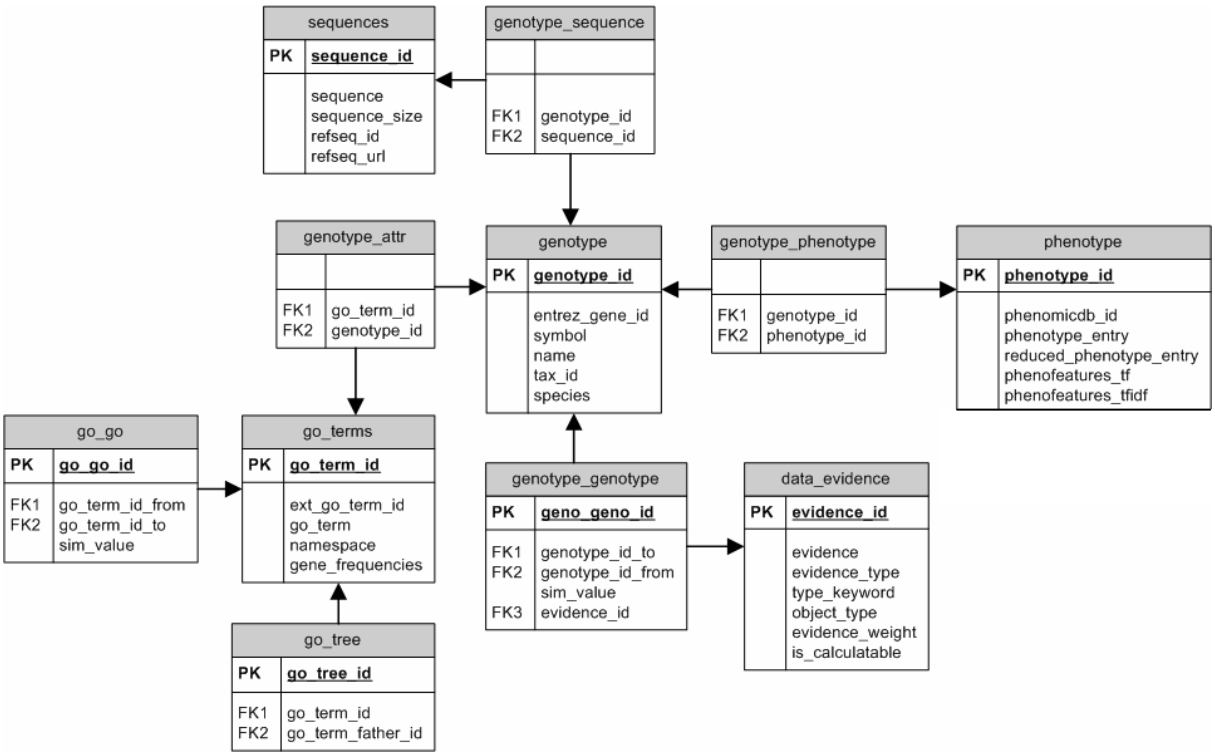


A1.4 PhenoMIX database scheme in PhenomicDB version 3.x

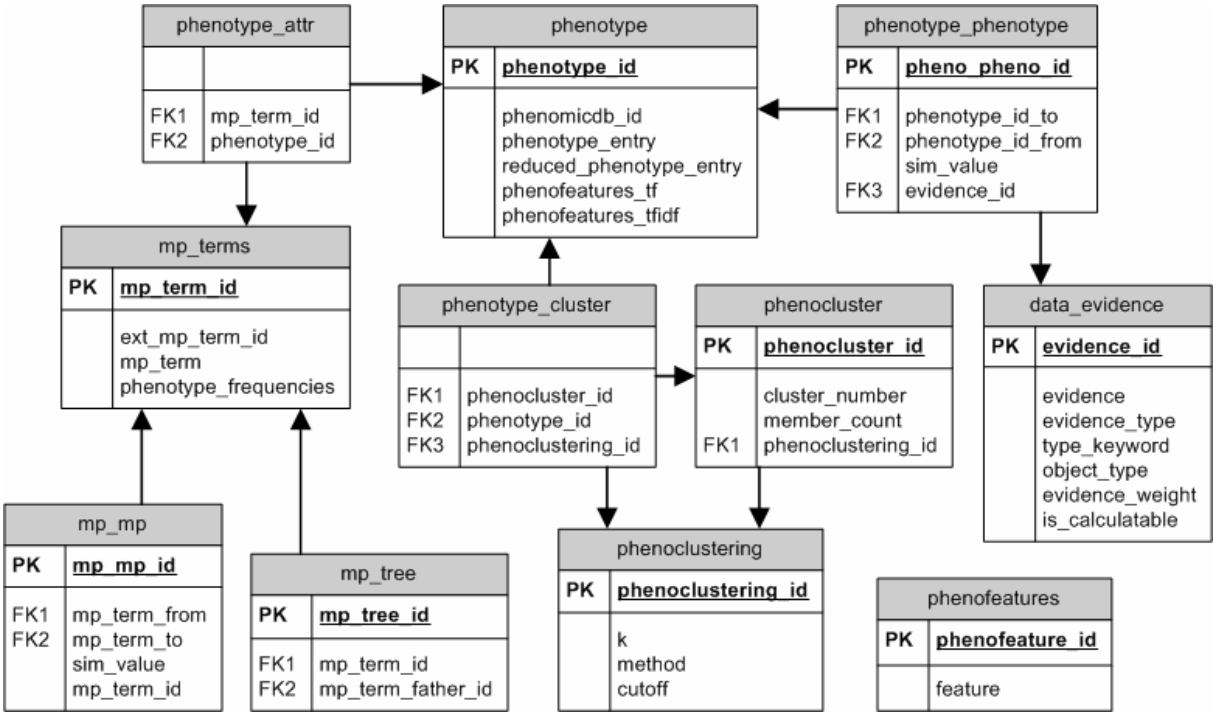


A2 PhenoMIX: Database scheme, packages and classes

A2.1 Database scheme 'Genotype'



A2.2 Database scheme 'Phenotype'



A2.3 Packages and classes

A2.3.1 The 'calc' package

A2.3.1.1.1 The class 'Distributions'

For any given evidence code, this class calculates a reasonable similarity score threshold which if applied would return a good portion of the available entities, but not too many of them (in view of retrieval inefficiencies). It also suggests an integer value k to return a reasonable group size (a certain quantile) if there is no applicable similarity threshold (i.e. when all similarity values are equal to '1').

Example:

```
Distributions dist = new Distributions(evidence);

int myK = dist.suggestK();

float myThreshold = dist.suggestThreshold();
```

A2.3.1.1.2 The class 'Phenoclustering'

This class does the actual phenoclustering by interacting on command-line with the vcluster algorithm from the CLUTO package. It needs just three input parameters; the minimum lengths of the input vectors, the value for k and the scoring method (TF or TFIDF). The class interacts with a method from the 'DBInterface' class which retrieves for a given (integer) length all phenodocs that are of at least this given length.

Example:

```
PhenoClustering phenoclustering = new PhenoClustering();

phenoclustering.do_phenoclustering(5000, 50, "tfidf");
```

A2.3.2 The 'calc.predict' subpackage

A2.3.2.1.1 The class 'AnnotationProbabilities'

This class calculates the probabilities of annotation for all annotations associated with a given list (ArrayList) of entities. It does so by calculating a Hypergeometric Distribution over the input set of associated annotations, depending on the type of the input entities. It retrieves the entire set from the database and the number of times this term has been associated and thus calculates the probability for this term to have ended up in this grouping by

chance. This class is used by the classes 'FunPred' to calculate the annotation probabilities of the lists of associated terms. The Hypergeometric Distribution is calculated by using the appropriate method from the 'MathUtils' class.

Example:

```
AnnotationProbabilities annoprob = new AnnotationProbabilities();

annoprob.setSourceEntities(listofentities);

annoprob.calc_annotation_probability();

Hashtable<java.lang.String,java.lang.Double> myresultlist = anno-
prob.getAnnotationProbabilities();
```

A2.3.2.1.2 The class 'FunPred'

This class calculates a prediction of genotype annotations for a network of phenotypes or vice versa. It calculates for each associated ontology term of each phenotype associated to genotype (and vice versa) from the input network a probability score and returns this list of terms together with this score. Furthermore, it calculates a precision and recall values for these observations with the methods discussed in section 2.2.8. The result is a list of words that are most likely to represent the ontology annotation of a phenotype for each associated gene from the network, as well as the precision and recall values of the calculations.

Example:

```
FunPred funpred = new FunPred();

funpred.setSourceNetwork(entitynetwork);

funpred. do_function_prediction();

Hashtable<java.lang.String,java.lang.Double> myresultlist = fun-
pred.getAnnotationProbabilities();

double myRecall = funpred.getRecall();

double myPrecision = funpred.getPrecision();
```

A2.3.2.1.3 The class 'PhenoPred'

This class calculates a prediction of phenotypes for a network of genotypes. It calculates for each feature of each associated phenotype of each genotype from the input network a

probability score and returns this list of features (words) together with this score. The result is a list of words that are most likely to represent the outcome of a phenotype experiment for each of the genes from the network.

Example:

PhenoPred phenopred = new PhenoPred();

phenopred.setSourceNetwork(genotypeentitynetwork);

phenopred. do_phenotype_prediction();

Hashtable<java.lang.String,java.lang.Double> myresultlist = phenopred.getPhenoScore();

A2.3.3 The 'calc.sims' subpackage

A2.3.3.1.1 The class 'Alignment'

This class calculates the alignment of two sequence strings and returns their similarity value (float). It could – theoretically – also return an ASCII representation of the alignment, showing in the top line the query sequence aligned with the subject sequence in the line below. Since this has no application, however, this output is discarded.

Example:

Alignment align = new Alignment();

align.setQuerySequence('ACGT');

align.setSubjectSequence('ACCT');

align.calc_alignment();

float myScore = align.getSimilarityScore();

A2.3.3.1.2 The class 'Cosine'

This class calculates the overall cosine similarity (float) of two vectors from phenodocs. The vectors can be given as strings or instances from the SparseVector class (otherwise they are transformed into such instances). The class itself uses the method cosine() from the SparseVector class to calculate the similarity values.

Example:

Cosine cosine = new Cosine();

cosine.setQueryVector(queryVector);

cosine.setSubjectVector(subjectVector);

cosine.calc_cosine_sim();

float myScore = cosine.getSimilarityScore();

A2.3.3.1.3 The class 'Ontology'

This class calculates the overall similarity score of two sets (ArrayLists of Integers, representing ontology term identifiers) of ontology terms. It contains a getter method for the calculated ontology score (float).

Example:

Ontology ontology = new Ontology();

ontology.setSubjectOntologyTerms(subjectOntologyTerms); // ArrayList<Integer> - a list of term identifiers

ontology.setQueryOntologyTerms(queryOntologyTerms); // ArrayList<Integer> - a list of term identifiers

ontology.setOntology("GO");

ontology.calc_ontology_sim();

float myScore = ontology.getSimilarityScore();

A2.3.4 The 'common' package

A2.3.4.1.1 The class 'EntityFactory'

This class is a factory for creating and populating instances of entities. Factories are programmed in order to separate objects from their populating methods. By this, the objects can be populated more easily by different methods, e.g. via a database, a form from a web-interface, a command-line, etc. The class contains a constant INSTANCE which instantiates an object of this class without constructor. This helps to handle the calls to the factory more easily and helps preserving system resources.

Example:

EntityFactory.getINSTANCE().method();

A2.3.4.1.2 The class 'EntityNotFoundException'

This class instantiates an Exception message and returns the Exception when it occurs. Particularly, it throws an exception if an identifier for which an Entity object should be created is not in the database.

A2.3.4.1.3 The class 'ErrorMessage'

This class is instantiated with an error message (as string) and returns it to the specified output. The reason this class exists is that it is extensible for throwing error message for any type of output, whether it is command-line or web-interface. The difference between exceptions and errors is that exceptions are usually fatal to the program, while errors mostly allow the program to keep running but communicate some type of failure in the process of the program.

A2.3.4.1.4 The class 'EvidenceFactory'

This class is a factory for creating and populating instances of evidence. Factories are programmed in order to separate objects from their populating methods. By this, the objects can be populated more easily by different methods, e.g. via a database, a form from a web-interface, a command-line, etc. The class contains a constant INSTANCE which instantiates an object of this class without constructor. This helps to handle the calls to the factory more easily and helps preserving system resources.

Example:

EvidenceFactory.getINSTANCE().method();

A2.3.4.1.5 The class 'ExceptionMessage'

This class is instantiated with an exception message (as string) and returns it to the specified output. The reason this class exists is that it is extensible for throwing exception message for any type of output, whether it is command-line or web-interface. The difference between exceptions and errors is that usually, exceptions are fatal to the program, while errors mostly allow the program to keep running but communicate some type of failure in the process of the program.

A2.3.4.1.6 The class 'InvalidClusteringException'

This class instantiates an Exception message and returns the Exception when it occurs. Particularly, it throws an exception if a clustering is invalid, i.e. when the number of vectors to be clustered is smaller than K.

A2.3.4.1.7 The class 'MathUtils'

This class contains some very useful additions to the built-in functions in java.lang.Math. Especially, this class contains a highly efficient implementation of the Hypergeometric Distribution, returning the probability that from a population of size N of which M members have a specific property, x elements with that specific property are drawn in k trials. The problem with the Hypergeometric Distribution is that it needs binomial coefficients. In many implementations (and also in the built-in java implementations) of these binomial coefficients, the actual values of 'n choose k' are calculated with factorials. This means that very large numbers are calculated and then need to be broken down again into very small numbers, if n and k are very large numbers themselves, but have only a very small absolute difference. An easy workaround for this dilemma is using logarithmic values returning the correct result of 'n choose k' even for large values of n and k. All methods of this class are static for direct access to the method without constructor.

Example:

MathUtils.method();

A2.3.4.1.8 The class 'PhenoClusteringFactory'

This class is a factory for creating and populating instances of phenoclusterings. Factories are programmed in order to separate objects from their populating methods. By this, the objects can be populated more easily by different methods, e.g. via a database, a form from a web-interface, a command-line, etc. The class contains a constant INSTANCE which instantiates an object of this class without constructor. This helps to handle the calls to the factory more easily and helps preserving system resources.

Example:

PhenoClusteringFactory.getINSTANCE().method();

A2.3.5 The ‘database’ package

A2.3.5.1.1 The class ‘DBHandle’

This class is the actual handler of a database connection. It keeps the physical information for the connection, such as password, database, host and driver as private variables and receives only the SQL query as string in the constructor.

Example:

DBHandle dbh = new DBHandle(‘select * from anytable’);

A2.3.5.1.2 The class ‘DBInsert’

This class is part of the connection interface to the database and is responsible for passing through data to the database handle and the query from the database query instances. The difference between the class ‘DBInsert’ and the class DBReturn’ is the type of return value. While ‘DBInsert’ only returns a Boolean value, indicating success or failure of the query, ‘DBReturn’ returns actual data from the database (as ResultSet).

Example:

DBInsert dbins = new DBInsert(‘insert into anytable anything’);

A2.3.5.1.3 The class ‘DBInterface’

This class is the full-fledged interface to the database and the factories and objects. It is by far the class with the most implemented methods. Each method contains a pre-formulated SQL query that is passed on to the other classes from the ‘database’ package and each of these method returns either data from the database, or calls upon class methods from the packages ‘objects’ or ‘common’ to instantiate and populate them with data from the database. The class contains a constant INSTANCE which instantiates an object of this class without constructor. This helps to handle the calls to the database more easily and helps preserving system resources.

Example:

DBInterface.getINSTANCE().method();

A2.3.5.1.4 The class 'DBReturn'

This class is part of the connection interface to the database and is responsible for passing through data from the database handle and the query from the database query instances. The difference between the class 'DBInsert' and the class DBReturn' is the type of return value. While 'DBInsert' only returns a Boolean value, indicating success or failure of the query, 'DBReturn' returns data from the database (as ResultSet).

Example:

DBReturn dbret = new DBReturn('select * from anytable');

A2.3.6 The 'objects' package

A2.3.6.1.1 The class 'Connection'

This class represents a pair-wise similarity between two entities of the same kind (either two genotypes or two phenotypes). The class keeps a unique identifier for the connection, as well as the full information of the two entity objects. Furthermore, the similarity value (distance) and the amount of support for this connection are stored. This class has been implemented as comparable in order to be able to find equal connections in two networks.

Example:

Connection connection = new Connection();

connection.setConnectorA(genotype1);

connection.setConnectorB(genotype2);

connection.setConnectionDistance(0.9876);

A2.3.6.1.2 The abstract class 'Entity'

This class instantiates genotype and phenotypes into a more general (and thus easier to use) class, namely 'Entity'.

Example:

Entity entity = new Genotype();

OR

Entity entity = new Phenotype();

A2.3.6.1.3 The class 'Evidence'

This class instantiates the evidences (hence: the similarity calculations) for entities. It contains basic setters and getters, like the type of evidence that has been instantiated (e.g. interaction from IntAct), the type of entity (genotype or phenotype) and some basic 'statistics' on the instantiation, e.g. whether this evidence is calculable (like sequence or ontology similarity) or just retrievable (like interaction or orthology), the number of available similarity values of this type of evidence, etc.

Example:

```
Evidence evidence = new Evidence();

evidence.setNumEntities(10);

int        numOfEntites = evidence.getNumEntities();
```

A2.3.6.1.4 The final class 'Genotype'

This class represents a Genotype. It is meant to instantiate an Entity as such.

Example:

```
Entity phenotype = new Genotype();
```

A2.3.6.1.5 The class 'Network'

This class represents a Network. In its basic form, it represents a list of entities (members of the network, either genotypes or phenotypes) and a list of connections (pair-wise similarities of genotypes and phenotypes). However, there have been many methods for network manipulation implemented in this class. One of the basic manipulation methods lies in adding or subtracting a network member (and thus also its connection and the other member of this connection, unless this member is active in another connection). Another implemented basic manipulation strategy is the removal of a connection (and thus the removal of both its members, unless they are involved in other connections). The class also contains advanced manipulation algorithms, such as finding a connected subnetwork for an entity, where, given an entity and an evidence, all other entities that have a connection with the given entity and evidence are returned as a network. Furthermore, this class also returns networks limited by a threshold or a size (ordered by similarity values). Lastly, it is responsible for the creation of consensus networks, which are networks of different evidences and overlapping members and connections.

Example:

```
Network network = new Network(entitylist, evidence);

ArrayList<Connection> nwconnections = network.getConnections();

ArrayList<Entity> nwmembers = network.getMembers();
```

A2.3.6.1.6 The class 'PhenoCluster'

This class instantiates objects of phenoclusters. Currently, these phenoclusters are Hash tables with integers for keys and entities (genotypes or phenotypes) for values. The class comes with a conversion method to convert a phenocluster to its group of associated genes and vice versa.

Example:

```
PhenoCluster phenocluster = new PhenoCluster();

Hashtable<Integer, ArrayList<Entity>> phenotypeclusters = phenocluster.getPhenotypeClusters();

Hashtable<Integer, ArrayList<Entity>> genotypeclusters = phenocluster.getGenotypeClusters();
```

A2.3.6.1.7 The class 'PhenoClustering'

This class represents an object keeping the parameter selection of a clustering of phenotypes. It is populated by retrieval from the database or from the result of a phenoclustering by CLUTO.

Example:

```
PhenoClustering phenocluster = new PhenoClustering(1000, 25, "tfidf");

phenocluster.setK(500);

phenocluster.setMinlength(50);

phenocluster.setMethod("tf");
```

A2.3.6.1.8 The final class 'Phenotype'

This class represents a Phenotype. It is meant to instantiate an Entity as such.

Example:

Entity phenotype = new Phenotype();

A2.3.6.1.9 The class 'SortableHashtable'

This class extends the normal Hashtable and represents a Hashtable which is sortable by values. It brings along the method getSortedValues() sorting the values of a Sortable-Hashtable. It inherits modifier methods from its superclass Hashtable. Equal values will be sorted by their keys. Keys and values have been generalized and are not limited to any particular object, type or class but must be of the Comparable class.

Example:

SortableHashtable sortableHashtable = new SortableHashtable();

sortableHashtable.put(1, 5);

sortableHashtable.put(2, 1);

sortableHashtable.put(3, 3);

<Iterator> i = labelfrequency.getSortedValues().iterator();

Entry entry = null;

while(i.hasNext()) {

entry = i.next();

System.out.println(entry.getKey() + " " + entry.getValue());

A2.3.6.1.10 The class 'SparseVector'

This class represents a SparseVector, i.e. a vector that has only few non-zero entries, as is the case for vectors from phenodocs, derived from a large body of documents. This class supports special (space) optimized routines for binary vectors and also supports k-means with Euclidean and cosine distance. Furthermore, one of the class constructors has been optimized to allow for a simple string to be transformed into an instance of this class.

Example:

String s1 = "1.0 1:2.523 3:7.81 4:2.313"; // (s = "label dim:val ...")

String s2 = "2.0 1:1.23 2:0.453 4:4.7235"; // (s = "label dim:val ...")

SparseVector vector1 = new SparseVector(s1);

SparseVector vector2 = new SparseVector(s2);

double cosineSimilarity = vector1.cosine(vector2);

A2.3.6.1.11 The final class 'SubstitutionMatrix'

This class represents an object needed for the calculation of sequence similarities and brings along the access method getValueAt(String querystring, String subjectstring).

Each value in a substitution matrix describes the rate at which one character in a sequence changes to other character states over time.

Example:

SubstitutionMatrix matrix = new SubstitutionMatrix();

int substitutionValue = matrix.getValueAt("A", "C");

A3 Numbers on phenotypes

This file contains some basic information on word and character sizes of phenotypes and the number and sizes of features in the resulting vectors.

Number of phenotypes/phenodocs: 39,610

Phenotypes:

Words: Mean: 169.99, Min.: 27, Max.: 985

Characters (excluding whitespaces): Mean: 1,135.29, Min.: 251, Max.: 6,528

Number of unique words: 113,283

Phenodocs:

Features: Mean: 67.87, Min.: 12, Max.: 364

Number of unique features: 73,188

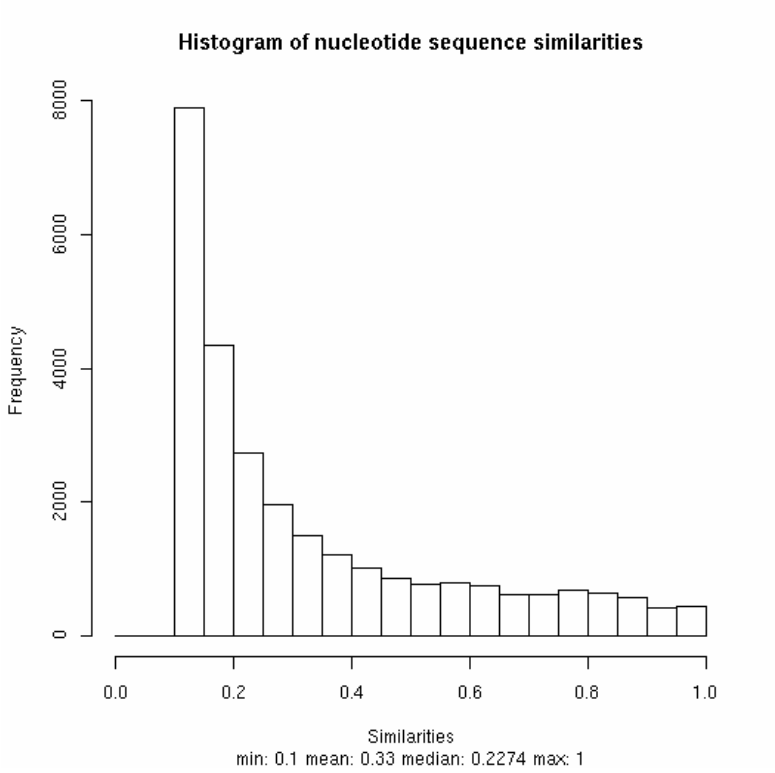
A4 Listing of and evidence for phenocopies

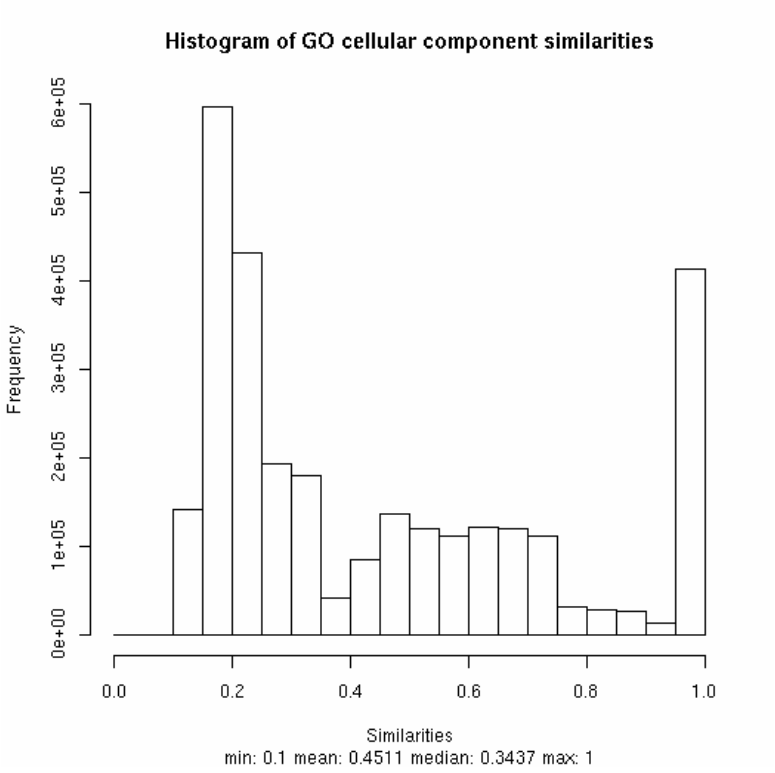
The following table lists all 27 phenocopies that have been identified from literature, including the Entrez gene IDs of each gene which is a phenocopy and the PubMed ID of the evidence for why exactly those genes are said to be phenocopies.

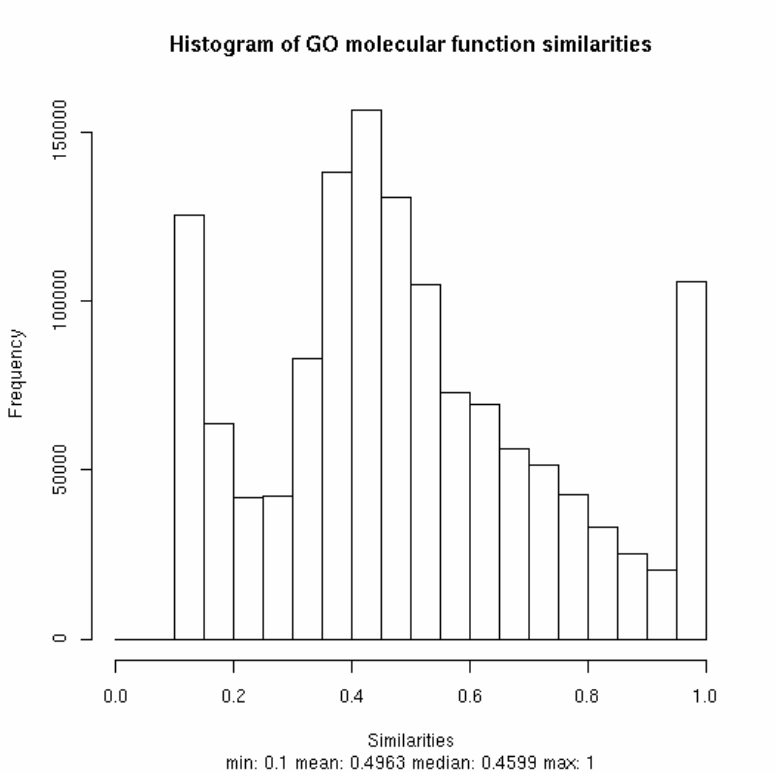
Number	Entrez gene identifier for phenocopies	PubMed identifier
1	38879, 36538	16129829
2	42445, 31293	12049762
3	15460, 22337	11564167
4	16835, 100017	15728179, 17200716, 14717060, 16343504
5	18516, 27140	15466398
6	36775, 43916	11877391
7	41885, 39844	17353360
8	11491, 13649	16079154
9	19378, 15426	15753214
10	44279, 34009	11029007
11	35197, 41363	16774999
12	3346167, 31816	16862128
13	12802, 14678	16924491
14	856580, 850675	17553781
15	13433, 20937	16980612
16	13711, 19116	15650748
17	14179, 14182	15221377
18	14460, 14461, 22761	12077323
19	15111, 22160	16237669
20	17319, 15251	17142669
21	16449, 18129	12496659, 15509774
22	4000, 10269	16079796
23	17283, 214162	15199122
24	19164, 19165, 18128,	12424225, 15525534
25	17125, 55994	15899870
26	14950, 74585	15998642
27	69581, 22408	11459829

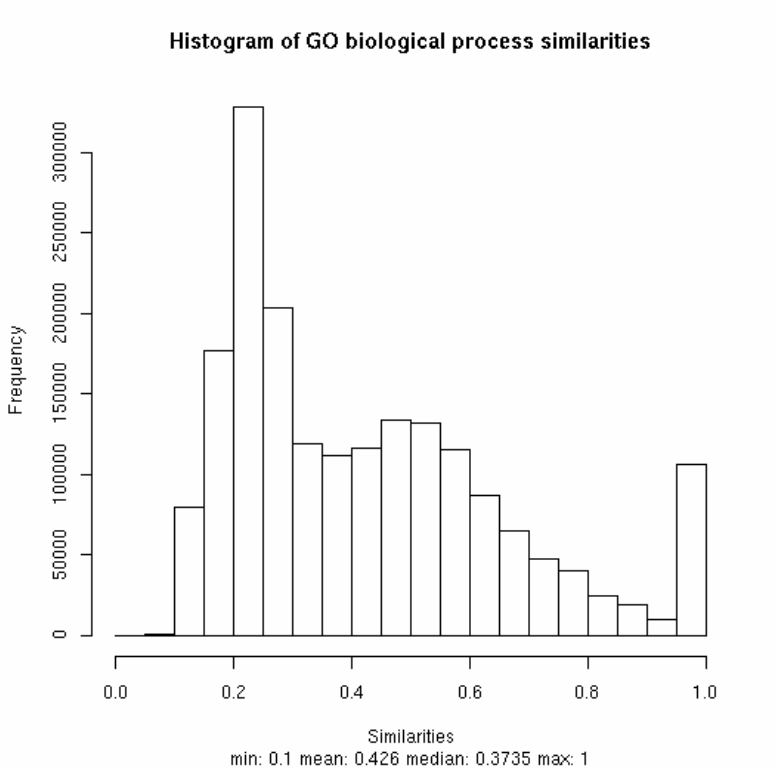
A5 Histograms of similarity data in PhenoMIX

A5.1 Similarities of genotypes where associated to phenotypes

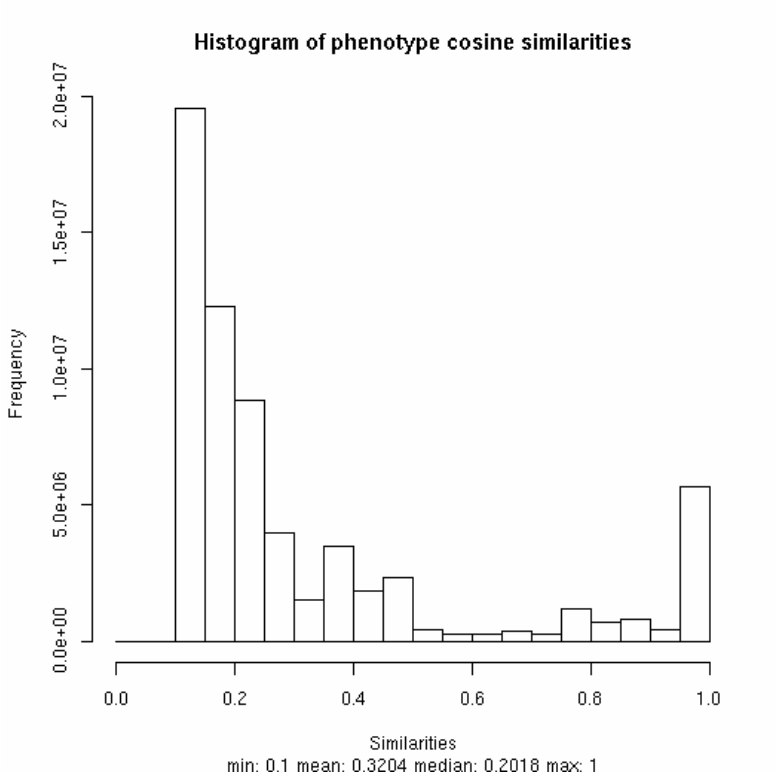


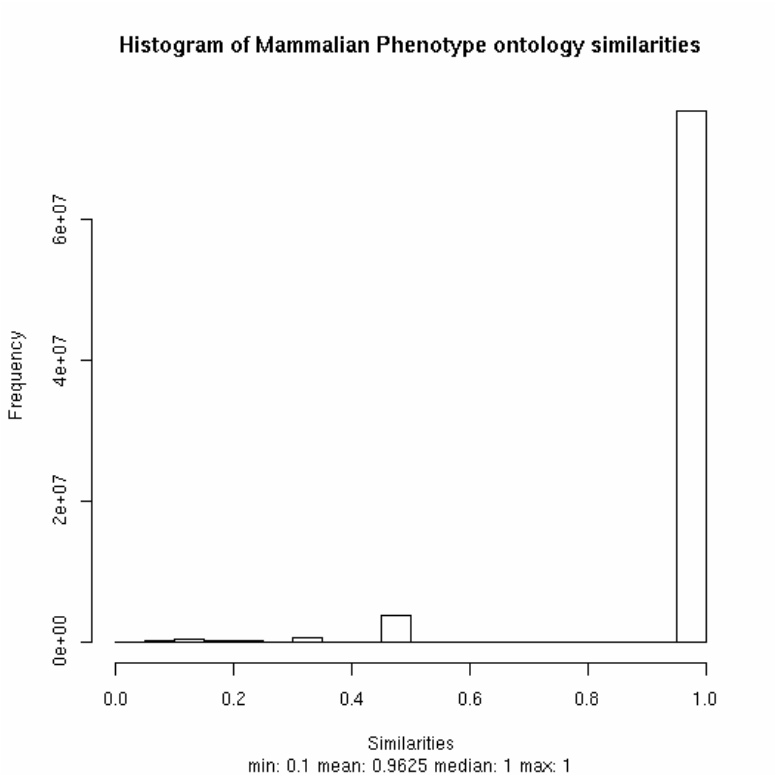



Histogram of GO molecular function similarities
Frequency
0
50000
100000
150000
0.0
0.2
0.4
0.6
0.8
1.0
Similarities
min: 0.1 mean: 0.4963 median: 0.4599 max: 1


Histogram of GO biological process similarities
Frequency
0
50000
100000
150000
200000
250000
300000
0.0
0.2
0.4
0.6
0.8
1.0
Similarities
min: 0.1 mean: 0.426 median: 0.3735 max: 1

A5.2 Similarities of phenotypes where associated to genotypes





A6 Figures for GO- and phenotype term prediction

A6.1 Figures for GO-term prediction from different similarities

Similarity measure	k	GO-term prediction # predicted terms (real)	GO-term prediction Recall (real)	GO-term prediction Precision (real)	GO-term prediction F-measure (real)	GO-term prediction # predicted terms (random)	GO-term prediction Recall (random)	GO-term prediction Precision (random)	GO-term prediction F-measure (random)
GO biological process	1,560	1,313	13.7%	67.1%	0.228	243	3.6%	31.7%	0.065
	780	1,159	13.2%	65.0%	0.219	281	5.4%	42.5%	0.096
	390	1,025	13.8%	64.0%	0.227	310	7.6%	50.5%	0.132
GO molecular function	1,571	1,009	12.3%	59.8%	0.204	229	2.6%	27.6%	0.048
	786	1,018	11.5%	63.0%	0.194	254	3.3%	35.2%	0.06
	393	968	11.3%	58.6%	0.189	310	4.0%	36.0%	0.072
GO cellular component	1,286	602	8.5%	50.6%	0.146	238	3.5%	43.2%	0.065
	643	813	6.6%	53.3%	0.117	370	3.7%	38.8%	0.068
	321	819	5.1%	40.0%	0.09	410	5.0%	37.1%	0.088
Nucleotide sequence	1,071	763	22.3%	49.0%	0.307	124	2.5%	11.8%	0.041
	536	509	18.7%	54.7%	0.279	117	2.6%	19.1%	0.046
	268	215	16.8%	64.6%	0.267	62	3.8%	31.3%	0.068
Intact PPi	646	546	12.3%	49.5%	0.197	142	1.3%	11.8%	0.023
	323	330	13.1%	59.0%	0.214	132	2.3%	22.6%	0.042
	162	127	12.7%	66.2%	0.213	63	4.2%	35.8%	0.075
BioGrid PPi	768	825	12.3%	53.0%	0.2	189	1.1%	11.3%	0.02
	384	640	14.4%	63.8%	0.235	184	2.1%	20.1%	0.038
	192	403	14.9%	69.5%	0.245	167	3.6%	33.5%	0.065
Homologene	662	598	24.1%	43.0%	0.309	48	1.4%	3.3%	0.02
	331	425	22.4%	42.2%	0.293	41	1.2%	3.0%	0.017
	166	244	22.3%	45.4%	0.299	26	1.5%	3.6%	0.021
Linear combination	2,188	850	12.5%	60.1%	0.207	178	4.2%	32.8%	0.074
	1,094	824	12.0%	63.5%	0.202	204	6.0%	44.7%	0.106
	547	638	12.7%	62.3%	0.211	192	8.2%	52.6%	0.142
Phenocluster	1,000	345	23.0%	67.9%	0.343	51	5.0%	29.1%	0.085

A6.2 Figures for phenotype prediction from different similarities

Similarity measure	k	PT-term prediction # predicted terms (real)	PT-term prediction Recall (real)	PT-term prediction Precision (real)	PT-term prediction F-measure (real)	PT-term prediction # predicted terms (random)	PT-term prediction Recall (random)	PT-term prediction Precision (random)	PT-term prediction F-measure (random)
GO biological process	1,560	9,219	9.9%	54.9%	0.168	3,176	4.3%	18.3%	0.070
	780	8,564	11.6%	64.3%	0.197	3,292	5.9%	32.7%	0.100
	390	8,200	12.8%	67.4%	0.215		7.5%	47.4%	0.130
GO molecular function	1,571	7,092	9.4%	57.9%	0.162	2,499	4.0%	14.6%	0.063
	786	8,000	10.1%	62.3%	0.174	3,131	5.8%	28.5%	0.096
	393	7,776	10.4%	62.6%	0.178	3,260	8.2%	41.6%	0.137
GO cellular component	1,286	6,556	7.6%	52.9%	0.133	3,147	5.2%	17.3%	0.080
	643	8,030	7.3%	55.8%	0.129	4,525	7.0%	36.4%	0.117
	321	8,271	9.0%	56.9%	0.155	4,182	7.3%	46.2%	0.126
Nucleotide sequence	1,071	6,821	15.4%	49.0%	0.234	2,093	5.3%	13.7%	0.076
	536	5,032	13.2%	58.8%	0.216	2,026	5.4%	21.1%	0.086
	268	2,799	12.6%	67.7%	0.212	1,204	5.3%	30.2%	0.090
Intact PPi	646	5,554	16.8%	68.9%	0.270	1,751	5.1%	17.0%	0.078
	323	3,950	14.2%	70.2%	0.236	1,664	5.6%	22.2%	0.089
	162	1,467	15.1%	69.8%	0.248	534	6.5%	26.1%	0.104
BioGrid PPi	768	9,243	16.0%	70.3%	0.261	2,374	5.3%	18.4%	0.082
	384	6,993	16.7%	71.9%	0.271	2,385	7.3%	32.1%	0.119
	192	4,791	16.6%	72.5%	0.270	2,014	9.2%	41.1%	0.150
Homologene	662	2,547	9.3%	23.0%	0.132	1,114	5.7%	8.7%	0.069
	331	1,992	9.3%	24.4%	0.135	981	5.9%	9.7%	0.073
	166	1,450	8.7%	23.0%	0.126	722	5.1%	9.0%	0.065
Linear combination	2,188	6,397	11.8%	57.5%	0.196	2,097	3.8%	13.8%	0.060
	1,094	6,229	13.1%	66.7%	0.219	2,239	5.7%	27.9%	0.095
	547	5,019	13.6%	69.4%	0.227	2,241	6.9%	40.9%	0.118
Phenocluster	1,000	12,896	27.1%	74.2%	0.397	5,589	6.2%	25.1%	0.099

A7 List of stop-words

The following list of 348 unique stopwords has been used to create phenodocs as described in sections 2.1.7 and 2.2.2.

these	none	toward	yourselves	anyway	only	a	cant
mm	did	anything	he	indeed	very	would	can
you	inc	describe	until	herself	significantly	no	where
both	do	full	was	anyhow	may	seen	mill
put	we	whereby	serious	really	re	might	throughout
my	find	twelve	another	therein	me	in	into
many	ten	thick	which	something	by	quite	never
see	somehow	myself	two	whoever	hasnt	upon	above
couldnt	to	four	nobody	below	mine	sixty	at
give	from	they	if	during	thereupon	first	empty
co	towards	especially	least	hereby	otherwise	thin	had
what	thru	same	less	bottom	herein	between	seemed
although	her	whereupon	him	using	whose	than	whatever
wherever	any	fify	himself	amount	hers	without	take
them	detail	fill	keep	alone	after	overall	hereafter
seems	perhaps	fire	wherein	side	am	just	our
sometime	more	us	own	now	get	become	who
used	an	nearly	always	becomes	nine	somewhere	nevertheless
computer	the	his	also	then	hence	its	obtained
again	made	i	yours	afterwards	five	amongst	bill
twenty	against	anyone	due	pmid	cry	thence	were
though	done	make	each	onto	back	whenever	next
shown	noone	everywhere	name	fifteen	go	ml	yourself
several	yet	under	interest	since	there	one	eleven
km	or	latterly	your	elsewhere	hundred	some	ever
system	could	across	but	whole	anywhere	with	even
of	does	well	too	few	as	here	ltd
still	this	through	and	down	showed	eg	therefore
mainly	before	via	over	amoungst	why	almost	up
all	once	while	six	about	everyone	kg	front
being	so	themselves	found	whom	became	off	out
she	regarding	please	is	ie	shows	thus	formerly
will	for	because	already	show	itself	etc	thereafter
nor	per	nothing	becoming	how	sometimes	former	theirs
much	else	eight	have	enough	ourselves	latter	someone
neither	whither	however	con	whence	third	ours	must
when	among	mg	moreover	those	has	beyond	namely
various	whereas	around	meanwhile	de	on	having	top
hereupon	everything	move	it	rather	thereby	use	seem
nowhere	be	others	last	un	further	been	behind
part	seeming	along	besides	other	sincere	should	
every	such	cannot	forty	their	mostly	together	
beforehand	most	that	within	whether	beside	except	
either	are	not	often	three	whereafter	call	

Eidesstattliche Erklärung

Ich, Philip Groth, geboren am 15.12.1978 in Berlin, erkläre hiermit, dass ich

- die vorliegende Dissertationsschrift ‚Knowledge Management and Discovery for Genotype/Phenotype Data' selbständig und ohne unerlaubte Hilfe angefertigt habe;
- sämtliche Quellen und Hilfsmittel an geeigneter Stelle referenziert habe;
- mich nicht bereits anderwärts um einen Doktorgrad in der Informatik beworben habe oder einen solchen besitze;
- die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät II der Humboldt-Universität zu Berlin gemäß amtlichem Mitteilungsblatt Nr. 34/2006 gelesen und verstanden habe und sie mir somit bekannt ist.

Berlin, Unterschrift: _____

P.Groth