

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

DATA INTEGRATION TECHNIQUES FOR MOLECULAR BIOLOGY RESEARCH

Promotoren:
Prof. dr. ir. B. De Moor
Prof. dr. ir. Y. Moreau

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Bert COESSENS

Juni 2006



KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT TOEGEPASTE WETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

DATA INTEGRATION TECHNIQUES FOR MOLECULAR BIOLOGY RESEARCH

Jury:
Prof. dr. ir. Y. Willems, voorzitter
Prof. dr. ir. B. De Moor, promotor
Prof. dr. ir. Y. Moreau, co-promotor
Prof. dr. ir. J. Vanderleyden
Prof. dr. B. Van den Bosch
Prof. dr. J. Vermeesch
Prof. dr. ir. K. Marchal

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de toegepaste wetenschappen

door

Bert COESSENS

U.D.C. 681.3*J3 Juni 2006

© Katholieke Universiteit Leuven – Faculteit Toegepaste Wetenschappen
Arenbergkasteel, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotocopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm or any other means without written permission from the publisher.

D/2006/7515/41

ISBN 90-5682-707-3

Dankwoord

Ik ben vermoedelijk het meest gelezen onderdeel van menig thesis, wat een nogal zware druk op me legt. Doel is steeds het bedanken van mensen: bedankt. Maar wat als ik mensen vergeet? Wat als ik saai ben, te lang, of te kort?

Vandaag ben ik tot Bert gekomen. Via zijn handen vind ik een weg naar buiten. Ik lees mezelf en vraag me af hoe ik er uit wil zien. In elk geval dankbaar: heil aan de promotoren! Voor hun inzet, hun hulp, de *pep talk* op moeilijke momenten. Een doctoraat maken doe je nooit alleen.

Ik denk terug aan de gebeurtenissen die geleid hebben tot mijn geboorte. Het begon allemaal in juli 2001, na een aangenaam gesprek ten huize ESAT. Met het volste vertrouwen, werd Bert opgenomen. Vruchtbare samenwerkingen ontstonden, vruchtbaar onderzoek ontsproot. Onderzoek verrichten doe je nooit alleen.

Onderzoek verrichten is geen beroep, het is een manier van leven, een manier van zijn. Leven is leren, zijn is zin. Je bent wat je eet. Zo wordt Bert gevormd, door familie, vrienden en collega's. Leven doe je niet alleen.

En zo ontsta ik, uit dankbaarheid voor alles wat dit werk heeft mogelijk gemaakt.

Bedankt Bart, bedankt Yves! Jullie waren mijn rots in de branding, mijn centraal massief.

Bedankt Kristof, Stein, Pat en Steven! Waar was ik zonder jullie gebleven?

Dank aan bioi! Werken op ESAT is als God in Frankrijk leven...

Bedankt aan al mijn vrienden, die me hielpen worden wie ik ben!

Jullie, door wie ik mezelf het beste ken.

Bedankt mams en paps, Liesje en Saartje! Jullie zijn mijn thuis, jullie wonen in mijn huid.

Bedankt Ellen! En onze kleine spruit, vol ongeduld kijken we naar je uit.

Bedankt!

Abstract

The availability of entire genomes caused a general adoption of high-throughput techniques like microarrays. With them the focus of molecular biology research shifted from the study of a single gene (*one gene, one Ph.D.*) to the functional analysis of large groups of genes. As the amount of raw data grows, so does the need for methods to automate the analysis and integrate the results with existing knowledge. The process of gaining insight into complex genetic mechanisms depends on this data integration step in which bioinformatics can play an important role.

The structure of this thesis follows the cyclic nature of knowledge acquisition. Acquiring knowledge in scientific practice always starts with testing a hypothesis. In molecular biology, a wet-lab experiment is performed and the results interpreted in the context of well-known information. This (hopefully) leads to improved insights that will allow new hypotheses to be formulated.

In the different chapters of this thesis, several methods are presented that allow inclusion of biological knowledge in the analyses of high-throughput experimental data. A distinction is made between early, intermediate, and late integration based on when in the analysis pipeline the knowledge is included. In the first place, only one source of knowledge is used to validate the experimental results. Afterwards, a myriad of complementary information sources is combined to discover new relations between genes. Finally, a web services architecture is presented that was developed to enable an efficient and flexible access to several information sources.

Korte inhoud

Moleculaire biologie wordt heden ten dage gedomineerd door hoge-doorvoertechnologieën zoals microroosterexperimenten, waarbij de expressie van duizenden genen tegelijk gemeten wordt. Dergelijke technologieën zijn een gevolg van de algemene beschikbaarheid van steeds meer DNA sequenties van uiteenlopende organismen. Terwijl tot voor kort het brandpunt van veel moleculair biologisch onderzoek gericht was op het bestuderen van individuele genen, zijn er steeds meer mogelijkheden om de aard van groepen van genen te bestuderen. Deze ontwikkelingen hebben tot gevolg dat steeds meer gegevens bij analyses betrokken worden en er in toenemende mate nood is aan automatisatie van analyses enerzijds, en aan integratie van de resultaten met de bestaande kennis anderzijds. Het is op dit punt dat bioinformatica een belangrijke rol te spelen heeft.

De opbouw van deze thesis volgt het cyclisch verloop van het verwerven van kennis. In de wetenschappelijke praktijk start de zoektocht naar kennis steeds met het stellen van een hypothese. Om de hypothese te testen wordt een experiment opgezet (in het geval van de moleculaire biologie is dit doorgaans een laboratoriumonderzoek). De resultaten van het experiment worden geanalyseerd en in de context van algemeen aanvaarde kennis getoetst. Dit leidt dan tot nieuwe inzichten waarop nieuwe hypotheses kunnen gebaseerd worden die dan in het laboratorium getest kunnen worden.

In de verschillende hoofdstukken worden methoden besproken voor het gebruik van algemeen aanvaarde kennis bij het analyseren van experimentele data. Afhankelijk van het moment waarop deze kennis bij de analyse betrokken wordt, spreekt men van vroege, intermediaire of late integratie. Eerst wordt slechts informatie van 1 bron gebruikt om resultaten van experimenten te valideren. Dan wordt gekeken hoe een groot aantal complementaire informatiebronnen gecombineerd kan worden om nieuwe verbanden tussen genen aan het licht te brengen. Tot slot wordt een web-service-architectuur voorgesteld die ontwikkeld werd om een efficiënte en flexibele toegang tot verschillende databronnen te verschaffen.

Notation

Abbreviations

ANOVA	ANalysis Of VAriance
API	Application Programming Interface
AQBC	Adaptive Quality-Based Clustering
AUC	Area Under the Curve
BIND	Biomolecular Interaction Network Database
BiNGO	Biological Networks Gene Ontology tool
BLAST	Basic Local Alignment Search Tool
BN	Bayesian Network
BP	Biological Process (part of the Gene Ontology)
CC	Cellular Component (part of the Gene Ontology)
CDF	Cumulative Distribution Function
CDS	CoDing Sequence
CNS	Conserved Non-coding Sequence
DAG	Directed Acyclic Graph
DAS	Distributed Annotation System
DNA	DeoxyriboNucleic Acid
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EMBOSS	European Molecular Biology Open Software Suite
ER	Entity Recognition
ESS	Error Sum of Squares
EST	Expressed Sequence Tag
FN	False Negatives
FP	False Positives
GBA	Guilt By Association
GO	Gene Ontology
GUI	Graphical User Interface
HGNC	HUGO Gene Nomenclature Committee

HUGO	HUman Genome Organization
HTTP	HyperText Transfer Protocol
IDF	Inverse Document Frequency
IE	Information Extraction
IR	Information Retrieval
JSP	Java Server Pages
JWS	Java Web Start
KD	Knowledge Discovery
KEGG	Kyoto Encyclopedia of Genes and Genomes
LSI	Latent Semantic Indexing
MeSH	Medical Subject Headings
MF	Molecular Function (part of the Gene Ontology)
MGI	Mouse Genome Informatics
MIAME	Minimum Information About a Microarray Experiment
NCBI	National Center for Biotechnology Information (US)
OMIM	Online Mendelian Inheritance in Man
PDF	Probability Density Function
POS	Part Of Speech
PRM	Probabilistic Relational Model
PWM	Position Weight Matrix
RMI	Remote Method Invocation
ROC	Receiver Operating Characteristic
SC	Silhouette Coefficient
SGD	Saccharomyces Genome Database
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
SRS	Sequence Retrieval System
SVD	Singular Value Decomposition
TAIR	The Arabidopsis Information Resource
TF	Transcription Factor
TN	True Negatives
TP	True Positives
TFBS	Transcription Factor Binding Site
UDDI	Universal Description, Discovery, and Integration
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium
WSA	Web Services Architecture
WSDL	Web Service Description Language
XML	eXtended Markup Language

Gene nomenclature

All gene symbols are italicized and protein symbols are normally the same as the encoding gene symbols but not italicized. Human gene symbols[1] are designated by uppercase Latin letters or by a combination of uppercase letters and Arabic numerals, for example *BRCA1*, *CYP1A2*. To identify human genes either HUGO symbols as found in the Entrez Gene and Ensembl databases or Ensembl gene identifiers (ENS*) are used.

[1] Guidelines for human gene nomenclature can be found on `http://www.gene.ucl.ac.uk/nomenclature/guidelines.html` [147].

Related publications

- Stein Aerts, Gert Thijs, **Bert Coessens**, Mik Staes, Yves Moreau and Bart De Moor (2003) TOUCAN: deciphering the *cis*-regulatory logic of coregulated genes. *Nucleic Acids Research*, 31(6), 1753-1764.

- Kristof Engelen, **Bert Coessens**, Kathleen Marchal, Bart De Moor (2003) MARAN: normalizing microarray data. *Bioinformatics*, 19(7), 893-894.

- **Bert Coessens**, Gert Thijs, Stein Aerts, Kathleen Marchal, Frank De Smet, Kristof Engelen, Patrick Glenisson, Yves Moreau, Janick Mathys, and Bart De Moor (2003) INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Research*, 31(13), 3468-3470. (*)

- Patrick Glenisson, **Bert Coessens**, Steven Van Vooren, Yves Moreau, Bart De Moor (2003) Text-based gene profiling with domain-specific views. In *Proceedings of the First International Workshop on Semantic Web and Databases* (SWDB 2003), Berlin, Germany, 15-31.

- Patrick Glenisson, **Bert Coessens**, Steven Van Vooren, Janick Mathys, Yves Moreau, Bart De Moor (2004) TXTGate: Profiling gene groups with text-based information. *Genome Biology*, 5(6), R43.1-R43.12.

- Stein Aerts, Diether Lambrechts, Sunit Maity, Peter Van Loo, **Bert Coessens**, Frederik De Smet, Leon-Charles Tranchevent, Bart De Moor, Peter Marynen, Bassem Hassan, Peter Carmeliet, Yves Moreau (2006) Gene prioritization via genomic data fusion. *Nature Biotechnology*, 24, 537-544. (*)

(*) First author publications

Contents

Dankwoord **i**

Abstract **iii**

Korte inhoud **v**

Notation **vii**

Related publications **xi**

Contents **xiii**

1 Bioinformatics and its role in biological research **1**
1.1 From *in vitro* to *in silico* and back 1
1.2 Biological research in the post-sequence era 2
1.3 Towards systems biology 4
1.4 Integration of heterogeneous data 6
1.5 Early, intermediate, and late data integration 7
1.6 Web services integration 10
1.7 Using textual knowledge in biological analyses 10
1.7.1 Short overview of molecular biology text mining . . . 12
1.7.2 The vector space model 14
1.7.3 Document similarity 16
1.7.4 Construction of an entity index 17
1.7.5 Dimensionality reduction 17
1.7.6 Domain-specific views 18
1.8 Thesis overview 19

2 Grouping genes **21**
2.1 General-purpose data set 22

2.2 Grouping genes based on expression data . . . 23
2.2.1 Preprocessing . . . 23
2.2.2 Cluster analysis . . . 24
2.2.3 Cluster quality . . . 25
2.2.4 Discussion . . . 27
2.3 Grouping genes based on textual information . . . 28
2.3.1 Cluster analysis . . . 32
2.3.2 Cluster quality . . . 32
2.3.3 Comparison with grouping based on expression . . . 32
2.3.4 Discussion . . . 35
2.4 Combining expression and textual data . . . 35
2.4.1 Early integration . . . 37
2.4.2 Cluster quality . . . 38
2.4.3 Discussion . . . 38
2.5 Conclusion . . . 41

3 Gene group validation **43**
3.1 Gene Ontology to characterize gene groups . . . 44
3.1.1 Statistically over-represented GO terms . . . 46
3.1.2 Distances between GO terms . . . 52
3.2 Textual profiling of gene groups . . . 61
3.2.1 Profiling gene groups with text-based information . . . 63
3.2.2 Subclustering gene groups based on textual profiles . . . 67
3.3 Conclusion . . . 68

4 Expanding groups of genes **71**
4.1 Gene co-citation and co-linkage . . . 71
4.1.1 Examples . . . 74
4.1.2 Discussion . . . 76
4.2 Computational prioritization . . . 77
4.2.1 Methodology . . . 79
4.2.2 Data sources . . . 81
4.2.3 Computational techniques . . . 84
4.2.4 Statistical validation . . . 87
4.2.5 Discussion . . . 98
4.3 Conclusion . . . 100

5 Web services integration **101**
5.1 Web services technologies . . . 103
5.1.1 The web services architecture . . . 103

5.1.2 SOAP and WSDL . . . 103
5.2 Bioinformatics and web services . . . 106
5.2.1 BioMOBY . . . 106
5.2.2 ^{my}Grid . . . 108
5.3 Web services integration . . . 109
5.3.1 Computing architecture and technicalities . . . 109
5.3.2 INCLUSive . . . 110
5.3.3 Toucan . . . 114
5.3.4 Endeavour . . . 115
5.4 Conclusion . . . 121

6 Conclusions and prospects **123**
6.1 Accomplishments . . . 124
6.2 Future work . . . 126
6.3 Outlook . . . 127

A Order statistics **129**

B Supplementary material **137**

Nederlandse samenvatting **147**

Bibliography **164**

Chapter 1

Bioinformatics and its role in biological research

THIS introductory chapter points out the importance of bioinformatics, and of the work described in this thesis, for molecular biology research. This thesis deals with computational methods to integrate high-throughput experimental data and high-level biological knowledge. Through *proof-of-concept* studies and biological validations, it is shown that these methods have the potential to speed up analyses considerably. Besides, a computing architecture is proposed based on web services technologies to enable efficient access to heterogeneous data sources.

In Sections 1.1, 1.2, and 1.3, the context of the presented work is described. Section 1.4 overviews the current status of *integromics*, a term used to denote the integrated use of heterogeneous data sources in molecular biology. Sections 1.5 and 1.6 give an overview of the methods and main methodological results described in this thesis. Since a lot of biological knowledge is captured in free text (textual descriptions, scientific abstracts, full papers, and so on), several text mining methods are frequently used throughout this thesis. Therefore, a more detailed description of these methods is given in Section 1.7.

1.1 From *in vitro* to *in silico* and back

In the context of this thesis, the term *knowledge* has to be interpreted as a type of information that is useful in practice; knowledge is information that can be applied. *Data*, on the other hand, is a passive type of information that needs processing and analysis to gain knowledge from. The term *information*

is often used to denote the continuum of more or less structured information in the phase between *data* and *knowledge*. Figure 1.1 lists the characteristics of this *information space*. The scientific challenge is to gain new knowledge by analyzing data.

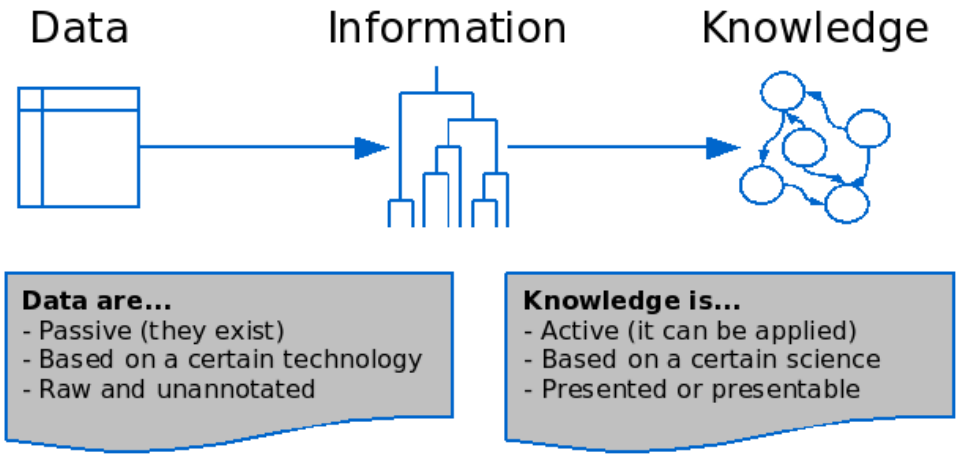


Figure 1.1: The difference between *data* and *knowledge*, the two extreme ends of the *information space*.

In molecular biology, a biological phenomenon is traditionally studied by performing *in vitro* experiments according to certain standard or custom protocols. The outcome of the experiment is then analyzed and interpreted in the context of the existing knowledge. This is called the *in silico* step, because of the important role computers play in it. Based on the results of the previous experiment, new experiments are designed until the biological observation of interest can be explained and new knowledge is obtained. Thus, knowledge acquisition in molecular biology research is a cyclic process in which new knowledge is created in an incremental way (see Figure 1.2).

1.2 Biological research in the post-sequence era

In the post-sequence era, the traditional way of biological experimentation changed completely. The availability of complete genome sequences led to an explosion of high-throughput techniques (like microarrays, yeast-two hybrid assays, and so on) resulting in an ever growing amount of raw data to be analyzed. This trend caused a shift in focus from the study of a single gene or process to the analysis of the behavior of large groups of genes [59, 13].

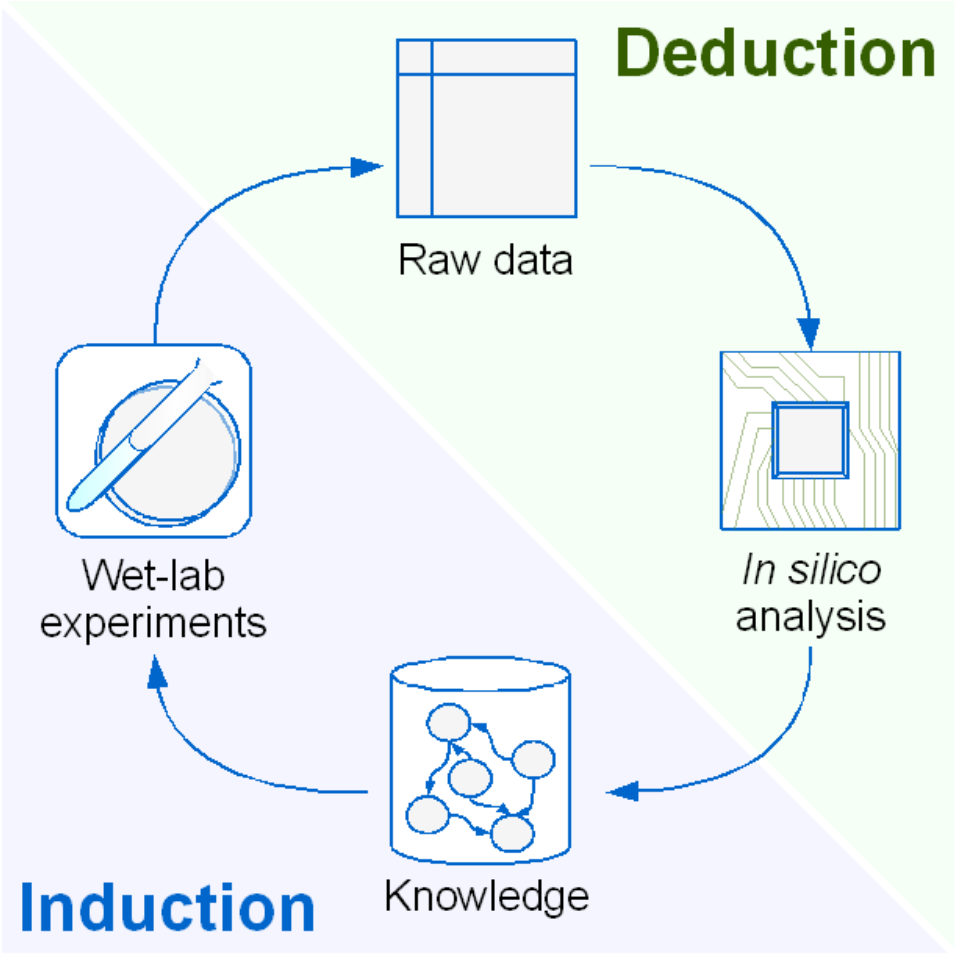


Figure 1.2: Knowledge acquisition is a cyclic process. During the induction step, a new hypothesis is formulated starting from a specific scientific question. In the deduction step, an experiment is set up to prove the hypothesis. The results of the experiment are then interpreted in the context of the existing knowledge, new insights are formulated, and a new hypothesis can be postulated.

In other words, biology moved from a data-limited to an analysis-limited science [94]. High-throughput techniques make exploratory research possible (as opposed to hypothesis-driven research), but at the cost of an increased need for standards in design, execution, and interpretation of experiments. As the price of acquiring biological data lowers, so does the data quality and it just gets harder to come to sensible conclusions.

Apart from the changing focus of biological research, advances in information technology enabled large amounts of data to be shared world wide. The rise of bioinformatics as a discipline is tightly connected with the upcoming of the Internet [70]. Especially the Human Genome Project (HGP) [99] sparked research into huge and interconnected biological databases.

As a consequence of these developments, bioinformatics has become an indispensable part of the knowledge acquisition cycle, not only to speed up the analysis of raw data, but more important, by coping with the huge amount of heterogeneous information available on the Internet.

1.3 Towards systems biology

The next challenge in biology is to wrap up all gathered information into workable models. Reductionist approaches made biological research successful in the last century. Currently, high-throughput technologies make possible a move towards more integrative approaches and the study of biological systems as a whole. The challenge is now to model biological processes globally rather than break them apart to explain their elements (see Figure 1.3). This is what so-called *systems biology* is all about.

Research in systems biology is either principle-driven or data-driven. Because of its complex intracellular physicochemical environment, a biological system is hard to describe in terms of mathematical equations. This explains the lack of a sound theoretical basis behind biology. However, the tendency towards high-throughput experimentation in molecular biology research enables data-driven models to be worked out for biological systems [97]. Both principle-driven and data-driven approaches can now complement each other. While the quality of high-throughput data will improve, and new (and better) technologies will arise to measure cellular properties, better parameter estimations might lead to improved mathematical models. These models could then be used to interpret the high-throughput data on a more qualitative level, thus bringing the biological knowledge to a systems level.

The remaining interests and challenges to enable true *in silico* biology

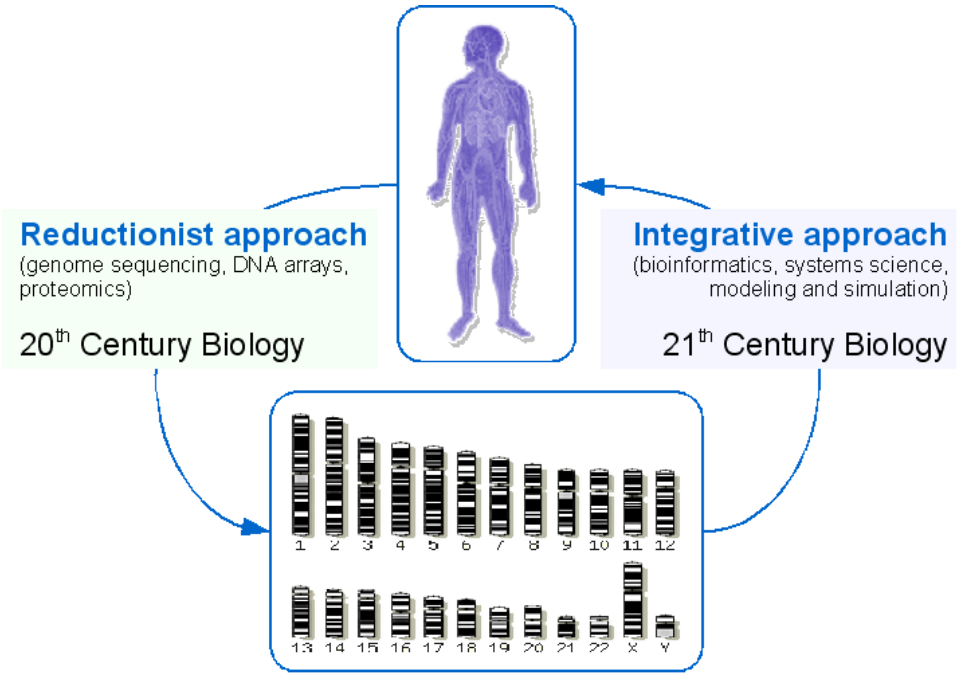


Figure 1.3: Biological research is shifting from reductionist towards integrative approaches. In the past, research in molecular biology focused on studying individual cellular components. Current high-throughput technologies enable the study of thousands of genes or proteins simultaneously. This causes a shift from reductionist biology towards more integrative approaches. Figure adapted from Bernhard Palsson [96].

can be grouped in three categories [90]:

- Integration of biological data
- Creation of a uniform and scalable systems view
- Promotion of science networking

The challenge of biological data integration is the main focus of this thesis and will be explained in more detail in the next section.

1.4 Integration of heterogeneous data

As outlined in the previous sections, the process of successfully gaining insight into complex genetic mechanisms increasingly depends on a complementary use of a variety of resources. Drilling down into the dispersed database entries of hundreds of genes is notably inefficient and shows the need for higher-level integrated views that can be captured more easily by an expert's mind.

Analogous to the different *-omics* terms used to denote, for instance, the study of the genes (genomics), transcripts (transcriptomics), or proteins (proteomics) in the cell, the term *integromics* [143] was introduced to describe the research into integration of data from molecular biology. Integromics can be divided in two main areas of research: *conceptual* or *qualitative data integration* versus *algorithmic* or *quantitative data integration*.

Conceptual data integration is concerned with combining data from different databases, in different formats, into a global (conceptual) scheme. As biology is a knowledge-driven discipline, access to information is of utmost importance. However, the exploding number of biological databases on the Internet has made manual integration of relevant biological information infeasible. The goal of this type of research is to provide scientists with a platform to retrieve the information they need as fast as possible and with a minimum of user intervention [61].

Algorithmic data integration comes down to the use of different data types in an experiment's analysis pipeline. In general, raw experimental data is combined with annotated information using mathematical or statistical approaches to find biologically meaningful results. Combining raw and annotated data can occur at different levels of the analysis, as outlined in Figure 1.4. During *early integration*, different types of data are transformed and combined into a common format as input of the analysis. An *intermediate integration* happens when analysis results are combined with

another type of information in a subsequent analysis step. Meta-clustering analyses are an example of this type of integration, in which two clustering results based on different data sources are combined. *Late integration* occurs when analysis results are interpreted and verified using relevant annotated information. This late integration coincides with the deduction step of the knowledge acquisition cycle (see Figure 1.2) and is, of course, related to conceptual data integration.

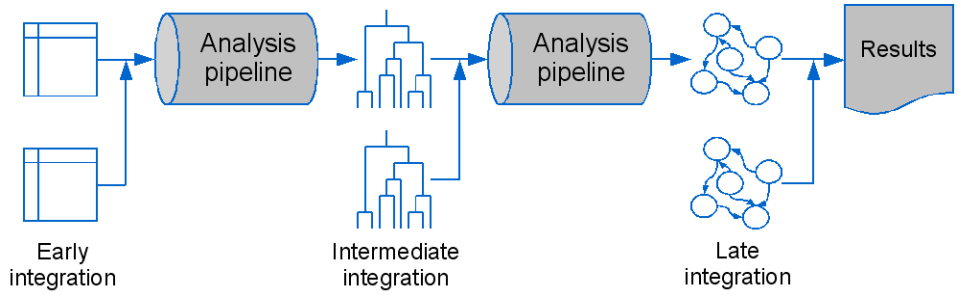


Figure 1.4: The different levels at which data integration can occur. During biological data analysis three phases of data integration can be distinguished: early, intermediate, and late integration. The three phases correspond to the distinction between data, information, and knowledge as depicted in Figure 1.1.

1.5 Early, intermediate, and late data integration

In summarizing the context of the presented work, high-throughput experimental technologies spawn ever growing amounts of data about genes and proteins. This causes a shift in focus towards the functional characterization of groups of genes. Hence, efficient data integration becomes the bottleneck of biological research. The downside of high-throughput analyses is the introduction of noise in the data. Therefore, better (statistical) validation procedures become necessary. Furthermore, availability of more data and broadening of the research scope towards the study of complex biological processes make data reduction, like data and text mining approaches, indispensable in future biological research.

With this context in mind, different data integration approaches for early, intermediate, and late integration were developed, all in the framework of characterizing large groups of genes. The different stages of integration correspond to the different stages in the knowledge acquisition cycle to go from experimental data to new biological knowledge.

Exploration of a large gene-centered data set almost always starts with a cluster analysis. This is done to find similar patterns in the data that can give a clue about, for instance, shared functionality between genes, or about possible connections between genes and the biological process or disease under investigation. Existing knowledge about the genes can be used to supervise the cluster analysis and improve the functional coherence of the obtained clusters. In the framework of this thesis, a method was developed to combine gene expression and literature data (see Chapter 2), but the proof-of-concept study was unable to verify improvement of the results.

Once interesting gene groups are found (for instance, based on statistical properties of the clusters), they can be further validated from a biological point of view. In most of the cases, a researcher wants to establish the biological properties of a gene group in a fast and efficient way. Because information about a group of genes is rarely available, most methods to characterize gene groups rely on the properties of its constituent genes.

In the framework of this thesis, two methods were developed to characterize gene groups. The first uses statistical analysis of the Gene Ontology annotations of genes to define the most characteristic properties of the group. This method was implemented by the author as a web service and integrated in the INCLUSive suite of services for gene expression and regulatory sequence analysis, which has been published in Nucleic Acids Research [28].

The other method combines textual information about individual genes to create a textual profile of a gene group. The method efficiently visualizes the most important terms of a gene group and even allows a closer examination of subgroups through subclustering. Figure 1.5 shows an example of the typical output of TXTGate, a web-based application implementing this method. Both the method and web interface were developed by the author in a collaboration with Patrick Glenisson and Steven Van Vooren. The work has been published in Genome Biology [56] and was presented by the author at the First International Workshop on Semantic Web and Databases [55]

After interesting gene groups are validated with existing biological information and a former research question is potentially answered, the time comes to start generating new hypotheses. Starting from a validated gene group, the question rises what other genes might also be part of the biological process the group represents.

Up to now, only two types of information were integrated: one type of experimental data with one type of existing knowledge. Part of this thesis work went into investigating if it is possible to combine numerous complementary data sources to get a more holistic model of a gene group

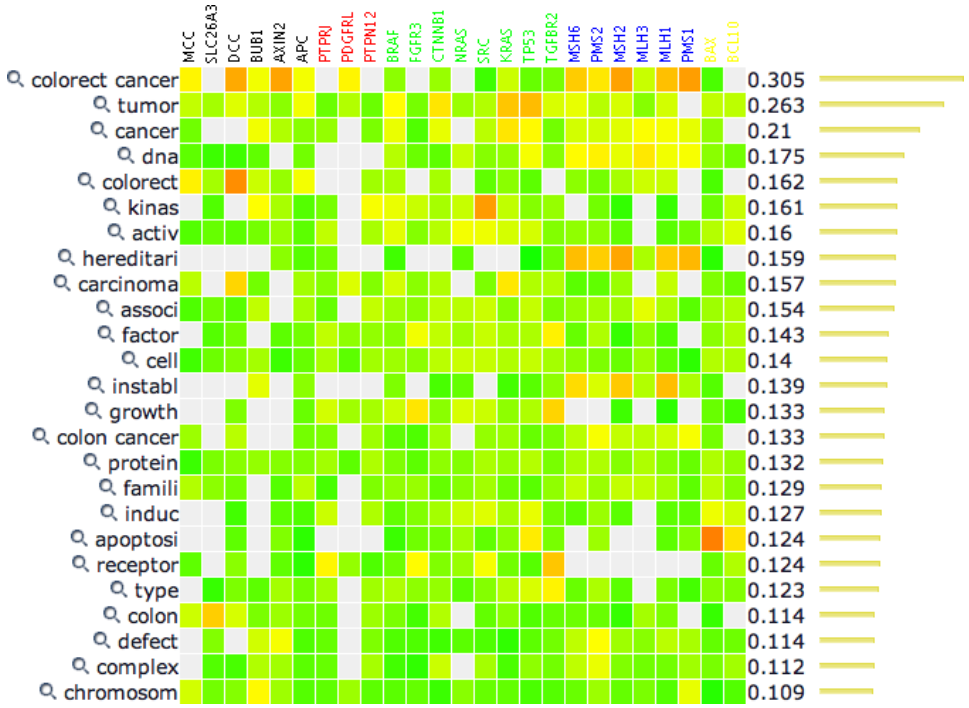


Figure 1.5: Example textual profile from TXTGate. This visualization was created by profiling a gene group involved in colon and colorectal cancer (see Appendix B) with the TXTGate application. TXTGate provides a nice and quick overview of the most important features of the gene group and allows an in-depth inspection of the textual profile through subclustering.

and use this model to find new genes that might be involved in the same process. Exactly this was the goal of the Endeavour project that was worked out in close collaboration with Stein Aerts. A firm statistical framework based on order statistics was developed to reconcile various heterogeneous, and often contradictory, data sources. A large-scale cross-validation on 29 disease and 3 pathways was performed with promising results, as can be seen in the Rank ROC curve in Figure 1.6. This work has been published by the author in Nature Biotechnology [3].

1.6 Web services integration

The ever increasing amount of biological data and knowledge, its heterogeneous nature, and its dissemination all over the Internet, make efficient data retrieval a horrendous task. Biological research has to deal with the diversity and distribution of the information it works with. Yet, access to a multitude of complementary data sources will become critical to achieve more global views in biology, as is expected from systems biology. To tackle this problem, web services technologies were introduced in bioinformatics.

Web services enable a uniform way of communication between users and providers of biological data and analytical services. A formal web service description ensures correct invocation. Besides, many efforts are being made to add a semantical, ontology-based layer on top of the web services technology to allow automated discovery of data- and task-specific services.

In the framework of this thesis, many web services were implemented to support execution of the described methods. Several software platforms that were developed in collaboration with colleagues, rely heavily on the web services architecture that resulted from this thesis work. The web services give both access to several in-house developed algorithms (like the algorithms in the INCLUSive suite [28], the ANOVA-based Maran algorithm for normalization of microarray data [39], and the algorithms for regulatory sequence analysis within the Toucan application [7]), as well as to custom-built data representations (especially for building data models of groups of genes in the Endeavour application [3]).

1.7 Using textual knowledge in biological analyses

Despite the vast amount of raw data coming from high-throughput experimentation, biological research is still mainly knowledge rich and data poor [11]. This is reflected by the fact that most biological knowledge is cap-

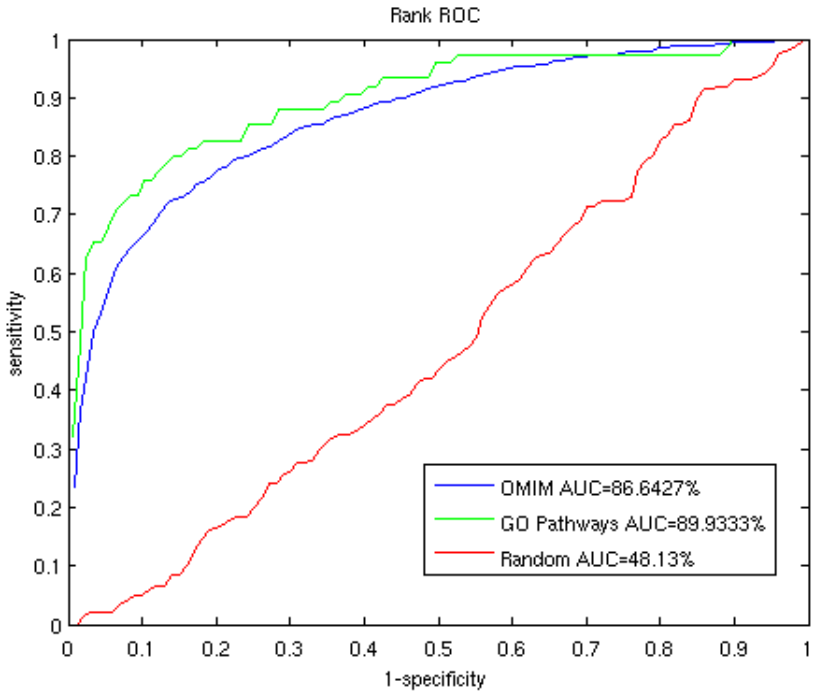


Figure 1.6: Rank ROC curve of the cross-validation. The figure shows the Rank ROC curves for the rankings of all leave-one-out cross-validations for the OMIM diseases and GO pathways study. The area under the curve of the plots is a measure of the performance of the method in finding back a gene that was left out of the original gene group and put in a group of 99 randomly selected test genes. The Rank ROC curve of the same leave-one-out cross-validation using random training sets is plotted in red. The cross-validation results in biologically meaningful results that are significantly better than random selections. Overall, the left-out gene ranks among the top 50% of the test genes in 85% of the cases in the OMIM study, and in 95% of the cases in the GO study. In about 50% of the cases (60% for the pathways), the left-out gene is found among the top 10% of the test genes.

tured in free-text descriptions and graphical representations, both knowledge representations that are hard to use in a formal, computational framework.

As the Internet became a widespread tool to share scientific knowledge, a big effort went into making knowledge captured in the scientific literature electronically available. The renowned PubMed system, for instance, contains already more than 15.5 million abstracts (as of April 2005) and is queried on average 60 million times a month. Moreover, there is a tendency towards new business models for publishers of scientific journals to have an open access policy. BioMed Central (BMC), for example, is a commercial publisher of online biomedical journals that provides free access to articles and even makes its entire open access full-text corpus available in a highly structured XML version for use by data mining researchers [19]. Open access publication guarantees that the published material is free of charge *and* available in a standard electronic format from at least one online repository (as described in the *Bethesda Statement on Open Access Publishing* [127]). An example of such a repository is NCBI's PubMed Central (PMC) [46] that contains over 350,000 full-text articles of over 160 different journals (as of April 2005).

With scientific papers publicly available, the difference between fetching the results of a database query and retrieving an article from an online repository is fading [52]. In fact, ongoing data integration efforts will result in the combined representation of database entries with knowledge captured in free-text descriptions. The manually obtained GeneRIFs (Gene Reference Into Function) present in the Entrez Gene database are a preview of this approach. GeneRIFs are concise functional descriptions of genes that link directly to the articles outlining these functions. Another example of this trend are the richly documented web supplements accompanying a scientific publication that allow a virtual navigation through the presented results (see for example the publication by Dabrowski *et al.* [32]).

It can be stated that a vast (and ever growing) amount of biological knowledge is captured in specialized literature and free-text descriptions. This information steadily becomes more accessible, not only to interested readers, but also to computerized analyses.

1.7.1 Short overview of molecular biology text mining

The efforts in biological text mining fall into four different categories: Information Retrieval (IR), Entity Recognition (ER), Information Extraction (IE), and Knowledge Discovery (KD). A basic overview of the different methods used in these categories is given by Shatkay and Feldman [119]. For a

more comprehensive overview, the reader is referred to Jensen *et al.* [68], and Krallinger and Valencia [76].

Information retrieval

Information retrieval (IR) is concerned with the identification of text bodies or segments relevant to a certain topic of interest. The identification can be based on a keyword query or on one or more related papers. Without any doubt the best-known and most-used biomedical IR system is PubMed, the official query interface to the MEDLINE database. Some research groups tried to improve the retrieval capabilities by adding query expansion rules, part-of-speech tagging, and entity recognition [129, 93]. Others tried to expand the functionalities of the interface by building a layer on top of the PubMed system (most notably HubMed [102]).

Entity recognition

Entity Recognition (ER) focuses on identifying biological entities in text (the names of genes or proteins, for instance). Methods are either based on machine-learning algorithms or on working with dictionaries. Often dictionary matching is combined with rule-based or statistical methods to reduce the number of false positives. Evaluation of the current status of ER was one of the two tasks of the BioCreAtIvE initiative [62]. ER's main problem is the lack of standardization in naming biological entities. Standardization of human gene names is the main focus of the HUGO Gene Nomenclature Committee (HGNC). By giving every human gene a unique and meaningful name and symbol, they hope to achieve less ambiguity and facilitate entity retrieval from publications considerably. The gene symbol list provided by the HGNC will be used further on in this thesis.

Information extraction

In Information Extraction (IE), the purpose is to derive predefined types of relations from text. This can be done based on gene/protein co-occurrence or on Natural Language Processing (NLP). In co-occurrence analysis the nature of the relation between two entities is less important than the fact that they are related. In Chapter 4 this concept of co-occurrence is extended to retrieve indirect but potentially interesting relations between human genes, thus being a means for knowledge discovery. NLP methods rely on part-of-speech tagging and ER to identify the syntax and semantic constituents of individual sentences. The method is unable to extract relations that span

multiple sentences. It is foreseen that IE will probably play an important role in systems biology, because of its ability to identify diverse types of relations on a large scale (the entire MEDLINE collection, for instance) [68].

Knowledge discovery

The Holy Grail of Knowledge Discovery (KD) is to discover new, previously unknown information through textual analysis of written information sources. KD's focus is on inferring indirect relations between genes or proteins (rather than relations between co-occurring genes, which is the focus of IE). The field can be divided in *closed* (Arrowsmith [120] and HyBrow [105], for instance) and *open* discovery approaches (which are much more challenging)[1]. Practice learns that KD through text-based analysis alone has a hard time coming up with unknown, non-trivial relations. Integrated approaches, being the topic of this thesis, are believed to have a much greater potential in discovering new biologically relevant relations.

1.7.2 The vector space model

To use the knowledge captured in biomedical literature during the analysis of biological data, it needs transformation into a format amenable to computation. A computational approach that appeared quite successful in transforming textual information is based on the concept of a vector space. In this vector space a document is represented as a vector, which allows the application of standard linear algebra techniques [16]. The *vector space model* allows extraction and transformation of information from a set of documents, referred to as the *corpus*. A document is transformed into a vector of which each component contains a weight that indicates the importance of a certain term with respect to the document. In other words, a literature corpus comprising n documents and k different terms can be represented as an $n \times k$ document-by-term matrix of which each component w_{ij} (with $0 < i < n$ and $0 < j < k$) is the weight of term t_j in document d_i (Figure 1.7). A term can be either a single word or a so called *phrase*, a sequence of words that represents a single concept. Calculation of the weights for all terms in the corpus is called *indexing*. The dimension k depends on the number of terms that are considered during the indexing process. Since all

[1] A closed discovery approach starts with two topics and tries to find indirect and yet unknown connections between these topics. An open discovery approach starts with only one topic and tries to find indirectly connected topics via the topics directly connected to it.

structure in the text is obliterated, this procedure is called the *bag-of-words* approach.

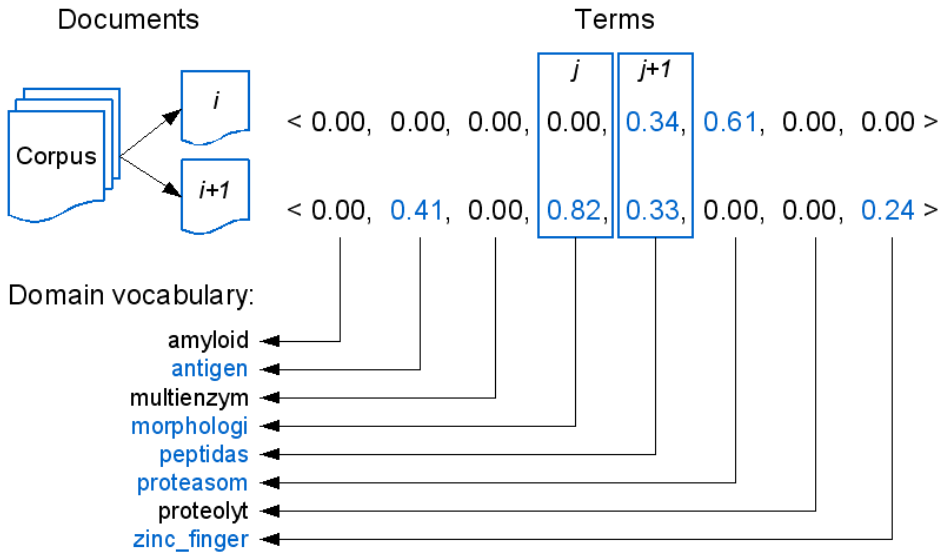


Figure 1.7: Illustration of the term index of a given document. Document i contains the terms *peptidase* and *proteasome* (the ones with non-zero weights). The set of all terms is called a vocabulary. Typically stop words such as *from*, *the*, *often*, etc. are removed. Note that keywords are matched according to their stemmed form.

To get a more precise reflection of the frequencies of a corpus' concepts, the morphological and inflectional endings (for instance, plurals, tenses, and so on) of all its terms can be removed in a process called *stemming*. Stemming helps to reduce to a certain extent the dimensionality as well as the dependency between words. In this thesis, standard English stemming with Porter's method [101] was applied on most occasions. A further noise reduction was achieved through the use of domain vocabularies (see below) and predefined *stop-word* and *synonym* lists.

Terms can be weighted according to a given weighting scheme that contains local weights (i.e., weights derived from term usage in one document), global weights (i.e., weights derived from term usage in the entire corpus), or a combination of both. *Boolean weighting* is the most straightforward scheme and is based on a local weight: if a term occurs in a document, w_{ij} is 1; if not, w_{ij} equals 0. A more refined local weight is the *Term Frequency* or TF that is defined as the number of times n_{ij} a term t_j occurs in

a document d_i, divided by the total number of terms N_i in that document:

$$w_{ij}^{\mathrm{TF}} = \frac{n_{ij}}{N_i}. \tag{1.1}$$

The weighting scheme used throughout this thesis is based on a global weight called the *Inverse Document Frequency* or IDF. The scheme proportionally weights down terms that occur often in the corpus and is defined as

$$w_{ij}^{\mathrm{IDF}} = log(\frac{N}{n_j}), \tag{1.2}$$

where n_j is the number of documents that contain term t_j in the collection of N documents. It accounts for the assumption that common terms (i.e., terms that recur in a lot of documents) are less interesting to characterize a document than rare terms that only occur in some documents. Since this weighting scheme is based on a global weight the term weights of a document are independent of the document's own term usage.

An more complex weighting scheme that is frequently used in information retrieval combines the TF local weight with the IDF global weight of a term to yield TF-IDF term weighting:

$$w_{ij}^{\text{TF-IDF}} = w_{ij}^{TF} w_{ij}^{IDF}, \tag{1.3}$$

Stemming a corpus and indexing with the IDF scheme is a reasonable choice for modeling pieces of text comprising up to 200 terms, as is observed in the database annotations and MEDLINE abstracts used throughout this thesis. Therefore, the IDF scheme was preferred over other weighting schemes in developing the methodologies described further on.

Once a corpus is represented this way, all basic vector operations can be used to work with the indexed information. The geometrical relations between document vectors can be exploited to model a document's semantics. Among the possibilities are similarity measurements (for searching or document retrieval), cluster analyses (see Section 2.3), creation of entity indices (see Section 1.7.4), as well as more advanced operations such as dimensionality reduction (see Section 1.7.5).

1.7.3 Document similarity

In the vector space model, the cosine of the angle between the vector representations of two documents d_1 and d_2 can be used to represent their semantic

similarity:

$$\mathrm{Sim}(d_1, d_2) = \cos(d_1, d_2) = \frac{\sum_j w_{1j} w_{2j}}{\sqrt{\sum_j w_{1j}^2}\sqrt{\sum_j w_{2j}^2}}. \tag{1.4}$$

This measure takes values between 0 and 1: the closer to 1, the more similar the two documents[2]. The underlying hypothesis is that documents sharing a lot of important words (i.e., with a high weight) are semantically connected.

1.7.4 Construction of an entity index

Depending on the research issue at hand, abstractions of different biological entities (such as genes, proteins, diseases, and so on) need to be made. An entity can be represented in the vector space model by combining all indices of the documents[3] that describe it into one summarized entity index. For instance, in the case of a gene, all documents describing it can be indexed. The average of the resulting term vectors can then be used as a textual profile to characterize this gene.

The text index of an entity i is defined here as the vector with terms t_j obtained by taking the average over the N_i indexed documents annotated to it:

$$g_i = \{g_i\}_j = \{\frac{1}{N_i}\sum_{k=1}^{N_i} w_{kj}\}_j. \tag{1.5}$$

Equation 1.5 pools the keyword information contained in all documents related to an entity into a single term vector. As a result, documents describing the same entity and containing different but related terms are joined.

1.7.5 Dimensionality reduction

Dimensionality reduction is the process of lowering the dimensionality of a matrix, thus removing redundant information and *noise* from it. In the context of text mining, this involves reducing the dimensionality of the term-by-document matrix (constructed as described in Section 1.7.2).

[2]In theory, a cosine can have values between -1 and 1. Since in this case a vector only consist of positive weights, all vectors are located in the first quadrant of the vector space. Hence, the cosine will never be negative.

[3]The term *document* has to be interpreted in a general sense. It denotes a journal publication as well as a functional summary, a paper abstract, an annotation description, etc.

Latent Semantic Indexing (LSI) is the best-known technique for reducing the dimensionality of a term-by-document matrix. It is based on a Singular Value Decomposition (SVD) of the matrix and was first described by Deerwester *et al.* [33]. LSI decomposes both the term and document space the matrix encompasses into linearly independent components or *factors*. The term space is the space where the terms are the dimensions and in which the document vectors lie. The document space is the space where the documents are the dimensions and in which the term vectors lie. To reduce the dimensionality of the new vector space that comprises the calculated factors, all reasonably small factors are ignored.

LSI takes advantage of implicit higher-order structure in the associations between terms and documents. It tends to map semantically similar terms into the same factor and identical terms with different meaning into different factors, thus resolving both *synonymy* and *polysemy* problems. Especially with respect to gene name synonymy, this is an important benefit. Table 1.1 lists, for example, several phrases used to denote the human gene *IFNB1*. If these phrases have a similar context of associated terms in different documents, their vectors will be mapped onto the same factor.

Table 1.1: Synonyms of the human gene *IFNB1*. Listed are several phrases that are used to denote the human gene *IFNB1*, as an example of the typical problem of gene synonymy biomedical text mining research faces. Latent Semantic Indexing is a methodology to decompose a term-by-document matrix into linearly independent components that tends to project synonyms onto the same component, thus also reducing the term space of the matrix.

interferon-beta, beta-interferon, fibroblast interferon, interferon beta, beta 1 interferon, interferon beta1, beta interferon, beta-1 interferon, interferon beta 1, interferon-beta1, ifn-beta, fiblaferon, interferon fibroblast, ifnbeta, interferon beta-1

In this thesis, reduction of the term space was done with domain vocabularies rather than with LSI. Working with domain vocabularies has several advantages, as explained in the next section.

1.7.6 Domain-specific views

The use of domain vocabularies to index a corpus can be seen as a way to reduce the dimensionality of the resulting vector space. A domain vocabulary determines the focus of the analysis by restricting the indexing process

to only the terms and phrases it contains. To show the effect of the use of a domain vocabulary on the indexing process, a group of genes related to colon and colorectal cancer was profiled with four different vocabularies. The complete list of used genes can be found in Appendix B. It was constructed by fetching all genes related to colon and colorectal cancer from the Online Mendelian Inheritance in Man (OMIM) database. The results are presented in Table 1.2.

The *GO* domain vocabulary is derived from the Gene Ontology (GO) [132] structured vocabulary and contains 17,965 terms. Since GO is considered the reference vocabulary for annotation purposes in the life science and in genetics in particular, it as an ideal source from which to extract a highly relevant and relatively noise-free domain vocabulary. All composite GO terms shorter than five tokens were retained as phrases. Longer terms containing brackets or commas were split to increase their detection. The *MeSH* and *OMIM* domain vocabularies are rather similar in scope but differ in size. The former is based on MeSH, the National Library of Medicine's controlled vocabulary thesaurus *Medical Subject Headings* [95], and counts 27,930 terms. The latter is based on OMIM's Morbid Map [88]. This is a cytogenetic map location of all disease genes present in the OMIM database. All disease terms were extracted to construct a 2,969-term vocabulary. The *eVOC* domain vocabulary was drawn from eVOC [74], a thesaurus consisting of four orthogonal controlled vocabularies encompassing the domain of human gene expression data. It includes terms related to anatomical system-, cell type-, pathology-, and developmental stage.

As can be seen, there is little difference between the *MeSH* and *OMIM* profiles, whose terms are mainly medical- and disease-related (*colorect cancer*, *colon cancer*, *colorect neoplasm*, *hereditari*), whereas the focus of the *GO* profile is on metabolic functions of genes (*mismatch repair*, *dna repair*, *tumor suppressor*, *kinas*) and the *eVOC* profile contains more terms related to cell type and development (*growth*, *cell*, *carcinoma*, *metabol*, *fibroblast*).

1.8 Thesis overview

The rest of this thesis is structured as follows: in Chapter 2 two example gene cluster analyses are performed. The first is based on experimental data, the second on known information about genes derived from paper abstracts. In a third cluster analysis, both experimental data and textual information of genes is combined and the results are statistically validated to proof the validity of this approach. Chapter 3 represents the step in

Table 1.2: Different domain vocabularies give various perspectives on textual information. The table shows how term-centric GO-, OMIM-, MeSH-, and eVOC-based vocabularies profile a group of genes involved in colon and colorectal cancer.

GO	**OMIM**	**MeSH**	**eVOC**
mismatch repair	colorect	colorect neoplasm	colorect
tumor	colorect cancer	mismatch	tumour
dna repair	tumor	cancer	malign tumour
mismatch	kinas	colorect	colon
pair	colon	mutat	growth
tumor suppressor	hereditari	repair	cell
apc	cancer	dna repair	carcinoma
kinas	colon cancer	colon	metabol
somat	associ	neoplasm protein	fibroblast
ra	on	tumor	chain

the knowledge acquisition cycle where experimental results are verified with existing knowledge. Several methods are presented to efficiently characterize groups of genes. To illustrate the methods, statistically validated gene groups from Chapter 2 are processed with the methods and the results are shown. Chapter 4 presents two methods designed to generate new hypotheses under the form of potential relations between genes and biological processes. The methods are illustrated with validated gene groups from Chapter 3. The groups are used to find other genes potentially related to the same biological process. Chapter 5 goes into detail about web services technologies and the important role they play in assuring access and efficient retrieval of biological data. In Chapter 6 the achievements of this work are presented together with future prospects.

Chapter 2

Grouping genes

WHILE in the recent past research was focussed on investigating functions of individual genes and proteins, the availability of entire genomes (311 completed, 244 draft assemblies, and 515 in progress, as of January 2006 [40, 15]) now allows adoption of more holistic approaches. When trying to understand functional behavior of genes at a higher level, the first endeavor is to group genes involved in the same biological pathways or processes. Cluster analysis of gene expression data is one way to do this. The rationale is that functionally related genes (i.e., involved in the same cellular process) might be co-regulated and, thus, have a similar gene expression profile; or, put the other way around, that genes with similar expression profiles might be functionally related. This way of inferring biological function of genes is known as the *guilt-by-association* (GBA) heuristic and seems to be broadly applicable in co-expression analyses [104, 151].

This chapter represents the first step in the knowledge acquisition cycle (Figure 2.1). An experiment is being set up and performed to gain new information about a certain biological process or about an entire genome. The purpose of this chapter is to exemplify this first step by describing the cluster analysis of a set of genes starting from several different data sources. The subsequent steps in those analyses are highlighted, from preprocessing over clustering to selecting gene clusters of high quality.

In Section 2.2, a genome-wide cluster analysis based on gene expression data is described by way of illustration. The gene expression data were taken from a microarray experiment conducted by Su *et al.* [126]. Section 2.3 describes the clustering of the same set of genes based on textual data to demonstrate that an *in silico* cluster analysis is as good an experiment as the microarray experiment which was conducted in a wet-lab environment.

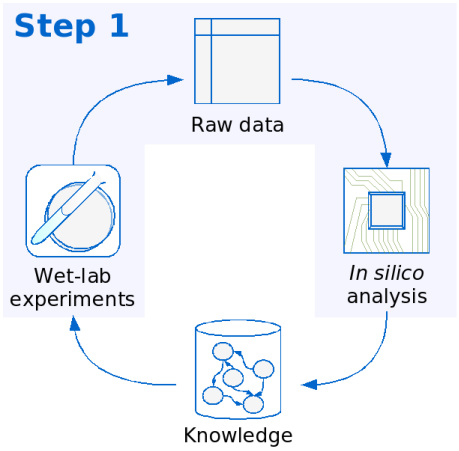


Figure 2.1: Step 1 in the knowledge acquisition cycle. The first step comprises preparation of experimental data and extraction of preliminary results for further validation.

As more data from high-throughput analyses come in the public domain, *in silico* experiments might become a major part of biological experimentation [58]. These two cluster analyses try to exemplify two different approaches towards grouping of genes: one based on experimental data that is equally valid for well-known as well as unknown genes; the other based on existing information about known genes only. Section 2.4 elaborates on combining expression and textual data to cluster genes. Combining experimental data (gene expression data, for instance) with biological knowledge (textual data, for instance) can be seen as a methodology in which the validation step (see Chapter 3) is inherently present in the cluster analysis. The method described here is an example of an early integration approach (see Figure 1.4).

2.1 General-purpose data set

Throughout this thesis, the same data set will be used in examples. This data set is derived from the experiments done by Su *et al.* [126]. They constructed a gene atlas of human (and mouse) protein-encoding transcriptomes by measuring expression patterns of 44,775 transcripts in 79 different human tissues. From this atlas, a selection of 3,989 genes was made, mostly based on the availability of Gene Ontology and literature annotations. This

set of genes will be referred to as the *general-purpose gene corpus*.

2.2 Grouping genes based on expression data

From the introduction of microarray technology in the beginning of the nineties, grouping genes based on expression data was believed to have the potential of identifying efficiently genes of similar function. This was discussed in a landmark paper by Eisen *et al.* [38] in which hierarchical clustering was combined with the presently famous visual red-green representation (see Figure 2.2).

It is not the purpose of this thesis to detail out all possible strategies for analyzing microarray data and clustering genes based on expression data. Rather, a practical example of a common analysis is given for illustration purposes. The outcome of this analysis will be used in the next chapters. For a more elaborate discussion, the reader is referred to the review papers by Quackenbush [103] and Moreau *et al.* [89].

To obtain groups of functionally related genes, the expression profiles of all 3,989 genes of the general-purpose data set were retrieved from the Su *et al.* gene atlas. After preprocessing the data, the profiles were used to perform a hierarchical clustering.

2.2.1 Preprocessing

Microarray measurements are known to be of low *absolute* quality. Therefore, prior to cluster analysis, some additional data manipulation steps are necessary.

First, all missing (or NaN) values present in the expression profiles of the general-purpose gene corpus were replaced by the profile's mean. If a gene was measured more than once (i.e., if more than one gene expression profile was available) the average of all profiles was taken.

Secondly, all profiles were mean-centered and variance-normalized to remove all absolute differences in gene expression behavior. It is believed that functionally related genes share the same relative behavior because they are up- and down-regulated together, regardless of their absolute expression levels. The profile of gene i, $x_{i.} = (x_{i1}, x_{i2}, \ldots, x_{ip})$ with p elements, is rescaled by subtracting from each element x_{il}, $l = 1 \ldots p$, the profile's mean $\mu_i = \bar{x_i} = \frac{1}{p}\sum_{l=1}^{p} x_{il}$ and dividing the result by the profile's standard deviation $\sigma_i = \sqrt{\frac{1}{p}\sum_{l=1}^{p}(x_{il} - \bar{x_i})^2}$:

$$\hat{x}_{il} = \frac{x_{il} - \mu_i}{\sigma_i} \tag{2.1}$$

The resulting profile has zero mean and unit variance.

2.2.2 Cluster analysis

Cluster analysis was performed with a hierarchical clustering methodology. The distance measure used was the Pearson correlation between two expression profiles. For two genes i and j with expression profiles $x_{i\cdot}$ and $x_{j\cdot}$, the Pearson correlation is defined as

$$s_{\text{Pearson}}(i,j) = \frac{\sum_{l=1}^{p}(x_{il} - \bar{x}_i)(x_{jl} - \bar{x}_j)}{\sqrt{\sum_{l=1}^{p}(x_{il} - \bar{x}_i)^2 \sum_{l=1}^{p}(x_{il} - \bar{x}_i)^2}} \tag{2.2}$$

with $\bar{x}_i$ and $\bar{x}_j$ the mean of $x_{i\cdot}$ and $x_{j\cdot}$, respectively. Because the profiles have zero mean and unit variance, s_{Pearson} is equivalent to s_{Cosine} in this context.

Hierarchical clustering organizes elements into a binary tree in a process called *linkage*. In this case, an agglomerative method was used (i.e., a method that starts with all elements in a separate cluster and gradually combines these atomic clusters until all elements are merged). The cluster analysis was started with the calculation of an upper-triangular distance matrix containing the mutual distances between all profiles, as given by $d_{\text{Pearson}} = (1 - |s_{\text{Pearson}}|)$. The distance matrix was then fed to the linkage algorithm. During every iteration of the algorithm the two closest clusters (i.e., the ones with the smallest distance between them) were grouped and the distance matrix was updated according to Ward's minimum variance method. This method specifies the distance between two elements/clusters as the increase in the *error sum of squares* (ESS) when they are combined. The ESS of a cluster x is the sum of squares of its n_x elements' deviations from the mean and can be written as

$$\text{ESS}(x) = \sum_{i=1}^{n_x} |x_i - \frac{1}{n_x}\sum_{j=1}^{n_x} x_j|^2. \tag{2.3}$$

Ward's linkage defines the distance $d[r,s]$ between two clusters r and s as

$$d[r,s] = \text{ESS}(r,s) - [\text{ESS}(r) + \text{ESS}(s)] \tag{2.4}$$

with $\text{ESS}(r,s)$ the ESS of the combined cluster of all elements in r and s.

Ward's linkage strives to minimize the increase in $d[r, s]$ during every iteration. The method creates a tree with evenly distributed branches from which compact, spherical clusters of similar size can be retrieved. The heat-map representations of certain parts of this tree are visualized in Figure 2.2.

Instead of searching for an optimal number of clusters to cut the tree, an optimal cluster size was chosen, acknowledging that a group of 100 or more genes rarely contains valuable biological information. To define a more interesting estimated number of genes per functional module, the average number of genes from all pathways in the HumanCyc Pathway/Genome Database [112] was calculated and found to be approximately ten genes. Gene groups of this size better reflect the complexity of biological processes at an intermediate level (i.e., the level of interest in this thesis). Therefore, all possible leaves in the cluster tree comprising 10 to 20 genes were retained for further analysis. A further selection was made based on the *Silhouette coefficient*, a statistical index of cluster quality, as described in the next paragraph.

2.2.3 Cluster quality

The *Silhouette coefficient* can assess the quality of a clustering. It is an internal index (i.e., a score that measures how good the clustering fits the original data based on statistical properties of the clustered data). External indices, by contrast, measure the quality of a clustering by comparing it with an external (supervised) labeling (see Section 2.3.3).

The Silhouette coefficient of an element i of a cluster k is defined by the average distance $a(i)$ between i and the other elements of k (the *intra*-cluster distance), and the distance $b(i)$ between i and the nearest element in the nearest cluster (i's minimal *inter*-cluster distance):

$$\mathrm{sc}_i = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{2.5}$$

An overall score for a set of n_k elements (a cluster or the entire clustering, for instance) is calculated by taking the average of the Silhouette coefficients sc_i of all elements i in the set:

$$\mathrm{SC}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathrm{sc}_i. \tag{2.6}$$

The Silhouette coefficient takes values between -1 and 1. The closer to 1, the better the clustering fits the data. Table 2.1 lists a general rule of thumb on how to interpret the Silhouette coefficient.

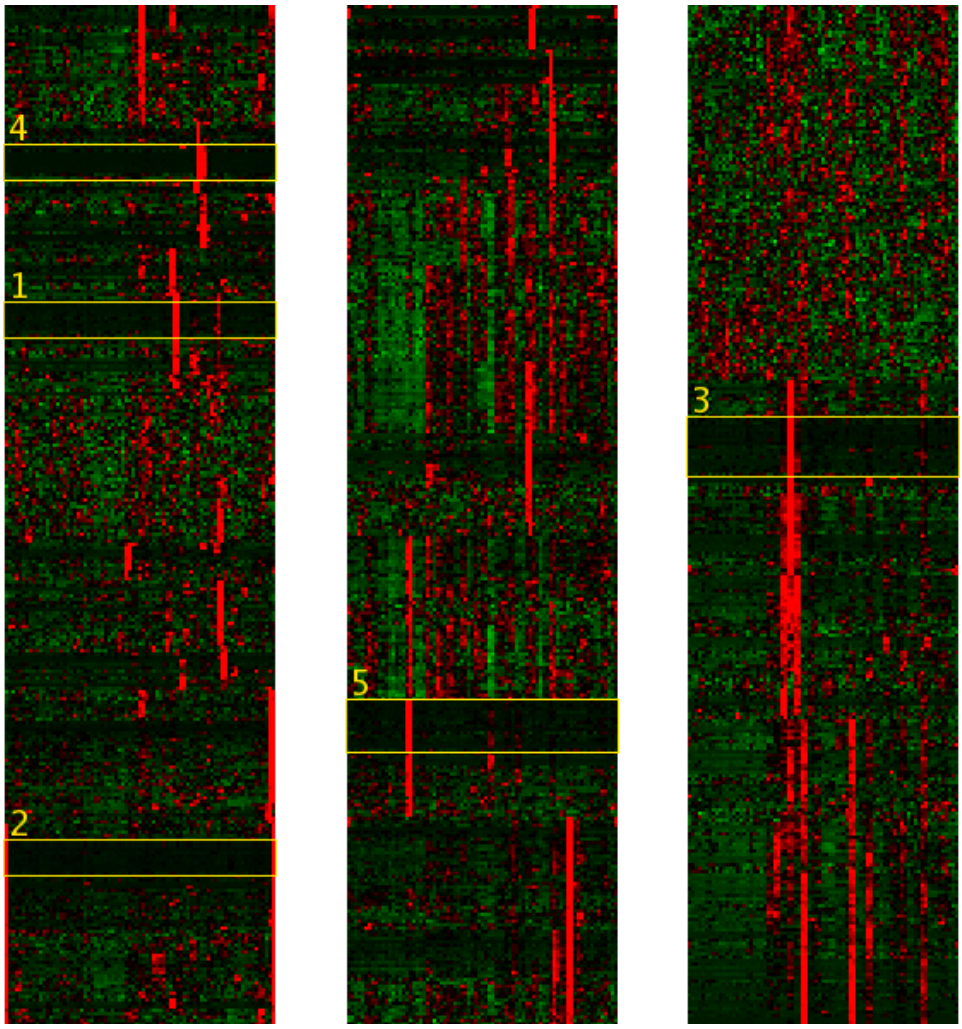


Figure 2.2: Heatmap visualization of the hierarchical tree based on expression data. The 3,989 gene expression profiles were linked using Ward's minimum variance method. The Pearson correlation between the profiles was chosen as the distance measure. Only a small part of the entire tree is shown. The rows represent the genes; the columns represent the conditions. The color at each position gives an indication of a gene's expression in a certain condition: green indicates the gene is down-regulated in this condition, red indicates the gene is up-regulated, black means the gene is not expressed. The five clusters with highest Silhouette coefficient are marked in yellow. The visualization was created with Java TreeView [114].

Table 2.1: Rule of thumb for the interpretation of the Silhouette coefficient.

Range	Interpretation
> 0.70	strong structure has been found
0.50-0.70	reasonable structure has been found
0.25-0.50	the structure is weak and could be artificial
< 0.25	no substantial structure has been found

The overall Silhouette coefficient of the clustering performed in Section 2.2.2 is 0.0896. This rather low figure indicates that the clustering does not fit the data well. Hierarchical clustering of microarray gene expression data forces every gene in a cluster, often resulting in heterogeneous clusters of low value. Nevertheless, some of the clusters will be coherent and suitable for further analysis.

For the selection of high quality clusters, the tree was cut at all possible levels to yield a number of clusters from 1 (all genes in one cluster) up to 3989 (all genes in a separate cluster). At every level, all clusters that contained 10 to 20 genes were recorded together with their Silhouette coefficients. Note that the exact same cluster can have different Silhouette coefficients for different clustering results of the same set of genes. The 5 clusters with the highest average coefficient are depicted in Table 2.2. In the case of clusters with the same base (i.e., clusters that share the same set of 10 genes) only the cluster with the highest average Silhouette coefficient is shown. These clusters are selected for later use, on the one hand to illustrate the methods described in the following chapters, on the other hand to investigate the correlation between the statistical quality of a cluster and its functional coherence.

2.2.4 Discussion

Following the GBA heuristic, analysis of gene expression data at first sight yields biologically relevant gene groups. However, it is clear that manual investigation of every cluster is not only very labor intensive, but always biased by the investigator's own background knowledge. A first selection can be made based on statistical properties, as was done above using the Silhouette coefficient. However, because gene expression data is known to be of low quality, a proper biological validation is mandatory. Validation defines how biologically meaningful a gene group is. This is the topic of the next chapter. In the next chapter, the correlation between the statistical

Table 2.2: Gene clusters with highest average Silhouette coefficient based on expression data. The table contains information about the five clusters with the highest average Silhouette coefficient. The clusters contain 12.4 genes on average. For genes without HUGO gene symbol, the Ensembl identifier is given.

Nr.	SC	Size	Genes
1	0.7499	10	*C1orf10*, *EVPL*, *KRT13*, *LY6D*, *RHCG*, *S100A7*, *SLURP1*, *SPRR1A*, *SPRR1B*, *SPRR2B*
2	0.6698	10	*CGB1*, *CGB2*, *CRH*, *CYP19A1*, ENSG00000124467, ENSG00000183668, *KISS1*, *PSG1*, *PSG4*, *PSG5*
3	0.6376	17	*CD160*, *CST7*, *CTSW*, ENSG00000129277, *GNLY*, *GZMA*, *IL18RAP*, *IL2RB*, *KIR3DL3*, *KLRC1*, *KLRD1*, *KLRF1*, *PTPN4*, *SPON2*, *TBX21*, *XCL1*, *XCL2*
4	0.6333	10	*AMY1A*, *AQP8*, *CPA1*, *CPA2*, *CTRC*, *CTRL*, *ELA3B*, *PLA2G1B*, *PNLIPRP2*, *SERPINI2*
5	0.6123	15	*ACTC*, *CASQ2*, *CKMT2*, *COX6A2*, *COX7A1*, *CSRP3*, *HRC*, *HSPB7*, *ITGB1BP3*, *MYBPC3*, *NKX2-5*, *NPPA*, *TNNC1*, *TNNI3*, *TNNT2*

quality of a gene group and its *biological* quality will also be investigated.

2.3 Grouping genes based on textual information

As discussed before, the electronic availability of large amounts of biological data rapidly increases. This poses an unprecedented opportunity for biologists to perform *dry-lab* bioinformatics research. One of the challenges is to exploit the information captured in biomedical papers.

Figure 2.3 shows a histogram of the number of times a document is *annotated* to a gene in *Entrez Gene*, a database with information on genes defined by sequence that is part of the *Entrez* system. Clearly, most documents are annotated to only very few genes. On average a document is linked to 9.4 different genes. This rather high average is caused by several very general papers that are linked to as many as 40,000 genes but contain very few gene-specific information. Examples of these kind of publications are gene sequencing and identification efforts or large-scale functional annotation studies. The median of the number of links per document is only one. It can be stated that there is a lack of textual information describing the functionality of gene groups larger than 10 genes. Thus, the question rises if textual data can be used to find functionally related gene groups.

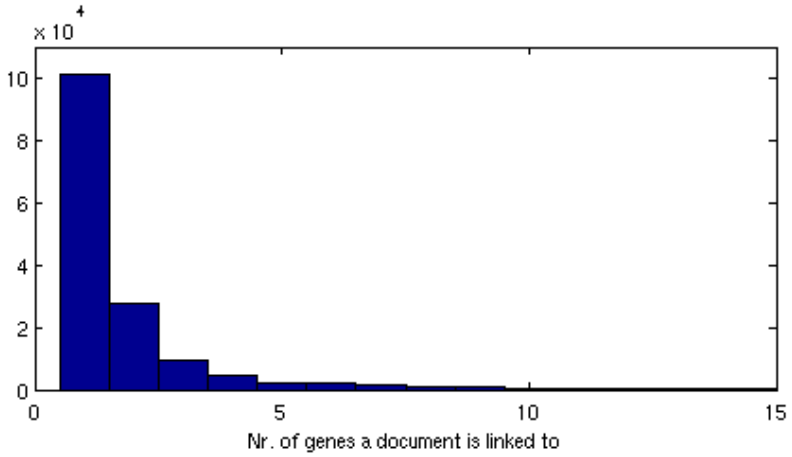


Figure 2.3: Histogram of the number of genes per publication. The histogram shows the distribution of the number of times a publication is annotated to a gene in the Entrez Gene database. The histogram only contains information about publications that both reside in PubMed and are connected to a gene in Entrez Gene. While some publications are linked to a large number of genes, most publications are linked to very few genes.

Several methods are described in the literature to group genes based solely on textual information. In most cases, these methods are used to help interpreting high-throughput data analysis results. Nevertheless, useful information can be derived from this kind of *in silico* analyses. Three categories can be distinguished:

Grouping based on co-occurrence This type of methods, also called bibliometric approaches, is based on the statistical analysis of co-occurrence of genes or keywords. It is assumed that co-occurrence of gene or protein names in the same sentence, abstract, and so on indicates a biological relation. One of the hurdles in this domain is the correct identification of biological entities in free text, an area of investigation on its own (for a review on this topic, the reader is referred to the methodological review by Krauthammer and Nenadic [77]). One of the aims of the BioCreAtIvE initiative [62, 18] was to provide a way to assess the ability of automated systems in finding genes and proteins in written text and to bring transparency in this field. The two most common ways to find gene and protein names in biomedical text are the use of curated thesauri and of *named entity tagging*, a method

that combines rule-based recognition with external knowledge.

Several approaches to grouping genes based on co-occurrence can be found in the literature. Stapley and Benoit [122] describe an approach to cluster yeast genes based on a dissimilarity matrix derived from the joint and individual occurrences of gene names in MEDLINE abstracts. They conclude that the retrieved associations between genes, although often not directly related to *in vivo* relationships, carry accurate information about biological processes. Jenssen *et al.* [69] create a genome-wide gene-to-gene co-citation network of human genes by linking all genes that co-occurred in titles and abstracts from MEDLINE records. They use the network to perform a supervised clustering of gene expression data and prove that it adequately represents the current knowledge about human genes. However, they do not use the network to group genes based on literature data. Wilkinson an Huberman [148] describe a method to partition a similar co-citation network into communities of related genes. Alako *et al.* [9] perform gene clustering based on an improved gene co-citation network. Since they also include gene-keyword co-occurrences, they can extend their analysis to find gene-pathway and gene-disease associations. The gene-keyword associations are also used in supervised clustering of gene expression data to improve the clustering results significantly.

Grouping based on linguistics In this type of methods, the nature of the biological connection between two entities is inferred via grammatical interpretation of sentences, also called Natural Language Processing (NLP). Cohen and Hunter [29] wrote a nice overview of the use of NLP techniques within genomics. As with the co-occurrence methods, the bottleneck for good performance is often accurate detection of biological entities.

The strength of NLP methods lies in the identification of the nature of relationships between small groups of co-cited biological entities, rather than in grouping large numbers of genes (possibly because NLP is a computationally intensive technique and because more efficient methods for finding groups of genes exist). Chen and Sharp [25] demonstrate a straightforward methodology in which they build biological networks between genes based on the interpretation of abstracts from MEDLINE. The abstracts are retrieved after querying PubMed with a set of user-specified terms.

Numerous other papers describe methodologies to extract from the

literature interactions between biological entities (in most cases gene-gene or protein-protein interactions, as with Textpresso [93] and PreBIND [35]) or to reconstruct pathways (as done by Friedman *et al.* [48]). In most cases the purpose is to help researchers cope with the vast amounts of literature coming to them (like with the iHOP system [63]), sometimes to automate population of specialized databases (like in the case of PRIME [75]). However, to quote Cohen and Hunter [29], "Unfortunately for impatient consumers—perhaps fortunately for curious scientists—NLP is approximately as difficult as it is important."

Grouping based on profiling The previous two categories heavily depend on correct identification of gene or protein names. This dependency can be removed by using explicit links between genes and the documents describing them. Profiling methods are based on the similarities between the information contained in these documents to connect genes. Hence, a relation between two genes can be extracted even if they do not co-occur. As described before, explicit links between genes and documents can be used to create textual profiles of genes (see 1.7.4). The vector space model then allows for efficient computation of similarities between genes.

Shatkay *et al.* [118, 117] introduce the concept of the *kernel document* (i.e., a document that can be treated as a representative of a certain gene). Based on a set of characteristic *Bernoulli* distributions that model the term usage in the kernel document, the likelihood that another document was generated by sampling from the same distributions is calculated. Based on this likelihood, a set N documents relevant to each gene is retrieved from MEDLINE. The PubMed identifiers of this set of documents are then put into a kernel vector that is used to calculate the similarity between the gene characterizing it and other genes that were represented in the same way. Glenisson *et al.* [54] investigated the use of typical information retrieval techniques in clustering genes. Instead of using kernel documents, a gene's functional information is retrieved from specialized databases and compiled into a vector representation. Several sources of information and weighting schemes are investigated. Homayouni *et al.* [65] use the technique of Latent Semantic Indexing (LSI) to represent a gene's information. The technique relies on a Singular Value Decomposition (SVD) to create a vector subspace in which genes are characterized by concepts rather than terms. The information about a gene is taken from the abstracts of MEDLINE documents linked to it in Entrez Gene.

The cluster methodology used in this thesis is part of this third category. To exemplify the methodology, the same 3,989 genes from the general-purpose gene corpus are clustered using literature data. For each of the 3,989 genes, all titles and abstracts of the MEDLINE documents linked to them in Entrez Gene, are retrieved. The information is indexed using a domain vocabulary derived from the Gene Ontology (GO). The resulting document profiles are then combined and normalized to obtain true gene profiles. These profiles are clustered as described below.

2.3.1 Cluster analysis

The genes are clustered using the cosine of the angle between their textual profiles as similarity measure. (Since the profiles are normalized, the cosine similarity is equal to the Pearson correlation and equivalent to the Euclidean distance.) The cosine similarity between two gene profiles g_1 and g_2 containing terms w_{ij} is defined as

$$\mathrm{s}_{\mathrm{Cosine}}(g_1, g_2) = \cos(g_1, g_2) = \frac{\sum_j w_{1j} w_{2j}}{\sqrt{\sum_j w_{1j}^2}\sqrt{\sum_j w_{2j}^2}}. \tag{2.7}$$

Linkage analysis was performed with Ward's minimum variance method and all possible clusters with 10 to 20 elements were selected for quality assessment and further analysis (as was done above based on the gene expression profiles). The complete tree is visualized in Figure 2.4.

2.3.2 Cluster quality

The statistical cluster quality is again determined with the Silhouette coefficient (see above). Table 2.3 shows the five clusters with the highest average Silhouette coefficient. The selected clusters will also be used later on to illustrate the developed methods for fast and efficient characterization of gene groups. It is expected that the clusters obtained from literature data will be more functionally coherent than their counterparts obtained via clustering of gene expression data. This will be investigated in the next chapter.

2.3.3 Comparison with grouping based on expression

The quality of a clustering can be measured by comparing it to a reference labeling. This kind of measurement is called an *external index* (as compared to an internal index that only takes into account the inner statistical properties of a clustering, see above). The *Rand index* is a well-known external

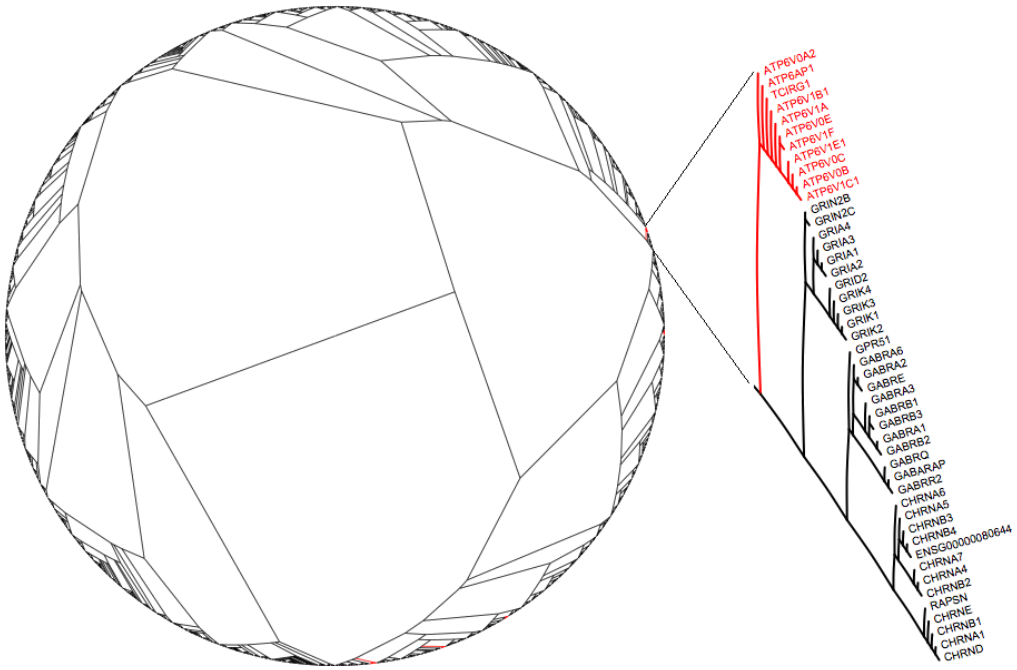


Figure 2.4: Radial visualization of the hierarchical tree based on textual data. The 3,989 genes of the general-purpose gene corpus are linked using Ward's minimum variance method based on the similarity of their textual profiles. The similarity between two profiles was measured via the cosine of the angle between them. Although visualization of the entire tree does not contain useful information, zooming into the tree can reveal more detail about the context of genes each cluster is embedded in. The five clusters with highest Silhouette coefficient are marked in red. The enlarged part of the radial tree contains one of the clusters listed in Table 2.3. The visualization was created with TreeIllustrator [140].

Table 2.3: Gene clusters with highest average Silhouette coefficient based on textual data. The table contains information about the five clusters with the highest average Silhouette coefficient. The clusters contain 11 genes on average. For genes without HUGO gene symbol, the Ensembl identifier is given.

Cluster	SC	Size	Genes
1	0.8252	11	*ATP6AP1*, *ATP6V0A2*, *ATP6V0B*, *ATP6V0C*, *ATP6V0E*, *ATP6V1A*, *ATP6V1B1*, *ATP6V1C1*, *ATP6V1E1*, *ATP6V1F*, *TCIRG1*
2	0.8176	14	*EFNA1*, *EFNA2*, *EFNA3*, *EFNA4*, *EFNB1*, *EFNB2*, *EFNB3*, *EPHA2*, *EPHA3*, *EPHA4*, *EPHA5*, *EPHB1*, *EPHB2*, *EPHB3*
3	0.8009	10	*CCNT1*, *CDK9*, *POLR2B*, *POLR2C*, *POLR2D*, *POLR2E*, *POLR2H*, *POLR2I*, *POLR2J*, *POLR2K*
4	0.5390	10	*PKN2*, *PRKCA*, *PRKCB1*, *PRKCD*, *PRKCE*, *PRKCG*, *PRKCH*, *PRKCI*, *PRKCN*, *PRKCZ*
5	0.4115	10	*ANGPT1*, *ANGPT2*, ENSG00000118257, *FIGF*, *NRP1*, *TEK*, *TIE*, *VEGF*, *VEGFB*, *VEGFC*

index that counts the proportion of cases where two elements are either part of the same cluster, or of different clusters in both the clustering and reference labeling. For both an external partitioning $P = P_1, \ldots, P_u$ and a clustering $C = C_1, \ldots, C_v$ of the same n elements, a matrix M can be compiled consisting of

$$M_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

Defining that N_{11}, N_{01}, N_{10}, and N_{00} count the number of times M_{ij}^P and M_{ij}^C exhibit the element-wise patterns {1,1}, {0,1}, {1,0}, and {0,0}, the Rand index can be written as

$$R = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}}. \tag{2.9}$$

The Rand index can be corrected for random partitioning by normalizing it so that its value equals zero for randomly selected partitions:

$$R_{\text{adj}} = \frac{R - E(R)}{\max(R) - E(R)}. \tag{2.10}$$

The corrected Rand index takes values between 0 and 1. If the correspondence is completely random, $R_{\text{adj}} \to 0$.

An alternative index is called the *Jaccard* coefficient. It measures the correlation between the two binary matrices M_{ij}^P and M_{ij}^C. Using the same notation as above, the *Jaccard coefficient* can be written as:

$$J = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}. \tag{2.11}$$

To compare the results of the linkage analyses based on expression data and textual information, the former was considered the clustering and the latter the reference labeling. For all possible clustering results of both that had an equal number of clusters, the Jaccard and corrected Rand external indices were calculated[1]. Figure 2.5 depicts the variation of these two indices over all possible numbers of clusters.

It is clear that the clustering based on expression data differs a lot from the clustering based on textual data. Even the peaks in Figure 2.5 do not reach higher than 0.02 and 0.04 for the Jaccard and corrected Rand index, respectively. In the following section, the complementarity of both data sources will be analyzed by integrating them in a combined clustering approach.

2.3.4 Discussion

At first sight the analysis of textual data, as with expression data, yields biologically relevant gene groups. However, since the clustering was based on a very limited and highly curated set of documents and gene-document connections, it is not expected to reveal yet unknown gene-gene relations. Its power lies in representing free-text biological knowledge in a computer-amenable way to help with the interpretation of high-throughput experimental data (see Chapter 3). Yet, its complementarity with gene expression data might be useful in combined analyses, as will be shown in the next section.

2.4 Combining expression and textual data

Microarray technology has enabled researchers to look at the expression of thousands of genes simultaneously. In most of the cases they like to know which genes are co-regulated and, hence, might share a functional relationship. This guilt-by-association principle is widely applied in molecular

[1]Only the clustering results with an equal number of clusters were compared because it is computationally intractable to calculate all possible combinations.

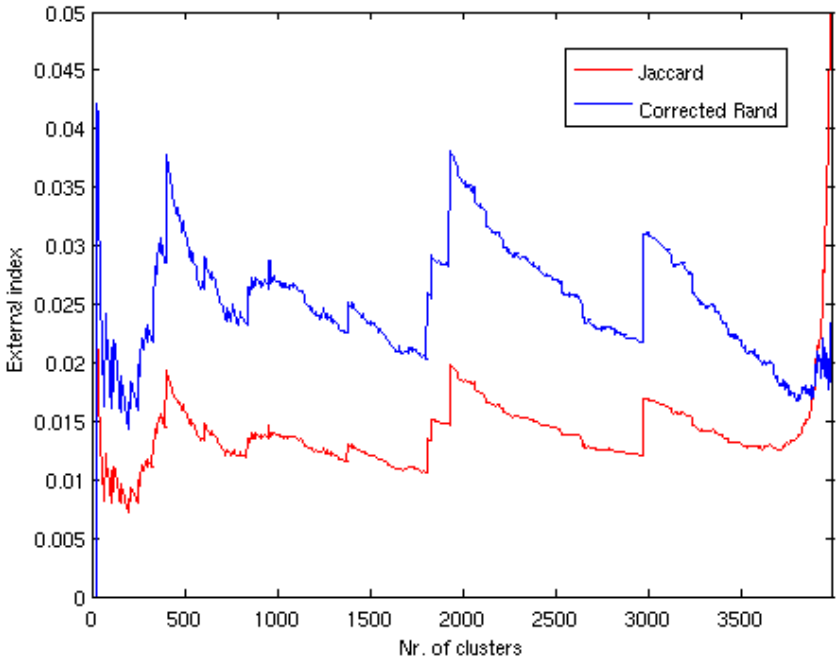


Figure 2.5: Comparison of clustering results based on gene expression versus textual data. Every possible clustering based on gene expression and textual data with an equal number of clusters was compared by calculating the Rand an Jaccard external indices. The plot shows these indices for every number of clusters.

biology research. But recently its applicability in microarray analysis was questioned [104]. Microarray data is generally know to be noisy and the experiments tend to be hard to reproduce. Nevertheless, useful information is captured in expression data that can shed a light on novel pathways. Acknowledging that expression data only contains one aspect of a gene's biological interpretation, the solution to enclose this information might lie in integration of complementary data sources or incorporation of background knowledge in the analysis [98, 108, 125].

Because of the great value of literature as a source of biological information, this section will focus on combining textual and gene expression data to obtain clusters with more functional coherence. Several strategies towards integration of literature and expression data are described in the literature. As already discussed in Section 2.3, Jenssen *et al.* [69] and Alako *et al.* [9] use their literature-derived co-occurrence networks to successfully supervise clustering of gene expression data. Raychaudhuri *et al.* [108] describe an *Optimal Scoring Projection* method that selects and optimizes linear projections in gene expression space. The optimization step is based on calculation of the functional coherence of gene groups according to the literature (see also Chapter 3). The approach of Glenisson *et al.* [57] is detailed below as the method described here is based on it.

2.4.1 Early integration

To combine two sources of data, they have to be brought to a certain level of abstraction on which they can be compared equally (a process also called standardization). In the case of vector-based data sets that can be represented in the same vector space (for instance, two standardized gene expression data sets), combination is as straightforward as concatenating the vector coordinates of both sets.

If the data sets belong to a different vector space, as is the case with gene expression profiles and textual profiles, combination is more complicated. Glenisson *et al.* [57] describe two different approaches. The first approach is based on a linear combination of the distance matrices derived from the data sets. It is equivalent to concatenating the original matrices if for both data sets the same distance measure was used (cosine similarity, for instance). However, because the gene expression and text data span a different vector space, the distribution of the cosine distances for both spaces is completely different. Combination of the distance matrices therefore necessitates use of a rather arbitrary scaling factor.

The second approach gets rid of the differences in distributions of the

distance measures by transforming the distances into p-values. At this level of abstraction, the *distances* can be combined with Fisher's omnibus. The combined statistic has a χ^2 distribution from which a new p-value can be derived for every combination of p-values from the old distance matrices.

The approach presented here does not rely on the original distance matrices, but rather on the results of the linkage analyses. The underlying thought is that linkage analysis removed most of the differences between the original distance matrices to yield hierarchical trees with similar properties. A combined distance matrix is constructed based on a combination of the ultrametric distances between elements in the respective linkage trees. The ultrametric distance between two elements in a binary linkage tree is given by the height of the lowest connection between them. This is clarified in Figure 2.6.

After two new distance matrices are composed based on the distances between the elements in the two linkage trees L_1 and L_2, a combined distance matrix is derived by calculating the average matrix $L_{\text{Comb}} = \frac{(L_1+L_2)}{2}$. This distance matrix is then fed to Ward's linkage algorithm and a statistical analysis of the resulting clusters is performed.

2.4.2 Cluster quality

To select the clusters of highest quality, the same procedure was followed as described in Section 2.2.3 and Section 2.3.2. From all possible clustering results the clusters of size 10 to 20 genes were selected. The clusters with the highest average Silhouette coefficient are listed in Table 2.4. The selected clusters will again be used later on to illustrate some of the developed methods for characterizing gene groups.

2.4.3 Discussion

The Silhouette coefficients of the selected clusters are a little lower than those of the expression and text clusters. Possibly, the reason for this is that the abstraction via the linkage trees removed a part of the discriminative power of the data. Nevertheless, the best clusters are still in the range of the Silhouette coefficient indicating that a reasonable to weak structure has been found (see Table 2.1). As will be shown in the next chapter, this method is able to generate clusters with high biological relevance.

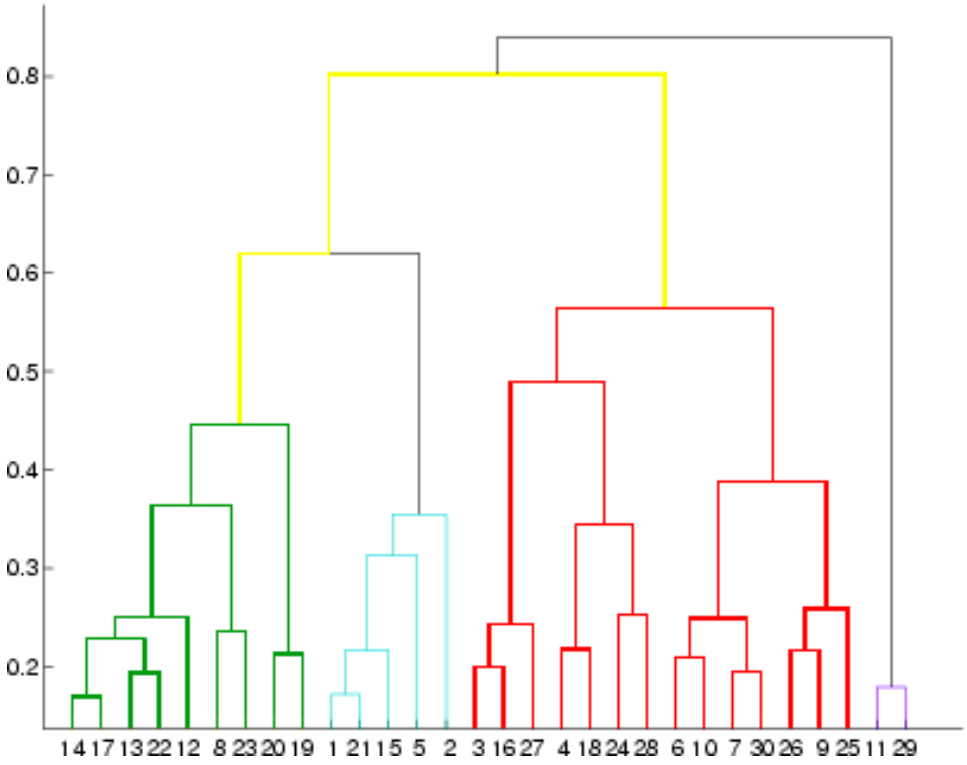


Figure 2.6: Ultrametric distances between elements in a hierarchical tree. The figure shows a dendrogram of a linkage tree. The height of the U-shaped connections between the elements represents the distance between those elements. The distance between the green and the red cluster is defined by the height of the U-shaped connection in yellow and equals 0.8. The distance between all elements of the green cluster and all elements of the red cluster is this distance of 0.8.

Table 2.4: Gene clusters with highest average Silhouette coefficient based on a combination of gene expression and textual data. The table contains information about the five clusters with the highest average Silhouette coefficient. The clusters contain 13.6 genes on average. For genes without HUGO gene symbol, the Ensembl identifier is given.

Cluster	SC	Size	Genes
1	0.5224	13	*C1orf10*, *DSC2*, *DSP*, ENSG00000153802, *EVPL*, *KRT13*, *KRT15*, *PITX1*, *S100A7*, *SCEL*, *SPRR1A*, *SPRR1B*, *SPRR2B*
2	0.5179	10	*ADAM28*, *CD19*, *CD37*, *CD79A*, *CD79B*, *CD83*, ENSG00000012124, *LTB*, *MS4A1*, *PFDN5*
3	0.4970	19	*ACTC*, *CRYAB*, *CSRP3*, *DES*, ENSG00000164082, *FABP3*, *HSPB3*, *ITGA7*, *TGB1BP2*, *ITGB1BP3*, *MYBPC3*, *MYL2*, *NKX2-5*, *NPPA*, *SGCA*, *SGCG*, *TNNC1*, *TNNI3*, *TNNT2*
4	0.4955	10	*ATPIF1*, *BCCIP*, *DDX11*, ENSG00000076043, ENSG00000111247, ENSG00000136824, *KNSL7*, *NOLC1*, *SMC4L1*, *TXNL1*
5	0.4953	16	*AKR7A3*, *ALDH8A1*, *BDH*, *ECH1*, *ECHS1*, ENSG00000176046, ENSG00000183048, *FTHFD*, *FXYD1*, *IFITM3*, *KCNA10*, *PEMT*, *PGM1*, *PRODH2*, *RARRES2*, *SPR*

2.5 Conclusion

In this chapter, three different approaches for clustering genes were described. A cluster analysis is always performed in an *ad hoc* way, in that it is important to *play* with the data to get certain insights. Depending on the purpose of the cluster analysis different approaches are more or less suitable.

In the cases described above, the purpose of the analyses was to determine clusters of genes of a controllable size, starting from two completely different data sources (and the combination of both). Because the size is fixed, clusters can be retrieved from the linkage trees without having to search for an optimal number of clusters or level at which to cut the tree. A statistical measure of how well the determined clusters fitted the data was used to find clusters of high (statistical) quality. Because all possible clusterings are taken into account, it might be interesting to investigate the variation of this Silhouette coefficient for one cluster. It is expected that a small variation throughout all clusterings could be correlated with the stability, and hence the quality, of a cluster.

One of the challenges is now to proof that this statistical measure is biologically relevant. Clusters can be selected based on their statistical quality, but in the end it is the biological quality that counts. Therefore efficient methods are needed to investigate the biological validity of a group of genes. This is the focus of the next chapter.

Chapter 3

Gene group validation

VALIDATION of a group of genes consists of investigating why these genes are grouped together from a biological point of view. Referring to the knowledge acquisition cycle, this investigation is the step in which the experiment's data analysis results are tested against and put into context with the established biological knowledge (Figure 3.1).

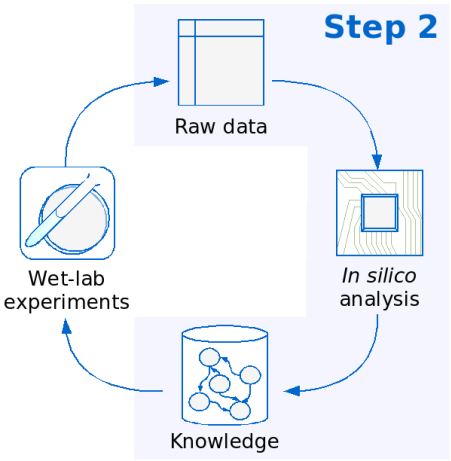


Figure 3.1: Step 2 in the knowledge acquisition cycle. The second step comprises removal of artifacts coming from noise in the original data, as well as interpretation of analysis results in the context of biological knowledge.

As described in the previous chapter, there are a lot of different ways to yield gene groups. Every method has its advantages and drawbacks, but they all share that at some level in the clustering analysis, noise in the

original data causes misclassification of genes. Hence, validation has two major goals:

- Unravel the gene group's underlying biological program.
- Define which genes do *not* belong to the gene group at hand.

Especially in the context of high-throughput experimentation there is a need for efficient methods to define the characteristics and biological quality of large sets of gene groups. The methods described in this chapter can be regarded as examples of intermediate integration (see Figure 1.4). In Section 3.1, functional annotations of genes are used to characterize gene groups and define their coherence. The biological information captured in free text documents like scientific papers can also be used for the purpose of validating gene groups. Subclustering of textual profiles can be used to spot outliers. This is elaborated in Section 3.2. The gene groups defined in the previous chapter are used to exemplify the different methods.

3.1 Gene Ontology to characterize gene groups

The Gene Ontology (GO) is a set of structured vocabularies that covers the domain of molecular and cellular biology [132, 30]. The three main and orthogonally defined ontologies are *Molecular Function* (MF), *Biological Process* (BP), and *Cellular Component* (CC). Each vocabulary consists of a tree of free text definitions (terms) connected through *is-a* or *part-of* relations. Since these terms are widely used in the annotation of genes, gene products, and sequences, they are an ideal and very specific information source for the characterization of groups of genes. Figure 3.2 visualizes a part of the *Molecular Function* tree.

The characterization of a group of genes usually involves answering two basic questions:

- What are the most prominent characteristics of the group?
- How coherent are these characteristics?

In the context of GO annotations the answer to the first question is the set of GO terms that best describes the properties common to all or most members of the gene group. The solution described in Section 3.1.1 uses a probability distribution to obtain the set of statistically over-represented GO terms for a group of genes. The second question can only be answered



Figure 3.2: Part of the *Molecular Function* tree of the Gene Ontology. The Gene Ontology (GO) consists of three orthogonal branches called *Molecular Function* (MF), *Biological Process* (BP), and *Cellular Component* (CC). The root part of the MF ontology is shown here. The ontology is hierarchically structured and its objects are connected via *is-a* or *part-of* relations. Each object has a stable and unique identifier and a free text description. Screenshot taken from the AmiGO browser [133].

by defining a distance measure between two GO terms. In Section 3.1.2, the distances between the terms in the GO trees are used to address the problem of establishing the functional coherence of a gene group.

The two methods described here use the information from the Gene Ontology to a different extent. The first method mostly uses the semantic information of the GO terms and the semantic links between them. The semantic information content of a term is determined by looking at the way it is used to annotate genes: how often does the annotation occur and together with what other terms? The second method obliterates all semantic information and only regards the structure of the GO trees. It is assumed that terms close to each other in a tree are functionally more related than distant terms.

3.1.1 Statistically over-represented GO terms

In the literature, several approaches to characterize groups of genes with Gene Ontology annotations are described. *FatiGO* [8] is probably the best known tool to find statistically significant GO terms for a group of genes. It uses Fisher's exact test to calculate which GO terms are over- and under-represented in a set of genes, as compared to a reference set. The resulting p-values are corrected for multiple testing in three different ways. Due to the used statistics, the analysis is restricted to a certain level of the Gene Ontology tree. Obviously this is a major drawback because an extra parameter has to be set in a rather arbitrary way. Analysis of lower levels will yield more specific, but much fewer terms. *GoMiner* [154] also uses Fisher's exact test but performs a multiple testing correction based on resampling. It also provides graphical representations of the GO tree to visualize the over-represented GO categories. *GOstat* [14] is yet another tool with a similar statistical basis as *FatiGO* and *GoMiner*. A χ^2 test is used to calculate a p-value that represents the probability to obtain the observed frequency of a GO term, given its frequency in a reference group. Again, the p-values are corrected for multiple testing. *BiNGO*, the Biological Networks Gene Ontology tool [82, 17], has the advantage to be implemented as a plugin for *Cytoscape* [116, 31]. Cytoscape is an open source software project oriented at the integrated visualization of large biomolecular interaction networks with high-throughput data from gene expression experiments or other phenotypical studies. BiNGO provides both hypergeometric and binomial statistical distributions and a variety of multiple testing corrections and its output can readily be mapped on top of the Cytoscape visualizations. Other applications to characterize gene groups based on GO annotations include *Ontolo-*

gizer [111], *Onto-Express* [37], *MAPPFinder* [36], and *GoSurfer* [155].

As said before, the method presented here takes a straightforward binomial approach with addition of a multiple testing correction to identify the most over-represented GO annotations in the complete Gene Ontology tree. The method was developed by Stein Aerts. Unique about this method is that it is available as a web service as part of the INCLUSive suite, a set of algorithms for the analysis of gene expression data and the discovery of *cis*-regulatory sequence elements. It was published as such by the author [28]. More information about web services and INCLUSive can be found in Chapter 5.

Statistical framework

Given a group of genes, all annotated with certain Gene Ontology terms, what terms give most information about the group? The solution described here uses a binomial probability distribution function to calculate a p-value for every term. This p-value allows ranking of all terms annotated to a group of genes and selection of the most significant ones, for instance by applying a significance level of $p < 0.05$.

First of all, the GO annotations of a large set of genes (the *corpus*; all known genes of a complete genome, for instance) are gathered. Because all terms in the three GO trees are connected through *is-a* or *part-of* relations, a more specific term automatically subsumes all its more general ancestor terms. If the term *transcriptional repressor activity* (see Figure 3.2) is annotated to a gene, for instance, the terms *transcription regulator activity* and *molecular function* are implicitly also annotated to this gene. Therefore, for every GO term, all ancestors up to but excluding the root of the tree are added to the annotations of a gene to obtain an *extended annotation*. Dividing the number of times a term is annotated to a gene by the total number of genes in the corpus results in expected frequencies for every term.

A binomial distribution function describes the probability to get n successes out of N *Bernoulli trials* with a success probability of $p = 1 - q$:

$$P_p(n|N) = \binom{N}{n} p^n q^{N-n}. \tag{3.1}$$

In this context, the Bernoulli trials are the selection of N genes from the corpus (i.e., the genes from the group under investigation), with a probability of p to select a gene with a certain GO term. The probability p equals the

expected frequency of the term, calculated as described above. The realization n is the number of times the term is present in the extended annotations of the genes in the group. The binomial distribution gives the probability of observing the GO term as often as seen in the extended annotations of the genes at hand. The p-value for this term expresses the probability to get even more occurrences by chance alone and is given by

$$p = 1 - \sum_{i=1}^{n} P_p(n|N). \tag{3.2}$$

For the binomial distribution to be statistically sound, the performed Bernoulli trials have to be independent. This is not the case because a gene will never occur twice in the same gene group. In other words, the gene is not replaced in the corpus. This kind of distribution *without replacement* can best be described with a hypergeometric distribution. However, since the number of genes in the group under investigation is much smaller than the number of genes in the corpus, the law of large numbers applies and the simpler binomial distribution can be used to approximate the hypergeometric distribution.

To control the number of false positives in the list of *significant* GO terms (the once above the cutoff), the obtained p-values need to be corrected for multiple testing. This is done with a Bonferroni step-down correction (Holm's correction). This correction is very similar to a Bonferroni correction, but slightly less stringent. The correction is performed by multiplying each p-value with the number of terms that have a higher p-value.

Examples

Tables 3.1, 3.2, and 3.3 show the top-15 statistically over-represented GO terms of the gene groups from the general-purpose gene corpus with the highest Silhouette coefficient after clustering based on gene expression and textual data and the combination of both, as described in the previous chapter. The annotations clearly identify the groups as being involved in epithelial and skin development for the expression and combined clusters, and in cell metabolism for the text cluster. The top-15 statistically over-represented GO terms of the other gene groups listed in Chapter 2 can be found in Appendix B.

Table 3.1: Statistically over-represented GO annotations of the gene group with the highest Silhouette coefficient after clustering based on expression data. Lower p-values indicate that an annotation is more over-represented and, thus, more characteristic for the gene group.

Attribute	**Description**	***p*-value**
GO:0008544	epidermal differentiation	1.53E-14
GO:0007398	ectoderm development	3.02E-14
GO:0009888	histogenesis	3.87E-13
GO:0009887	organogenesis	1.26E-06
GO:0009653	morphogenesis	5.57E-06
GO:0009506	plasmodesma	5.83E-06
GO:0015696	ammonium transport	2.30E-05
GO:0005198	structural molecule activity	3.13E-05
GO:0009408	response to heat	5.00E-05
GO:0008519	ammonium transporter activity	4.92E-05
GO:0001533	cornified envelope	4.84E-05
GO:0007275	development	1.08E-04
GO:0030506	ankyrin binding	1.30E-04
GO:0015695	organic cation transport	1.83E-04
GO:0030216	keratinocyte differentiation	2.45E-04

Table 3.2: Statistically over-represented GO annotations of the gene group with the highest Silhouette coefficient after clustering based on textual data. Lower p-values indicate that an annotation is more over-represented and, thus, more characteristic for the gene group. Note that in this case all calculated p-values are below machine precision (i.e., lower than 1.00E-16).

Attribute	**Description**	***p*-value**
GO:0016020	membrane	0.00E+00
GO:0042625	ATPase activity, coupled to transmembrane movement of ions	0.00E+00
GO:0006752	group transfer coenzyme metabolism	0.00E+00
GO:0009117	nucleotide metabolism	0.00E+00
GO:0015077	monovalent inorganic cation transporter activity	0.00E+00
GO:0005623	cell	0.00E+00
GO:0008151	cell growth and/or maintenance	0.00E+00
GO:0015078	hydrogen ion transporter activity	0.00E+00
GO:0046933	hydrogen-transporting ATP synthase activity, rotational mechanism	0.00E+00
GO:0006753	nucleoside phosphate metabolism	0.00E+00
GO:0009108	coenzyme biosynthesis	0.00E+00
GO:0005386	carrier activity	0.00E+00
GO:0009145	purine nucleoside triphosphate biosynthesis	0.00E+00
GO:0009206	purine ribonucleoside triphosphate biosynthesis	0.00E+00
GO:0015075	ion transporter activity	0.00E+00

Table 3.3: Statistically over-represented GO annotations of the gene group with the highest Silhouette coefficient after clustering based on the combination of expression and textual data. Lower p-values indicate that an annotation is more over-represented and, thus, more characteristic for the gene group.

Attribute	**Description**	***p*-value**
GO:0008544	epidermal differentiation	0.00E+00
GO:0007398	ectoderm development	0.00E+00
GO:0009888	histogenesis	2.02E-14
GO:0009887	organogenesis	2.38E-08
GO:0009653	morphogenesis	1.67E-07
GO:0005200	structural constituent of cytoskeleton	9.68E-07
GO:0005882	intermediate filament	3.29E-06
GO:0045111	intermediate filament cytoskeleton	3.38E-06
GO:0005198	structural molecule activity	4.45E-06
GO:0005856	cytoskeleton	5.94E-06
GO:0007275	development	6.49E-06
GO:0009506	plasmodesma	2.48E-05
GO:0016327	apicolateral plasma membrane	8.00E-05
GO:0005911	intercellular junction	2.93E-04
GO:0030054	cell junction	5.71E-04

Discussion

It is clear that the automated characterization of groups of genes can speed up biological investigation enormously, especially when dealing with large gene groups or huge amounts of data. Nevertheless, these kinds of methods obfuscate the contribution of individual genes to the resulting summary. Therefore it is of utmost importance to investigate groups in more detail, for instance through coherence studies or subclustering (see Section 3.2).

3.1.2 Distances between GO terms

Coherence is all about the variation of distances between elements in a certain space. The space of concern here is the one describing gene function. To define a measure of coherence, a distance between two genes needs to be defined in this functional space.

Lord *et al.* [81, 80] describe a statistical method to calculate a semantic similarity measure between two GO terms in the Gene Ontology. Their measure is based on the premise that a term annotated to a lot of genes or proteins is less informative than a term annotated to few. Starting from a corpus of biological entities annotated with GO terms, they calculate the information content of every term. A similarity measure between two terms is then calculated based on the information content of the parents they share. These distances are then used to derive a semantic similarity measure between two genes or proteins.

The approach described in this section takes advantage of the fact that the Gene Ontology is a structured vocabulary. Distances are derived directly from the Directed Acyclic Graph (DAG) representation of the GO using a shortest path algorithm. This is a novel approach for assessment of the biological coherence of a gene group. Although nice results are obtained, the method still has to be validated in a larger biological context, and benchmarked against similar approaches.

Dijkstra's shortest path algorithm

Dijkstra's Shortest Path Algorithm [34] is one of the best-known algorithms to solve the *single-source* shortest paths problem (i.e., to define the shortest path between a source node and all possible destination nodes of a directed and weighted graph). The algorithm starts with all graph nodes being *unsettled*. Once the shortest path from a certain node to the source is defined, it is moved to the set of *settled* nodes. This process goes on until all nodes are settled. Table 3.4 shows the algorithm's pseudo-code.

Table 3.4: Pseudo-code of Dijkstra's Shortest Path Algorithm. d stores the best estimate of the shortest distance from the source to each vertex. π stores the predecessor of each vertex on the shortest path from the source. S contains all settled vertices. Q contains all unsettled vertices. Pseudo-code was taken from Renaud Waldura [146].

```
// initialize d to infinity, π and Q to empty
d = ∞
π = ()
S = Q = ()

add s to Q
d(s) = 0

while Q is not empty {
        u = extract-minimum(Q)
        add u to S
        relax-neighbours(u)
}

extract-minimum(Q) {
        find the smallest vertex in Q (as defined by d)
        remove it from Q
        return it
}

relax-neighbours(u) {
        for each vertex v adjacent to u, v not in S {
                if d(v) > d(u) + [u, v] // a shorter distance exists {
                        d(v) = d(u) + [u, v]
                        π(v) = u
                        add v to Q
                }
        }
}
```

Gene group coherence

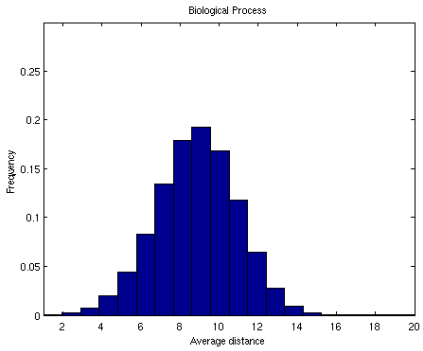
The distances between all terms contained by the three main branches of the Gene Ontology can be calculated with Dijkstra's Shortest Path Algorithm described above. To do this, the three structured vocabularies (MF, BP, and CC) are fed to the algorithm as directed graphs. For every term in the vocabularies, the shortest distance to all other terms in the same vocabulary is defined, not taking into account the direction of the relations between the terms. This means that for every directed edge between a parent and a child term, an inverse edge from child to parent is added to the graph representation. This is necessary because otherwise only the distances between every node and its descendants would be calculated and, for instance, not the distances between two different nodes located at the same depth. Figure 3.3 shows the distribution of these distances for MF, BP, and CC. Table 3.5 gives some statistics on the three GO branches.

Table 3.5: Statistics of the GO subtrees. The table lists the number of terms, the mean, median, and maximum distance (d) between two nodes, and the maximum depth from the root node of the subtrees Molecular Function (MF), Biological Process (BP), and Cellular Component (CC).

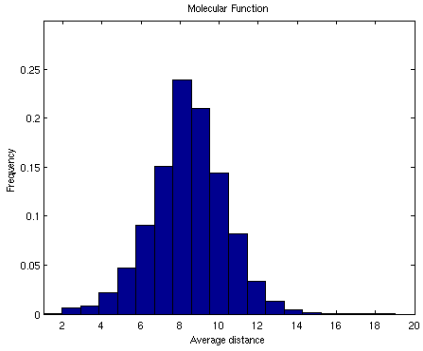
Tree	**Nr of terms**	**Mean** d	**Median** d	**Max** d	**Max depth**
MF	6,915	8.3240	8	19	11
BP	8,730	8.7480	9	20	13
CC	1,375	5.2721	5	14	9

Once all distances between all nodes in the trees are available, the coherence of an arbitrary group of GO annotations can be estimated by calculating the average distance between them and comparing this distance with the average distance between all nodes of the complete tree. This measure can readily be used to estimate the coherence of the annotations of a group of genes with respect to molecular functions, biological processes, and cellular components.

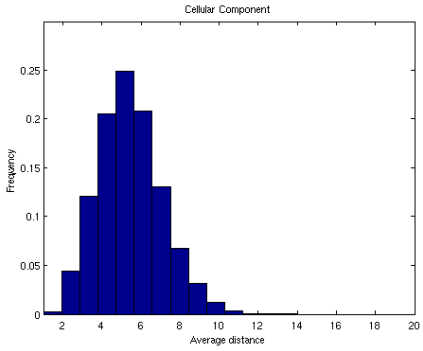
To obtain a measure independent of the size of the gene group and the number of associated GO annotations, a distribution of average distances is sampled for the group size and GO subtree at hand. This is done by random selection of an appropriate number of genes from the gene corpus at hand (from the 3,989 genes of the general-purpose data set, for instance). The GO annotations that are annotated to this randomly composed gene group, and



(a)



(b)



(c)

Figure 3.3: Distribution of distances in the Gene Ontology subtrees. The histograms show the distribution of the distances between all nodes in the subtrees Biological Process (a), Molecular Function (b), and Cellular Component (c).

that belong to the subtree of interest, are retrieved and the average distance between them is defined. This is done 1,000 times and the 1,000 resulting average distances are used to fit a normal distribution[1]. The distribution can then be used to estimate the probability of obtaining an average distance as low as seen for a certain gene group. The lower this p-value, the more coherent the gene group's GO annotations. Figure 3.4 gives an example of the distributions of the average distances between the annotations of a group of 20 genes randomly sampled from the general-purpose gene corpus.

As can be seen in Figure 3.5, the standard deviation for the Molecular Function subtree is larger than those for the Biological Process and Cellular Component subtrees. This result is consistent over all possible sizes of gene groups. It is probably due to the fact that most genes have more than one molecular function, while they are active in only one biological process or part of only one cellular component. The probability of finding a coherent gene group decreases with the size of the group, hence smaller standard deviations are found for larger cluster sizes. Yet, if the standard deviation is smaller, the coherence prediction will be more precise.

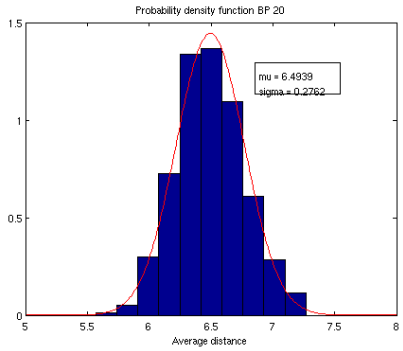
Examples

The described approach is exemplified with the 15 gene clusters that were selected after clustering based on gene expression data, textual data, and the combination of both in Chapter 2. All GO terms annotated to the genes in these clusters were retrieved and the average distance between them was calculated separately for each of the three GO branches. Based on the sampled distributions, a p-value was calculated for every gene group and every GO branch. The p-values are listed in Table 3.6.

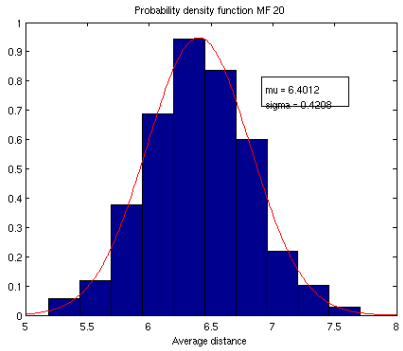
As expected, the clusters extracted from the textual data are (in general) more coherent than those extracted from the expression data. The clusters resulting from the combined approach, however, are not significantly more coherent.

At this stage it is interesting to know if there exists a correlation between the statistical measure of cluster quality (i.e., the Silhouette coefficient presented in the previous chapter) and this biological measure of cluster coherence. Table 3.7 lists the correlations between the Silhouette coefficients and the coherence measure described above. Only the clusters that contained 10 to 20 genes were included in the analysis.

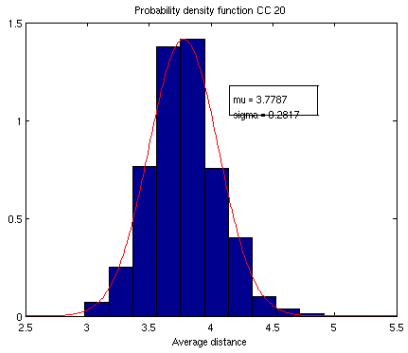
[1]Note that the number of annotations retrieved this way will vary between different randomly composed gene groups. However, because the genes are picked randomly, the number of annotations is expected to be more or less similar over all 1,000 samples.



(a)



(b)



(c)

Figure 3.4: Sampled distributions of average distances. The figures show the probability density plots (PDF) of the average distances for the Gene Ontology subtrees Biological Process (a), Molecular Function (b), and Cellular Component (c). The PDFs were fitted based on the average distances of the annotations of 1000 groups of 20 genes randomly selected from the general-purpose gene corpus.

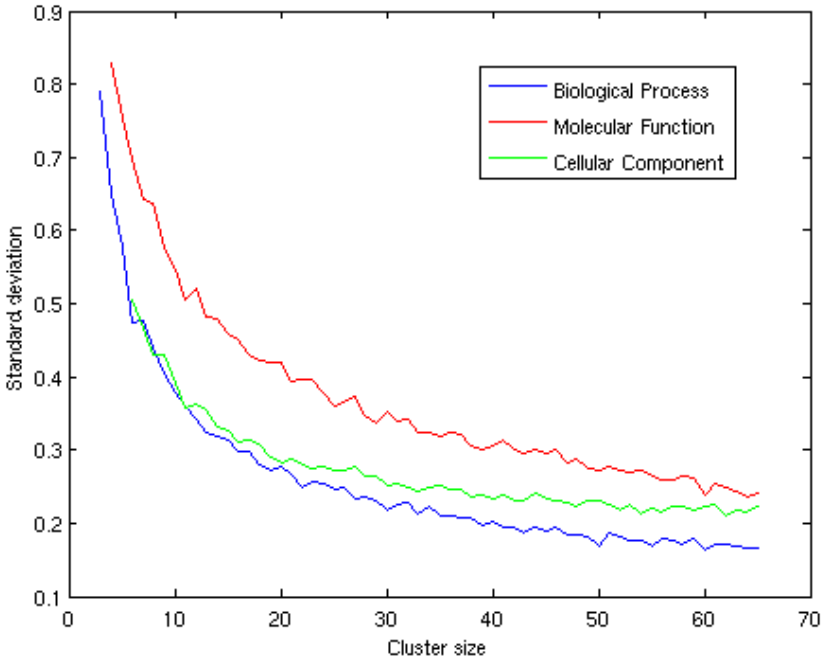


Figure 3.5: Standard deviations of the distributions of the average distance between the GO annotations of randomly composed groups of genes. First, a group of genes is composed by random selection of a number of genes from a gene corpus. Then, all GO annotations of these genes are fetched and for each GO subtree (Biological Process, Molecular Function, and Cellular Component) the average distance between these annotations in the trees is recorded. This is done 1,000 times and for every subtree a normal distribution is fitted through the resulting average distances. The plot shows how the standard deviations of these distributions vary across different cluster sizes. Predictions of coherence will be more precise for larger gene groups, but the chance of finding a coherent group will decrease.

Table 3.6: Results of average distance p-value analysis. For the five clusters with the highest *Silhouette* coefficients after clustering based on expression data (Expr), textual data (Text), and the combination of both (Comb), all annotations of the *Biological Process* (BP), *Molecular Function* (MF), and *Cellular Component* (CC) subtrees of the Gene Ontology were retrieved and the average distance between them was calculated. The listed p-values represent the probability to find a cluster of genes with a certain average distance between its annotations that is lower than the expected average distance for the cluster size at hand. All p-values lower than 0.01 are marked in bold.

Cluster	**BP**	**MF**	**CC**
Expr 1	**0.0060**	**0.0014**	0.9870
Expr 2	**0.0040**	**0.0001**	0.7431
Expr 3	**0.0007**	0.1192	0.2700
Expr 4	0.0314	0.6352	0.5740
Expr 5	0.0135	0.0405	0.6296
Text 1	**0.0003**	**0.0004**	**0.0000**
Text 2	**0.0000**	0.0640	0.0113
Text 3	**0.0000**	0.0575	**0.0001**
Text 4	0.4839	0.0138	**0.0038**
Text 5	**0.0065**	0.0184	0.1627
Comb 1	**0.0054**	**0.0047**	0.9019
Comb 2	**0.0000**	**0.0083**	0.0499
Comb 3	0.0712	**0.0029**	0.3802
Comb 4	**0.0098**	0.6941	0.2814
Comb 5	0.1678	0.5212	0.0505

Table 3.7: Correlation between Silhouette coefficient and gene group coherence. For all clusters with a number of genes between (and including) 10 and 20, the Silhouette coefficients were recorded. Then, the correlation was measured between those and the gene group coherence measures, calculated for each of the clusters as described in this section. The correlation gives an indication of the correspondence between a statistical measure of cluster quality (the Silhouette coefficient, derived from the underlying data source: gene expression data (Expr), textual data (Text), or the combination of both (Comb)) and a measure of functional coherence of a gene group (based on the genes' GO annotations from the Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) subtrees). The correlations are negative: clusters with a high Silhouette coefficient show more functional coherence (i.e., have lower gene group coherence p-values) than clusters of low statistical quality. The positive correlation between *Expr* and *CC* points out the fact that expression data is unable to capture information about the constituents of cellular components.

Clusters	**BP**	**MF**	**CC**
Expr	-0.2101	-0.2397	0.1245
Text	-0.1658	-0.2221	-0.1925
Comb	-0.1455	-0.1064	-0.1109

Although the correlations are small, some conclusions can be drawn. In most of the cases (and as expected), a negative correlation is seen. This means that clusters with a high Silhouette coefficient tend to have low p-values. In other words: they have the tendency to be more coherent from a biological point of view. The correlation between the Silhouette coefficients of the expression clusters and the p-values for the Cellular Component subtree is an exception in that it shows a positive correlation. Apparently, gene expression data is unable to capture information about the constituents of cellular components.

Another observation is that the gene groups from the clustering based on textual data are slightly less correlated with the Silhouette coefficient than those of the gene expression clustering (although the five selected clusters with the highest Silhouette coefficient are clearly more coherent in the case of textual data than in the case of expression data; see Table 3.6). As said before, almost no papers contain information about gene groups with more than 10 genes. Since the correlation study was based on gene groups of larger sizes (between 10 and 20), this might be the reason for the smaller correlations.

The correlations of the gene groups based on the combination of expression and text data are also smaller than the other correlations. Nevertheless, the clusters with the five highest Silhouette coefficients, selected in the previous chapter, display good coherence measures. As discussed previously (see Chapter 2.4), the combination method probably removed part of the discriminative power of the underlying data sources, resulting in fewer clusters with good Silhouette coefficients.

Discussion

Gene Ontology annotations are regarded as a rich and confident source of functional information about genes. However, they do not capture all functional intricacies about genes. The biomedical literature constitutes a much more comprehensive and up-to-date source of gene functional information. Acknowledging this, Raychaudhuri *et al.* [107, 106] developed a method to establish the functional coherence of a gene group based on the analysis of publications associated with the individual genes. Their *neighbor divergence per gene* (NDPG) method is based on counting the number of documents that are shared by more than one gene in the group. NDPG is successfully applied to both gold-standard and expression data, but has the slight drawback that it does not give information on the actual function.

There is a clear advantage in complementing conventional data mining statistics with functional data of genes to establish a gene group's coherence. Nevertheless, no exhaustive benchmark study was conducted and published to date to prove this.

3.2 Textual profiling of gene groups

As discussed before, a large amount of biological information resides in free-text format (such as textual annotations/descriptions, scientific abstracts, and full papers). New findings are accumulating and the number of publications to come out increases every year[2]. But this information is largely underutilized by researchers because of its highly unstructured format. To address this problem, a multitude of text-mining systems for biomedical research were developed (for an overview, see Chapter 1).

[2]The number of articles in MEDLINE published in 1980 is 272,826. In 2000, the number of new MEDLINE entries summed up to 520,189, a factor two increase in twenty years time. This number is steadily rising as in 2005 MEDLINE counted already 659,165 new entries.

Some of these text-mining approaches have been developed to assign biological meaning to a group of genes by summarizing and profiling the available literature, especially in the context of gene expression analysis. Masys *et al.* [86] link groups of genes with relevant MEDLINE abstracts through the PubMed engine. Each cluster is characterized by a pool of keywords derived from both the Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) ontology. The method described by Calogero *et al.* [23] relies on a gene name extractor to link genes and MEDLINE documents. From the titles and abstracts of these documents, keywords are derived using a part-of-speech (POS) tagger. The keywords are then used to characterize and cluster the set of documents associated to a set of genes. Kankar *et al.* [72] rely on the PubMed interface to retrieve all documents associated to the individual genes in a group. In the next step, the genes are represented by the MeSH terms associated with these documents, ordered according to their frequency count. Finally, an overall gene group relevance rank is calculated per MeSH term based on several statistical attributes. Chaussabel and Cher [24] also retrieve documents related to a gene via the PubMed interface, but only keep those that contain the gene's symbol in title or abstract. For each encountered term they define the *baseline occurrence* (i.e., the average frequency of a term in document collections associated with randomly selected genes). The difference between a term's *baseline occurrence* and occurrence in a subset of documents (related to a group of genes, for instance) is used to estimate a term's relevance.

Ongoing curated annotation efforts facilitate use of literature data to characterize biological entities or phenomena (genes, proteins, but also diseases, patients, and so on). The assumption is that the textual information describing a certain gene, for instance, can be used to represent this gene. The vector space model provides an ideal computational framework to work with free-text information and allows combination of all documents annotated to a gene to create a true *gene index* (see Chapter 1). The gene indices that are used in the method described below, were created using gene-literature associations present in the Entrez Gene database. Other examples of more specialized curated repositories that contain gene-literature mappings are the *Saccharomyces Genome Database* (SGD) [26, 130], *The Arabidopsis Information Resource* (TAIR) [110, 128], and so on.

The methodology and software described below were developed in close collaboration with Steven Van Vooren and Patrick Glenisson [53]. This involved developing the ideas, designing the domain vocabularies, and implementing a web-based application. Subclustering of textual profiles was added by the author. The novelty of the described work lies in the creative

use of domain vocabularies, the user-friendly and flexible implementation that allows subclustering and iterative investigation, and the link-out feature that can be used to generate queries to a myriad of biological resources.

3.2.1 Profiling gene groups with text-based information

In the same straightforward way as document indices are combined into gene indices, indices of groups of genes can be constructed by combining all gene indices of its constituent genes. This feature can be used to characterize groups of genes by creating group profiles and visualizing the most important terms and phrases.

As already mentioned, this idea was implemented in a web-based application called *TXTGate* [56]. TXTGate allows profiling of groups of genetic loci from all species present in Entrez Gene. On top of that, TXTGate also contains indices for *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. For most of the loci, indices created with different domain vocabularies are available (see Chapter 1). The terms can directly be used to query the databases from the National Center for Biotechnology Information (NCBI) (the *Entrez PubMed* database, for instance), the Gene Ontology controlled vocabularies, the GeneCards database (for information on human genes) [109], and SGD (for information on yeast genes). Four domain vocabularies (*GO*, *OMIM*, *MeSH*, and *eVOC*) are term-centric in that they profile genes with the most characteristic terms encountered in the documents describing them. Beside, a gene-centered analysis of human genes is made possible through the availability of the *HUGO* index. This index was created with a domain vocabulary comprising all human gene symbols of the HUGO Gene Nomenclature Committee (HGNC) [145]. The use of this index will be elaborated in the next chapter.

Gene information

The gene specific information to create gene indices was retrieved from the Entrez Gene database [83]. Entrez Gene provides a gene-based view of the information resulting from the sequencing and annotation of key genomes. It contains map, sequence, expression, structure, function, citation, and homology data. Every gene present in the database is annotated with zero or more publications relevant to it. Table 3.8 gives some statistics of the literature references present in Entrez Gene as of November 2004.

All abstracts of the publications linked to the genes in Entrez Gene were indexed with the four domain vocabularies described in Chapter 1.

Table 3.8: Entrez Gene database statistics. The table gives some numerical details with respect to the documents associated with the loci present in the Entrez Gene database.

	Mean	**Median**	**Min**	**Max**
Nr. of refs per gene	1.9401	1	1	2169
Nr. of genes per ref	9.3544	1	1	40816
Total nr. of genes	746559			
Total nr. of species	2533			
Total nr. of references	154833			

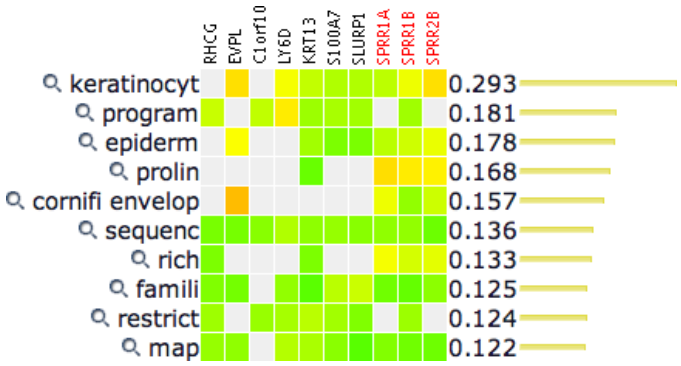
In the next step, the annotation information from Entrez Gene was used to combine the indices of all abstracts and create four gene indices, one for each domain vocabulary. The indexing was performed with the *Apache Lucene* package, an open-source high-performance text search engine library [131]. Customized package extensions were written to allow domain vocabulary sensitive indexing.

Examples

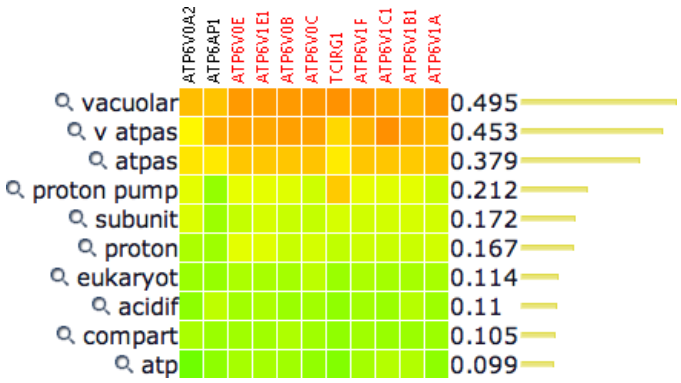
As an example of the typical output TXTGate generates, the three gene groups with the highest Silhouette coefficients from the previous chapter are profiled using the *GO* vocabulary. The terms in the profiles are sorted by their IDF weights, placing the terms best characterizing the gene group high in the visualization. The result is shown in Figure 3.6. There is a clear correspondence between the textual profiles and the statistically over-represented GO terms of the three groups (see Tables 3.1, 3.2, and 3.3). The reason for the homogeneity of the profile of *Text cluster 1* is that the genes in this group were clustered using the same textual profiles the group profile is generated from.

TXTGate can profile the textual information of a group of genes through the eyes of different vocabularies. This is especially useful when only a certain aspect of a gene group is being studied; its involvement in a disease, for instance. To exemplify this, a gene group from the combined clustering approach (see Chapter 2) was profiled with the *GO*, *OMIM*, *MeSH*, and *eVOC* domain vocabularies. Figure 3.7 shows the resulting profiles.

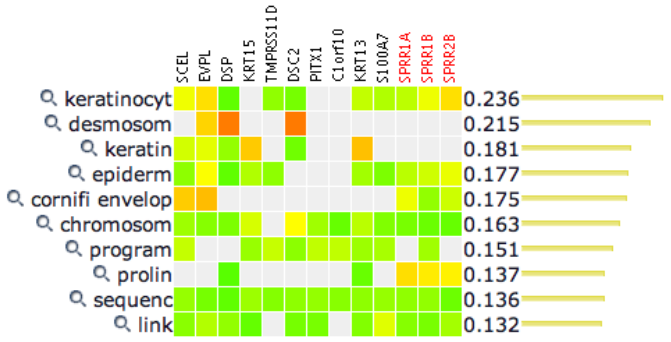
While the *GO* profile highlights the molecular and cellular functions of the gene group, the focus of the *OMIM* and *MeSH* profiles is more disease-



(a) Expression cluster 1 with *GO* vocabulary.

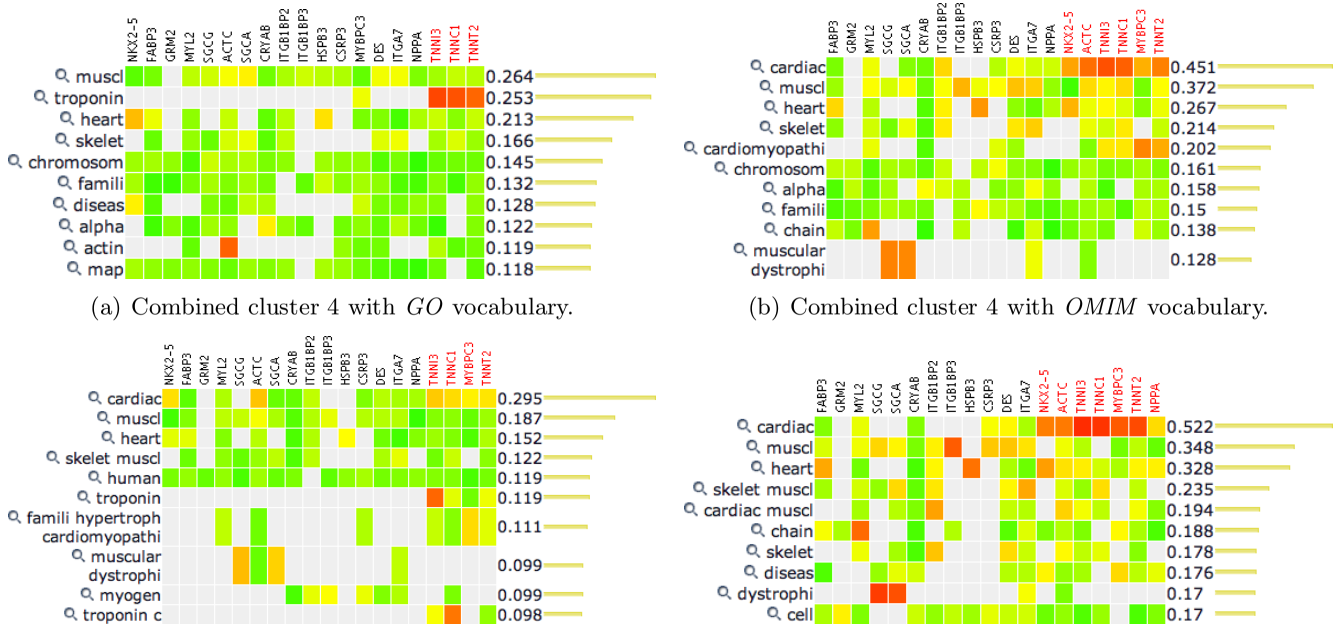


(b) Text cluster 1 with *GO* vocabulary.



(c) Combined cluster 1 with *GO* vocabulary.

Figure 3.6: Textual profiles of gene groups obtained after clustering based on expression and textual data, and the combination of both. The web-based application TXTGate generated the textual profiles of the three clusters with the highest Silhouette coefficient using the *GO* domain vocabulary.



(a) Combined cluster 4 with *GO* vocabulary.

(b) Combined cluster 4 with *OMIM* vocabulary.

(c) Combined cluster 4 with *MeSH* vocabulary.

(d) Combined cluster 4 with *eVOC* vocabulary.

Figure 3.7: Textual profiles of one of the gene groups with the highest Silhouette coefficients after clustering based on the combination of expression and text data. The group was profiled with different domain vocabularies to stress different aspects of the data.

related with terms like *cardiac cardiomyopathi* and *muscular dystrophi*. The *eVOC* vocabulary stresses cell and tissue types with terms like *skelet muscl* and *cardiac muscl*.

Discussion

Both the over-represented GO annotations as the textual profiles provide a means to characterize the biological features of a gene group. While GO annotations are less prone to noise, the textual profiles are much richer in content and give direct access to publications discussing the topic at hand. On top of that, the contribution of each of the genes can easily be visualized. Working with domain vocabularies also has a clear advantage by providing different views of the same information. It might be interesting to weigh and index individual sentences of the abstracts related to a group of genes (for instance based on the IDF weights of its constituent terms) and use those to create a sentence profile for the group. This profile is expected to be even more descriptive.

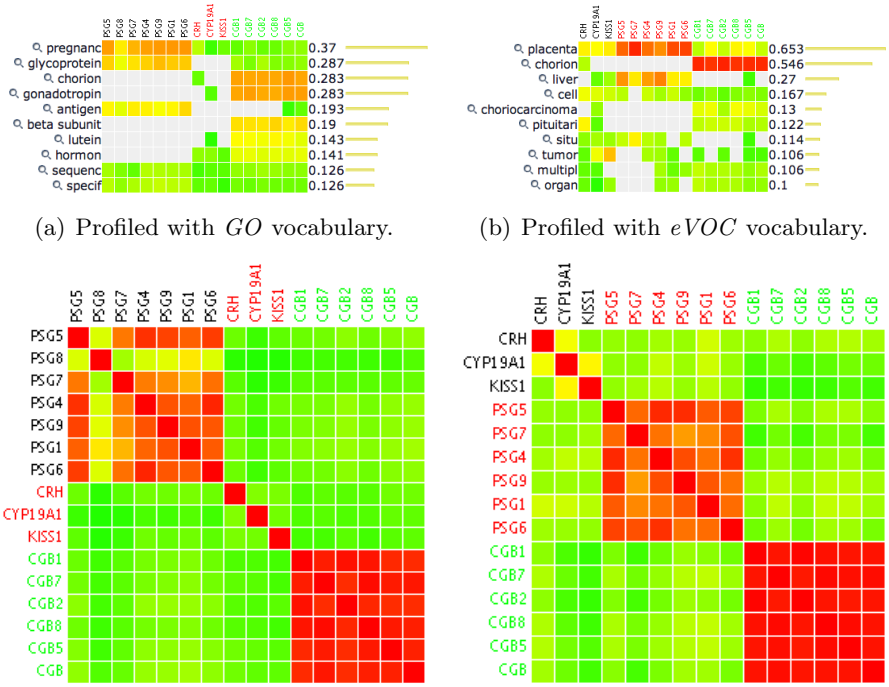
A deeper look into the constituent profiles of a gene group can shed light on the group's coherence. The next section will describe how the similarity between the textual profiles of individual genes can be visualized to study a group's structure.

3.2.2 Subclustering gene groups based on textual profiles

TXTGate automatically subclusters the profiled gene group according to the textual profiles of the individual genes. Clustering is performed on-line using Ward's minimum variance method. The linkage tree is cut to yield two clusters by default, but this setting can be adjusted by the user. TXTGate allows a fast and efficient investigation of a gene group's coherence through visualization of the clustered similarity matrix.

Figure 3.8 shows the *GO* and *eVOC* profiles and similarity matrices of the gene group with the second highest Silhouette coefficient after clustering based on expression data. The visualization allows a quick identification of the terms shared by the group's individual gene profiles. The group clearly consists of three subgroups: one with pregnancy-specific glycoproteins, a second with chorionic gonadotropin-beta polypeptides, and a third more heterogeneous group involved in placenta tumor. The *GO* profiles of the former two subgroups are shown in Figure 3.9(a) and 3.9(b). The latter subgroup shows more coherence when profiling with the *eVOC* vocabulary. Hence, it makes more sense to characterize it through its *eVOC* profile,

which is shown in Figure 3.9(c).



(a) Profiled with *GO* vocabulary.

(b) Profiled with *eVOC* vocabulary.

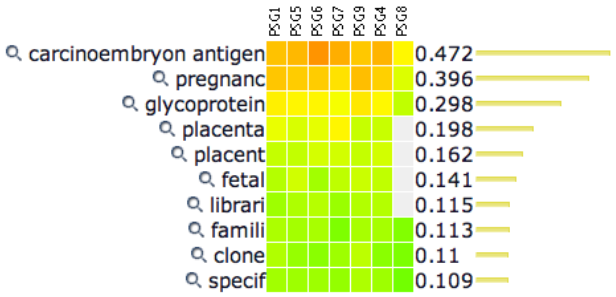
(c) Similarity matrix of *GO* profiles.

(d) Similarity matrix of *eVOC* profiles.

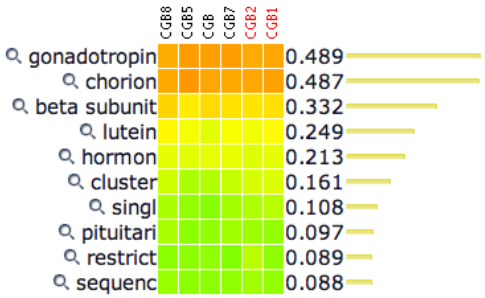
Figure 3.8: Profiles and similarity matrices. The figure shows the textual profiles and corresponding similarity matrices of all genes in the gene group with the second highest Silhouette coefficient after clustering based on expression data, as presented by TXTGate. The similarity matrices visualize the similarity between the individual genes of a gene group. The more red, the more similar two genes are.

3.3 Conclusion

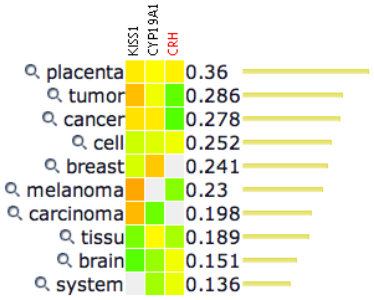
This chapter exemplified the characterization of groups of genes based on two different sources of information: Gene Ontology (GO) annotations and abstracts of literature references. Both allowed not only to determine the most prominent functional features, but also to establish a group's functional coherence (in terms of the average distance between the GO annotations) and to detect outliers (via subclustering of textual profiles).



(a) Subgroup 1 profiled with *GO* vocabulary.



(b) Subgroup 2 profiled with *GO* vocabulary.



(c) Subgroup 3 profiled with *eVOC* vocabulary.

Figure 3.9: Profiles of subgroups after sub-clustering. The figure shows the resulting TXTGate profiles after sub-clustering and textual profiling of the subgroups. Sub-clustering and profiling allows a more in-depth investigation of a group of genes.

From the moment a coherent group of genes is identified, it can be used to search the gene space for similar or related genes. This is elaborated in the next chapter.

Chapter 4

Expanding groups of genes

ONCE the rationale behind a group of genes is unveiled, the question rises which other known or unknown genes might also be part of the list. This is a well-known problem in association studies, where genes are linked with complex genetic disorders. The process of ranking genes according to their probability to be involved in a certain disease is called *prioritization*. In a way, a set of genes known to be involved in a certain biological process can be regarded as a seed, a small part of the puzzle ready to be enlarged. Every gene that might be involved in the biological process under investigation is a new starting point to set up a wet-lab experiment. This step closes the knowledge acquisition cycle (see Figure 4.1).

In this chapter two different approaches towards expansion of gene groups are described. The first is based on co-citation and co-linkage of genes in textual information, namely abstracts of scientific papers, as described in Section 4.1. In Section 4.2, the described method searches for genes that have similar properties as the genes in the seed group. The used properties are KEGG pathway membership, Gene Ontology (GO) annotations, textual descriptions from MEDLINE abstracts, microarray gene expression, EST-based anatomical expression, InterPro's protein domain annotation, BIND protein interaction data, *cis*-regulatory elements, and BLAST sequence similarity.

4.1 Gene co-citation and co-linkage

This paragraph deals with expansion of a group of genes through the analysis of gene co-citation in abstracts of publications. The hypothesis is that if two genes co-occur in an abstract of a scientific publication, they have a certain

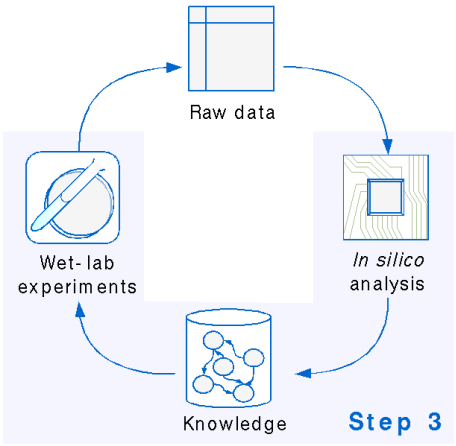


Figure 4.1: Step 3 of the knowledge acquisition cycle. In the third step, the newly gathered biological insights are used to postulate new hypotheses and design new experiments.

relationship, be it that they belong to the same functional category or that they are involved in the same or a similar biological process.

Several published co-citation approaches in the context of gene clustering were already reviewed in Chapter 2. Most of them stand or fall with correct identification of gene or protein names.

The approach taken here is conservative in that it only takes into account upper case human gene symbols and their synonyms from a curated list. It is based on the vector space model as described in Section 1.7. A gene-centered gene index was created by indexing all literature abstracts linked to human genes in Entrez Gene with a domain vocabulary comprising all uniquely defined human gene symbols. The vocabulary was derived from the lists provided by the Human Gene Nomenclature Committee (HGNC) [42, 67] of the Human Genome Organisation (HUGO) [134]. Gene symbol synonymy was resolved by mapping all known previous gene symbols and aliases to the HGNC approved symbol. The vocabulary was further pruned by removing the most frequently occurring symbols (1,000 times or more). In total the vocabulary consists of 11,623 unique gene symbols. Since these official gene symbols are frequently requested and used by scientists, journals and databases, it is assumed that they constitute a good first approximation to detect gene occurrence in abstracts from scientific papers.

The advantage of the vector-based approach over other approaches is that it allows to profile groups of genes. Thus it is able to characterize

gene pools consisting of all gene names that were present in the abstracts annotated to the genes in a group. In these gene pools, *co-linked* genes can be discovered. The term *co-linkage* refers to indirect links between genes. Every gene present via its symbol in an abstract annotated to a gene from the group, is co-linked with all other genes present in the pool of abstracts annotated to the gene group, and with all genes from the group, even if no other co-citations occur. The difference between *co-citation* and *co-linkage* is visualized in Figure 4.2.

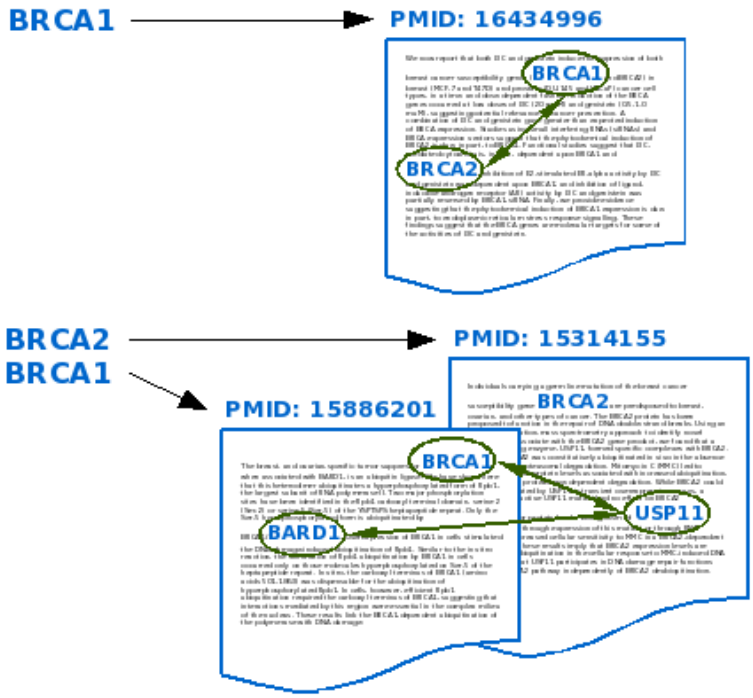


Figure 4.2: The concepts of co-citation and co-linkage. Two genes are co-cited if their symbols both occur in the same text body, as in the case of *BRCA1* and *BRCA2*. Genes are co-linked if there symbols occur in different but related documents, as in the case of *BRCA1* and *USP11*. Documents can be related because they are similar, or because they are connected to different genes in a set of functionally related genes. Co-linkage can reveal indirect links between genes that are never mentioned together in the same text.

This method was developed in the framework of the TXTGate project 3.2 together with Patrick Glenisson and Steven Van Vooren. The idea of co-linkage was introduced by Patrick Glenisson. The design of the gene-

centered domain vocabulary was done by the author. This gene-centered domain vocabulary is also available in the TXTGate application, allowing for online co-linkage analyses of gene groups.

4.1.1 Examples

Figure 4.3 lists the top-25 gene symbols from the average HUGO profile of the gene group with the highest Silhouette coefficient after clustering based on expression data (see Chapter 2). The TXTGate visualization gives a nice insight into which genes are co-cited and co-linked with each other. The group's profile gives an overview of all genes that are connected to the genes in the group. The profile consists of all gene symbols that were detected in one of the abstracts linked to one of the genes in the group. Thus it contains the complete pool of co-cited and co-linked genes.

As can be expected, most of the co-linked genes found this way are also present in the original gene group. For the other *new* genes, their role in skin development is worth investigating (see Chapter 3). The only input genes not present in the profile are *SPRR1A* and *SPRR2B*. Indeed, these symbols are not found in any of the abstracts of papers linked to them in Entrez Gene, which points out the benefit of this method over strictly co-citation-based approaches. In these abstracts, the gene symbols mentioned are *SPR*, *SPR2A*, *SPR2B*, *SPR3*, *SPRR*, *SPRR1*, *SPRR2*, *SPRR3*, *SPRC*, and *SPRK*. None of them is stated as being a synonym of *SPRR1A* or *SPRR2B* in the HUGO lists, but some of them clearly are. C1orf10 is a synonym of *CRNN*, thus also present in the profile.

TACR1 is the first *unknown* symbol in the profile. It stands for tachykinin receptor 1 and has no clear link with skin development. After investigation *SPR* seems to be an alias of *TACR1*, but this symbol was previously used instead of SPRR, the family of small proline-rich proteins to which the SPRR1A, SPRR1B, and SPRR2B proteins belong. Hence, *TACR1* is a false positive. The second *unknown* symbol is *SPRR3*, and an obvious symbol to pop up. *MAP3K11* and *PAEP* are again artifacts. *MAP3K11* shares an alias with a gene from the SPRR family (*SPRK*) and *K*PEP** is a consensus amino-acid sequence present in this same family, *PEP* being also a synonym of *PAEP*. *TIRAP* is the first interesting gene in the list as it seems to have a role in mediating LPS-induced NF-kappaB activation and apoptosis in human endothelial cells [12]. *SDS* is again an artifact as in the context of *EVPL* (with which it is co-cited) it is the abbreviation of *sodium dodecyl sulphate*. *TOC* or the tylosis oesophageal cancer gene is on the same chromosomal location as *EVPL*, explain-

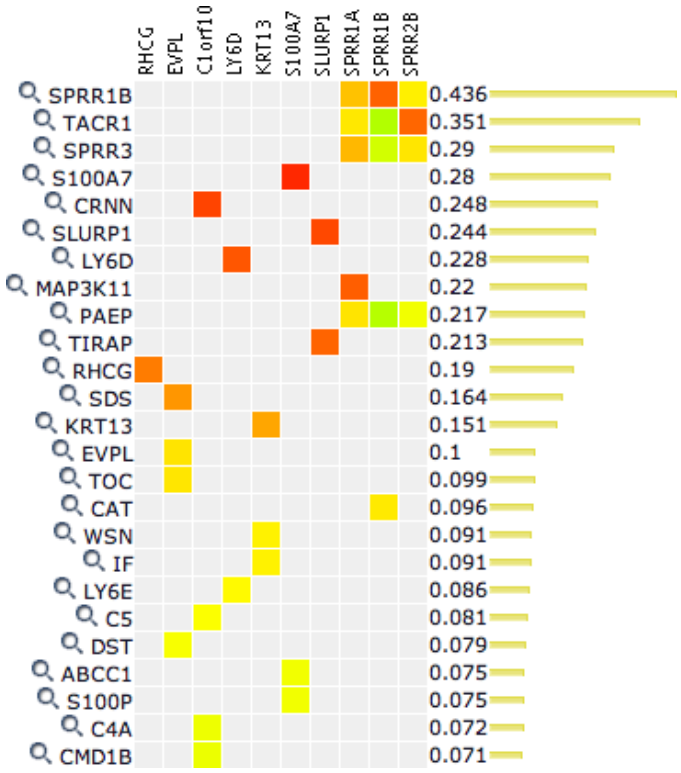


Figure 4.3: Co-citation profile of the gene group with the highest Silhouette coefficient after clustering based on expression data. The genes in the profiled gene group are on top of the figure. The genes co-cited and co-linked to them are at the left side. A colored box indicates that two genes are co-cited. If the box is grey, the two genes are co-linked. Only the 25 symbols with the highest weight are shown.

ing the association between focal nonepidermolytic palmoplantar keratosis (NEPKK; *tylosis*) and an increased risk of oesophageal cancer [113]. *CAT* comes up because it is mentioned in the abstract of a paper describing isolation and characterization of *SPRR1B*. The authors used the chloramphenicol acetyltransferase (*CAT*) reporter gene to study transcription regulation of *SPRR1B*. *IF* and *WSN* are not genes either, but abbreviations for *Intermediate Filament* and *White Sponge Nevus*, respectively. *LY6E* is again an obvious hit as, like *LY6D*, its encoding protein is part of the lymphocyte antigen 6 complex that plays a role in keratinocyte cell-cell adhesion [21]. *C5* and *C4A* encode for complement components described to be involved in psoriasis vulgaris, a chronic inflammatory dermatosis [91, 27]. The protein *DST* or dystonin encodes, belongs to the plakin protein family of adhesion junction plaque proteins. Some isoforms are expressed in epithelial tissue anchoring keratin-containing intermediate filaments to hemidesmosomes [87]. Both *ABCC1* and *S100P* are co-cited with *S100A7*. ABCC1 is a member of the superfamily of ATP-binding cassette (ABC) transporters. Both the gene products of *S100P* and *S100A7* belong to a family of low-molecular-weight calcium-binding proteins. *S100A7* is known to be involved in psoriasis [22]. *CMD1B* is again an artifact as its alias *FDC* coincides with the abbreviation of *follicular dendritic cell*.

4.1.2 Discussion

It is clear that the method described above has advantages over strictly co-citation-based methods for unraveling potentially interesting connections between genes. Nevertheless, it was unable to disambiguate homonyms, even in a very conservative setting. This resulted in a large number of *false positives*. The method did retrieve relevant gene symbols worth investigating further in the context of the gene group of interest, but careful investigation of each symbol was mandatory. Still, gene-centered co-linkage approaches can be a very efficient way to access the literature, as was illustrated by Hoffmann and Valencia [63, 64].

Because homonyms are the main problem, Latent Semantic Indexing (LSI) might give much more accurate results. LSI uses Singular Value Decomposition (SVD) to combine *synonymous* dimensions and separate *homonymous* dimensions. In this context, the domain vocabulary should probably be expanded and used without taking into account synonyms at the indexing level. Another concern is the quality and nature of the curated gene-document links from Entrez Gene. As discussed before, some linked documents are very general or not correctly annotated and can give rise to

noise in the results. As the amount of published literature increases every year, there is an urgent need for efficient tools that can help curators in providing correct annotations.

The next section describes an approach that takes into account a multitude of information sources to successfully reveal connections between candidate genes and a curated gene group.

4.2 Computational prioritization

The previous method only took into account information from the literature. In this section, a statistical procedure is described to combine different heterogeneous information sources to find genes with a high probability of belonging to a list of other genes. The method prioritizes lists of candidate genes based on the assumption that a new candidate gene has *similar properties* as a set of other genes that represents a certain biological case (a disease or pathway, for instance). The training set can be seen as a model for the biological case comprising multiple submodels, one for each information source that is included in the analysis.

In the field of linkage analysis and association studies researchers are often confronted with large lists of candidate disease genes, especially when investigating complex multigenic diseases. Investigating all possible candidate genes is a tedious and expensive task that can be alleviated by selecting for analysis only the most salient genes. The method was originally developed in this context, but it can be used to perform any kind of prioritization based on the similarity of a set of candidate genes with any (coherent) gene group. As today's high-throughput technologies (like the previously described microarray technology) spawn ever growing amounts of gene and protein data, it is clear that computational prioritization methods become increasingly important.

In the literature, three categories of computational candidate disease gene prioritization can be recognized:

***Ab initio* methods** *Ab initio* methods predict the association of a gene with a disease based on the values of a number of properties that are regarded suspicious. Examples of this are the location of the gene (does it lie within a region of linkage?), its sequence or sequence phylogeny (is the gene sequence well conserved?), annotation (what is the gene's molecular function?), expression level in certain conditions, and so on. *Ab initio* methods rely on information that is not related to previously known phenotypic and genotypic information about the disease.

Hauser *et al.* combined differential gene expression between diseased and control individuals with genomic linkage analysis to select the best candidate genes for a particular disease, as illustrated for Parkinson's disease, and called this procedure *genomic convergence* [60]. Similarly, Franke *et al.* combined linkage and association maps with microarray expression data in a tool called TEAM [45]. Turner *et al.* prioritized candidate disease genes based on the statistical over-representation of Gene Ontology and InterPro annotations [141], and van Driel *et al.* integrated location data, expression data, and phenotypic information in the GeneSeeker web application to filter genes based the co-occurrence of user-defined query terms [142].

Classification methods Classification methods attempt to classify genes or proteins based on their phylogenetic and physical properties. These methods do not try to associate a gene with a particular disease but to predict if a gene is disease-related or not. The use of these methods to associate genes with a particular disease is hampered by the small sizes of the training sets (in most cases only a few genes are associated with a disease) and the difficulty of defining negative training samples (genes that are not involved in the process).

There are two very similar reports on general disease probability prediction, both using decision trees as classification method: Lopez-Bigas and Ouzounis based their measure of disease probability purely on a gene's sequence and its evolutionary trace [79], whereas Adie *et al.* used gene features (gene length, for instance) and phylogenetic features in their software tool *Prospectr* [1].

Similarity-based methods Similarity-based methods prioritize genes by measuring the similarity of a set of properties with those of a disease or of genes known to be associated with a disease. The underlying assumption is that candidate disease genes are expected to have similar properties as the genes already associated with the disease. These methods rely on the existing knowledge of a disease, work well even with a small set of training genes, and do not need negative training samples.

The first systematic study using this type of gene prioritization was reported by Perez-Iratxeta *et al.* [100]. In a text-mining approach they used MEDLINE abstracts, Medical Subject Headings (MeSH) vocabularies, and Gene Ontology (GO) annotations to prioritize genes based on the similarity of these descriptions to those of a disease. Freuden-

berg and Propping measured the similarity of a gene with disease genes of related diseases using GO annotation data [47].

The method proposed here is a member of the last category. It differs from other described similarity-based methods in the way it combines all gene information to obtain one global measure of similarity. It is also extensible so that any new source of information can be added to the analysis and it is not restricted to disease-related prioritizations thanks to its generic nature. The methodology was developed, validated, and implemented in close collaboration with Stein Aerts. The work on *cis*-regulatory sequence elements was performed by Stein Aerts and Peter Van Loo. The methodology was tested and validated with real biological cases in close collaboration with Diether Lambrechts.

4.2.1 Methodology

Figure 4.4 overviews the different steps in the prioritization of genes with respect to a set of training genes using multiple heterogeneous information sources.

In the first step a training set `TRAIN` is compiled. All the 'properties' of the training genes are retrieved and prepared. In the case of Gene Ontology annotations, KEGG pathway membership, EST-based expression data, and InterPro protein domains, the statistically over-represented attributes are determined (see Section 4.2.3). For the textual information and the microarray gene expression data, the average profile of all individual gene profiles is taken. The BIND data are stored separately for each gene in `TRAIN`. The transcription factor binding site information of all training genes is compiled into one large vector. Also, the best combination of three transcription factors within human-mouse conserved non-coding sequences in the upstream sequences of the genes is recorded. For the sequence similarity a local BLAST database is created consisting of all coding sequences of the genes in `TRAIN`. All gathered data from one information source is a submodel for this source. All submodels together form a model for `TRAIN`.

In the second step a set of candidate genes `TEST` is compiled (for instance, candidate disease genes or candidate pathway members). All genes in `TEST` are then scored against the model for `TRAIN`. For each test gene the necessary information is retrieved and used to calculate a similarity by comparing it to the information contained in the submodels. This results in a list of scores for each submodel. Vector-based data are scored by the Pearson correlation between the test vector and the training average, while attribute-based data are scored by Fisher's omnibus meta-analysis (see Section 4.2.3),

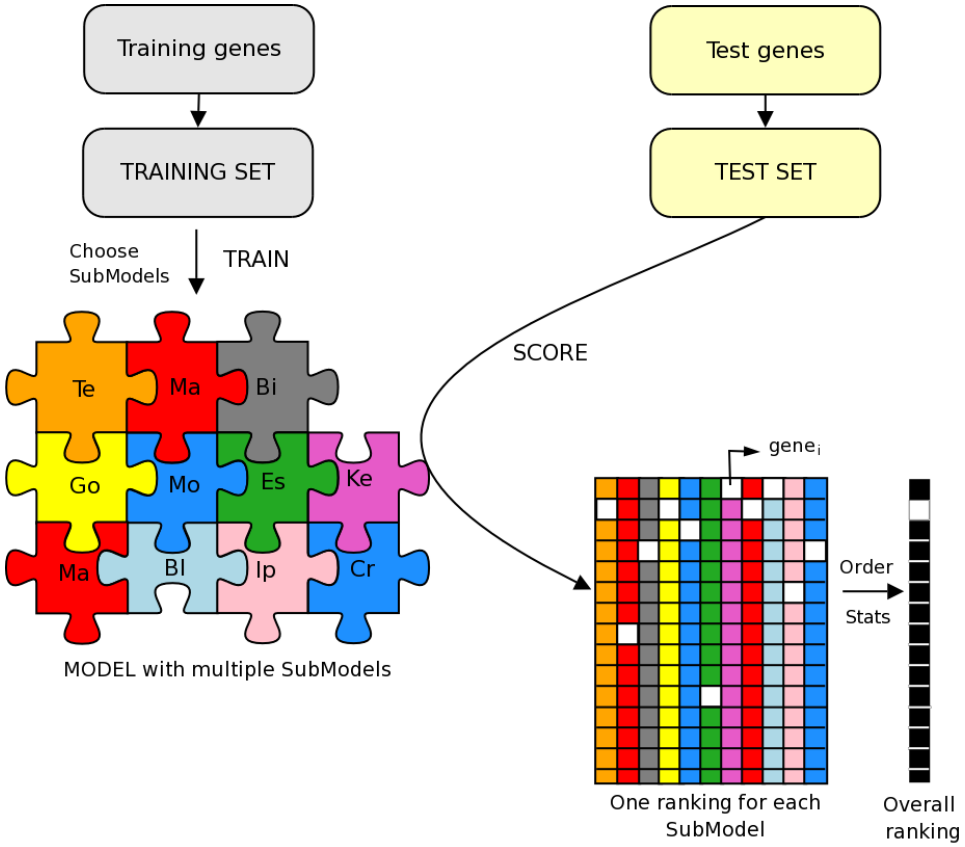


Figure 4.4: Candidate gene prioritization based on multiple data types. Currently the available data types are Textual data (Te), Microarray gene expression (Ma), Gene Ontology annotations (Go), BIND protein interactions (Bi), transcription factor binding sites or Motifs (Mo), *Cis*-regulatory modules (Cr), EST-based expression (Es), KEGG pathways (Ke), BLAST-based sequence similarity (Bl), and InterPro protein domains (Ip). There are two *Ma* models in the figure, illustrating the possibility of using multiple microarray gene expression data sets. The overall ranking of gene$_i$ (shown as a white square) is calculated from all individual rankings according to the different data types.

which combines the p-values for the over-representation of those training attributes that overlap with the test attributes. Taken together, all scores result in a matrix of scores, one for each gene and each submodel. Each list of scores (i.e., each column in the matrix, corresponding to one submodel) is then ranked independently from the other lists.

In the third step, all rankings according to the different submodels are combined to obtain an overall ranking of the candidate genes. This is done using the order-statistics formula for each gene separately (see Section 4.2.3). The formula takes as inputs the N rank ratios (i.e., the rank divided by the number of genes that have data available for this submodel), and gives as output a Q-statistic. This Q-statistic represents the probability that this gene is ranked at the observed positions by chance. The Q-statistic is then transformed into a global p-value using either a gamma distributions for which the parameters were estimated by random sampling. Finally, all TEST genes are ranked according to this global p-value, which results in the final prioritization.

4.2.2 Data sources

In practice, any information source that allows to define a similarity between a candidate gene and a set of training genes can be used. During the validation of the described method only ten sources of information were used. They are described below.

Text-mining: Entrez Gene and MEDLINE TXTGate [56] is a text-mining application designed towards the analysis of the textual coherence of groups of genes (see Chapter 3). TXTGate's textual profiles constructed with the Gene Ontology (GO) vocabulary are used for the prioritization. Gene prioritization based on textual data is done by calculating the cosine distance (see Chapter 1) between a test gene's textual profile and the average textual profile of the training genes. A high similarity between a test gene and the training genes means that the core of the literature abstracts that describe both, have a lot of terms in common, and they thus talk—in a general sense—about the same subject, no matter what the detailed messages in the abstract are. That is, textual profiles with contrasting statements that use the same words will still be similar. For example, if an abstract on gene x states that *protein X stabilizes tau plaques* and an abstract on gene y states that *protein Y solubilizes tau plaques*, the textual profiles of gene x and gene y could be similar due to the common occurrence of

the phrase *tau plaques*. In the context of the presented approach (allowing prioritization based on general similarity to an heterogeneous set of disease genes), this general measure of similarity may be an asset rather than a drawback.

Functional annotation: GO and KEGG In those cases where the textual profiles could suffer from the noise in the literature data, the curated Gene Ontology (GO) data brings salvation. GO is a manually curated vocabulary that is used for the functional annotation of genes [132] and is structured as a hierarchical tree. Prioritization is done by comparing the GO annotation of a test gene with the statistically over-represented GO terms in the training set (see Section 3.1.1). For example, if the proteins of most genes in a training set are in some way involved in linking cytoskeleton filaments to the plasma membrane, then GO terms like *cytoskeleton* (GO:0005856), *cytoskeletal anchoring* (GO:0007016), and so on, could be over-represented. If one of the test genes is annotated with any of these terms, it will get a high ranking according to the GO data.

The KEGG database [71] is an even more structured source of functional annotation. It contains the members of known biological pathways. Similarly as for GO, we calculate whether certain pathways are over-represented in the training set and will give a good score (i.e., a low rank) to those test genes that are involved in one of the pathways that is important for the training set.

Protein information: InterPro and BIND InterPro is a database of protein families, domains and functional sites [92]. For each training gene the InterPro attributes are retrieved from the Ensembl Mart database (the presented results are based on the ensembl_mart_25_1 database). An example of an InterPro attribute is IPR000418 (Name=Ets-domain) for which there are nineteen human proteins known to carry this domain. Scoring test genes using the InterPro protein domains is done by meta-analysis (see Section 4.2.3). If a certain protein domain is over-represented in the training set as compared to the full genome, and if a test gene also carries this particular domain, then it will get a good ranking according to the InterPro data.

Another interesting data type to score test genes is protein interaction data, for which data is taken from the Biomolecular INteraction Database (BIND) [10]. BIND contains interaction data from high-throughput experiments (yeast two-hybrid assays, for example) and

from hand-curated information gathered from the scientific literature. The idea behind using protein interaction data for gene prioritization is that one can expect a test gene to be more related to the training set if its protein directly interacts with one of the proteins of the training genes, or if it has a common interaction partner with one of them. In practice, all the proteins of the training genes and all their interaction partners are collected and the overlap between this set and the set containing a protein (encoded by a test gene) and its interaction partners is used to calculate a similarity score.

Gene expression: microarray data and ESTs Several research groups already described the use of microarray data for the prioritization of genes [142, 60, 45]. Here, the gene atlas of human protein-encoding transcriptomes measured in 79 normal human tissues [126] is used. However, it is obvious that disease- or process-specific microarray data are more informative for particular training and test sets. For example, if a geneticist has performed his or her own microarray experiment that measures gene expression in healthy versus diseased patients (or if such data are available in public repositories, such as ArrayExpress or GEO), then a prioritization based on these data is more likely to give good performance.

Next to microarray-based gene expression data, the large repositories of EST-based anatomical expression in the human body also contain valuable information that can be used for gene prioritization. The EST-based expression data available via the Ensembl Mart database are used. As is done for GO, model training consists of calculating a p-value for each anatomical site that measures its statistical over-representation within the training set. Scoring a test gene with this EST-based model is done by meta-analyis (see Section 4.2.3).

***Cis*-regulatory elements** The prioritization process uses *cis*-regulatory information in two different ways. Firstly, all (offline) predicted instances of a library of transcription factor binding models (position weight matrices or PWMs), in all human-mouse conserved non-coding sequences (CNS) upstream of a test gene (10 kilobases), is compared with the averaged instances of the training set. More information on this data set can be found in [6, 4]. The predicted binding sites of all available transcription factors are recorded in a vector (for instance, of length 400 if there are 400 PWMs), where each element represents the best score of this PWM in all human-mouse conserved sequence

blocks upstream of that gene. Comparison with the training vector is done by calculating the Pearson correlation.

Secondly, the best combination of three transcription factors within maximally 300 bp in the set of human-mouse CNSs is searched in the training set using the Genetic Algorithm version of the ModuleSearcher algorithm [4], using 20 generations. Scoring of a test gene is done by the ModuleScanner algorithm [6] that essentially sums up the best scores in all test gene CNSs of the three PWMs of the trained model.

Sequence similarity: BLAST There are examples of diseases that can be caused by proteins of the same family, for example Presenilin 1 and Presenilin 2 in Alzheimer's disease. The *e*-value of the BLAST between the (longest) coding sequences of these two genes is 10^{-133}, thus they are highly similar. One can imagine that a researcher would perform a BLAST search of a number of test genes and use his or her expert knowledge to judge whether the hits make sense. In the same sense a BLAST search is performed to score test genes against a set of training genes. Judging whether a hit is relevant is done automatically by restricting the BLAST search on an ad hoc created BLAST database consisting of all coding sequences of the training set. Test genes that have a low significance value of the BLAST are similar to one of the training genes and will get a low rank.

4.2.3 Computational techniques

Vector-based similarity measures

For information sources summarized using a vector representation, the Pearson correlation is used (in the case of microarray gene expression data and transcription factor binding sites)

$$\mathrm{r}_{\mathrm{Pearson}}(i,j) = \frac{\sum_q (x_{i,q} - \overline{x}_i)(x_{j,q} - \overline{x}_j)}{\sqrt{\sum_q (x_{i,q} - \overline{x}_i)^2}\sqrt{\sum_q (x_{j,q} - \overline{x}_j)^2}},$$

or the cosine similarity (in the case of textual information from MEDLINE)

$$\mathrm{sim}_{\cos}(i,j) = \frac{\sum_q x_{i,q} x_{j,q}}{\sqrt{\sum_q x_{i,q}^2}\sqrt{\sum_q x_{j,q}^2}}.$$

Meta-analysis

For the GO, KEGG, EST, and InterPro data types, the following meta-analysis is used to calculate a similarity score for a test gene compared to a set of training genes. For each gene in a set of training genes, all relevant attributes are collected (see before). Next, for each attribute a p-value is calculated using a binomial statistic that represents the statistical over-representation of this attribute within the training set (as is done for GO annotations in Chapter 3).

Coherent training sets for any of the four characteristics will contain statistically significant p-values. When a group of test genes is scored using these data types, the p-values p_i corresponding to the annotated attributes of a test gene are combined using the Fisher statistic:

$$S = \sum_{i=1}^{n} -2\log p_i$$

Under the null hypothesis of uniformly distributed p-values, the summary statistic S has a χ^2-distribution from which a meta-analysis p-value can be extracted. The test genes are then ranked according to this new p-value.

Order statistics

The heterogeneity of the scoring results of individual information models (correlations, p-values, counts, etc.), makes a meta(-meta)-analysis not trivial. A standard meta-analysis with Fisher's method requires a p-value for every submodel, which is difficult, because calculation of p-values from correlation measures, for instance, is not straightforward. The method presented here combines the n different rankings $R_1, R_2, ..., R_n$ of a gene (one for each of the n data types used) using order statistics.

Given a set of elements D and a sequence of n independent and identically distributed random variables $X = (X_1, X_2, \ldots, X_n)$ with $X_i \in D$, an order statistic of order i is defined as the random variable $X_{n,i}$ that represents every i'th smallest element of all possible combinations of size n chosen from D. The joint probability density function of all order statistics of X represents the probability to obtain an observed sequence of ordered elements by chance alone (see Appendix A and `http://www.math.uah.edu/stat/sample/OrderStatistics.xhtml` for a more elaborate description).

In the context of the described approach to prioritize genes, the n random variables correspond to the n different submodels and represent the gene rankings. The rankings R_i are divided by the total number of ranked

genes for a data source, excluding genes with no rank because of missing values. The resulting rank ratios r_i take values between zero and one. The cumulative probability distribution of all n order statistics can now be used to calculate the probability of observing the obtained and ordered rank ratios $r_1, r_2, ..., r_n$ of a gene, or a better sequence of ordered rank ratios (i.e., with smaller rank ratios), by chance:

$$Q(r_1, r_2, \ldots, r_n) = n! \int_0^{r_1} \int_{s_1}^{r_2} \ldots \int_{s_{n-1}}^{r_n} ds_n ds_{n-1} \ldots ds_1.$$

Stuart *et al.* [125] propose following recursive formula to efficiently compute this integral:

$$Q(r_1, r_2, ..., r_n) = \sum_{i=1}^{n} (r_{n-i+1} - r_{n-i}) Q(r_1, r_2, ..., r_{n-i}, r_{n-i+2}, ..., r_n),$$

with $r_0 = 0$. However, this formula is only tractable for $n < 12$ because its complexity is $O(n!)$. An alternative formula with complexity $O(n^2)$ allows computation for $n > 12$:

$$V_k = \sum_{i=1}^{k-1} (-1)^{i-1} \frac{V_{k-i}}{i!} r_{n-k+1}^i$$

with V_0=1. The solution of the integral can be found by calculating V_n. The proof that $V_n = Q(r_1, r_2, \ldots, r_n)$ is given in Appendix A.

Since the q-values calculated this way are not uniformly distributed, a p-value can only be obtained by sampling a distribution for every possible number of ranks, as elaborated in Appendix A. From the cumulative distribution functions of these distributions a p-value can be drawn for every q-value from the joint probability density of the order statistics of dimension n. Next to the original n rankings, the ordered p-values of all genes result in an $(n+1)_{th}$ ranking based on the combination of all data sources.

Handling missing values

The order statistics approach offers two advantages to deal with missing values. First, it allows to compare genes with a different number of rank ratios. This is useful as many genes have missing values in at least some of the data sources. Second, because the order statistics formula uses rank ratios instead of absolute ranks, the denominator of the ratio can be adjusted. To avoid artificially low rank ratios in data sources for which a lot of

genes have missing data, the number of test genes without missing values (i.e., the ones that actually received a ranking) is used as the denominator, instead of the total number of genes. This way, information sources with a different number of genes with missing values can still be combined.

Test genes for which there is data available, but for which there is no similarity with the training set, also have to be ranked with caution. Such genes have the highest (i.e., worst) possible score for a particular data source. Imagine the case where all test genes have the same extreme score. They will all get the same (best) rank of 1 and the order statistics will rank them high in the overall rank. To avoid this problem, all genes with maximal dissimilarity to the training set get a rank that equals $N - D/2$ with N the total number of test genes with information and D the number of genes with maximal dissimilarity.

4.2.4 Statistical validation

The performance of the methodology in prioritizing candidate disease genes on the one hand, and candidate members of biological pathways on the other hand, was tested with a large-scale cross-validation experiment. For the disease candidate approach, a list of 29 Online Mendelian Inheritance In Man (OMIM) diseases was compiled for which at least nine contributing genes were known. Automated HUGO-to-Ensembl mapping reduced the number of genes for a few diseases. The smallest gene set was the one for Amyotrophic Lateral Sclerosis (ALS) with only 4 Ensembl genes, and the largest one was the leukemia gene set with 113 genes. In total, the test comprised 627 disease associated genes with an Ensembl identifier and a disease gene set contained on average 19 genes. For the pathway candidate approach, three lists of genes were compiled from Gene Ontology annotations: one with Wnt pathway members (GO:0016055: Wnt receptor signaling pathway), one with Notch pathway members (GO:0007219: Notch signaling pathway), and one with EGF pathway members (GO:0007173: epidermal growth factor receptor signaling pathway). The latter two GO categories contained only a limited number of associated human genes. For these, the human orthologous genes of the fly pathway members were added to the set. For an overview of all genes included in the cross-validation the reader is referred to Appendix B.

Cross-validation procedure

Figure 4.5 describes the different steps in the cross-validation procedure. For each gene set, a leave-one-out cross-validation was performed: at each run, one gene was left out while the remaining genes were used to construct a training set. Each of the available data types (see Section 4.2.2) was used to train a submodel, summing to 10 submodels in total. Then, the left-out gene and 99 genes randomly selected from Ensembl were used as test set. 100 random sets were constructed, out of which for each tested left-out gene one set was randomly selected. The rankings for each separate submodel, as well as the combined ranking for all submodels based on the order statistics were recorded.

Rank ROC curves

The results of the cross-validation can be visualized in a Rank ROC (Receiver Operating Characteristic) curve, where the y-axis represents the sensitivity (i.e., the proportion of true positives) and the x-axis represents one minus the specificity (i.e., the proportion of true negatives):

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Because the ROC curve visualizes the performance of the applied method rather than the performance of one model, and because it is based on the rankings of multiple prioritizations with different models, it is called a Rank ROC. The values in the above formulas are calculated from all iterations and their interpretation is the following: (1) the number of true positives (TP) is the number of times that the left-out gene is ranked above the cut-off; (2) the false positives (FP) are all genes besides the left-out gene that are ranked above the cut-off (these can be thought of as being retained for further evaluation but they are probably not associated with the disease or pathway at hand); (3) the true negatives (TN) are those genes that are ranked below the cut-off and that are not the left-out gene; and (4) the number of false negatives (FN) is the number of times that the left-out gene is ranked below the cut-off (in these cases the real disease- or pathway-associated gene is not retained for potential further analysis). In a Rank ROC curve as in Figure 4.6, the sensitivity and (1-specificity) are plotted for each possible cut-off value. On the one hand, such curves can be used to

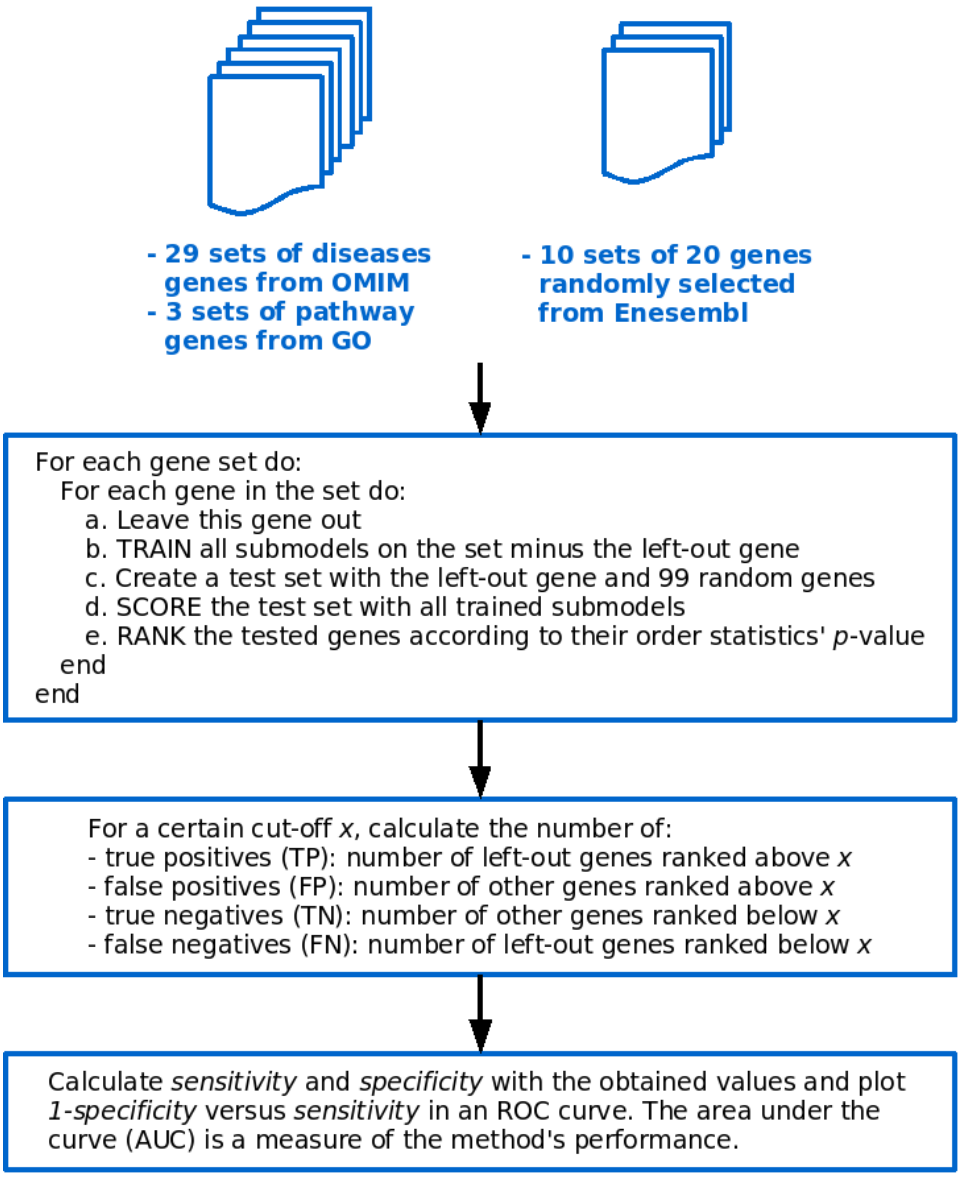


Figure 4.5: The cross-validation procedure. Schematic overview of the different steps in the large-scale cross-validation of the computational prioritization methodology.

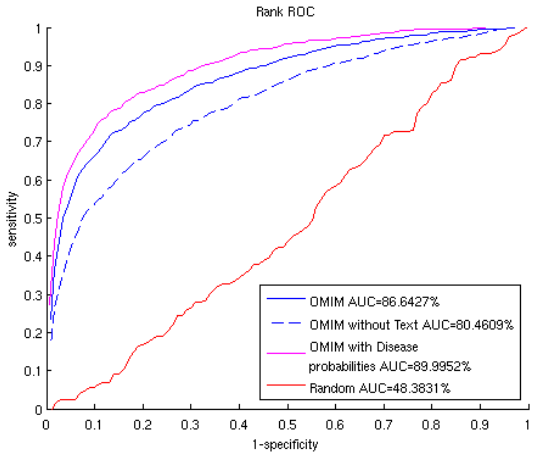
choose a cut-off value (giving a desirable balance between FP and FN). On the other hand they can be used to compare different kinds of prioritizations. The area under the curve (AUC) is a measure of the performance, which would equal one if every left-out gene (i.e., the wanted disease/pathway gene in each test set) is ranked first for all tested sets. The AUC would be 0.5 if the prioritization is not better than ranking the genes randomly. Although the Rank ROC is not an ROC per se, it is an appropriate measure of the proportion of genes correctly (incorrectly) included (or left-out) from a list of follow-up genes—as a function of the length of such a follow-up list.

Performance of the combined model

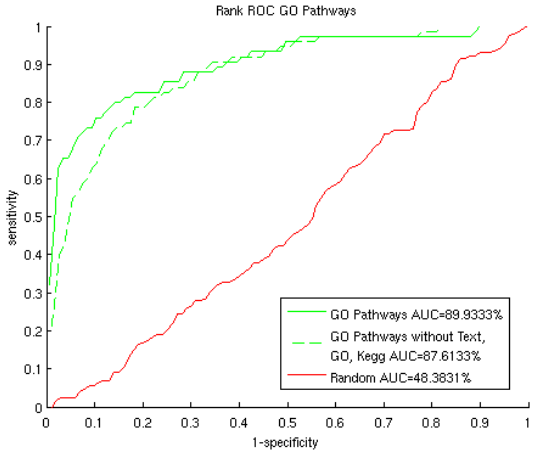
Figure 4.6 shows the Rank ROC curves for the rankings of all leave-one-out cross validations for the OMIM and GO-pathway study. In the same figure, the Rank ROC curve of the same leave-one-out cross-validation using random training sets is also plotted. The validation experiment results in a biologically meaningful prioritization that is significantly better than random prioritizations. Overall, the left-out gene ranks among the top 50% of the test genes in 85% of the cases in the OMIM study, and in 95% of the cases in the GO study. In about 50% of the cases (60% for the pathways), the left-out gene is found among the top 10% of the test genes.

Performance of individual submodels

Figure 4.7 shows the AUC values of all submodels individually. Every separate submodel performs better on real data than on randomized data. The best performing model for OMIM is the text model (93%), because of explicit co-occurrences of a gene and a disease in the same abstract, and which is therefore an artificially high percentage. In a real disease prioritization case, the text model can, in some cases, still capture the knowledge that is building up to the discovery of the disease association (see further). It is apparent that the text performance is lower in the pathway-study, pointing at a far less explicit mentioning of the pathway when the function of a pathway member is described. Expression-based data (both EST and microarray) generally perform well, both for diseases and pathways. This is however only measured with general microarray data (normal human tissues), and it is expected that the microarray performance will increase when disease- or pathway-specific expression data is used. Protein domains (InterPro) and sequence similarities (BLAST) are reasonably useful for diseases, but more for pathways. This might be caused by the high number of paralog-



(a) Rank ROC curves for the 29 OMIM diseases.



(b) Rank ROC curves for the 3 GO categories.

Figure 4.6: Rank ROC curves of the cross-validation. The genes for 29 diseases selected from OMIM and 3 pathways derived from GO are used as training sets. Each gene is used once as a left-out gene and scored together with 99 randomly chosen genes against the submodels trained on the remaining left-in genes. After calculation of the true and false positives, and true and false negatives, *1-specificity* is plotted against the *sensitivity* to obtain a Rank ROC curve. The area under the curve is a measure of the performance of the method with respect to identifying candidate disease genes and candidate pathway members.

ous pathway genes. Notably, also the motif models (both single motifs and *cis*-regulatory modules) are performing better for the pathways. It is indeed expected that members of the same pathway are more tightly co-regulated than genes that are linked to the same disease.

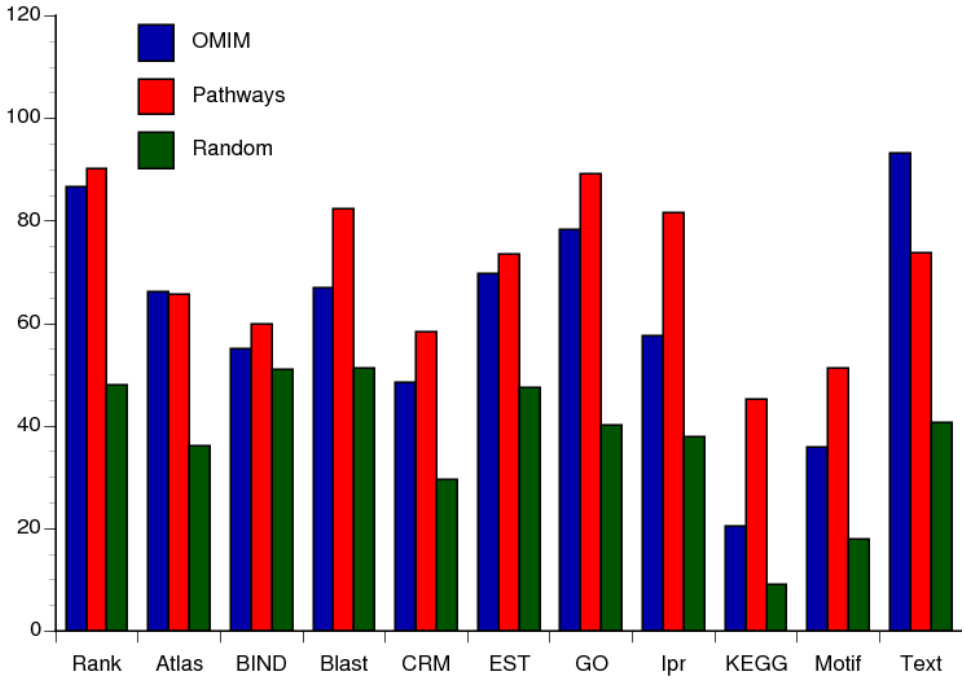


Figure 4.7: Area under the curve values of the different submodels. This bar graph visualizes the area under the curve as a measure of the performance of the different submodels with respect to the OMIM diseases and GO pathways cross-validation. Also plotted is the performance of submodels built using randomly selected genes.

The bad performance of the KEGG model is mainly attributable to the high numbers of missing values. Actually, if missing values are not taken into account in the performance calculations, the performance rises from 20,43% to 89,53%, meaning that if data is present, the prioritization is good. The bad performance of the BIND model (small difference with the random sets) could be caused by high levels of noise in protein interaction data (from yeast-2-hybrid experiments, for instance). Models like KEGG and BIND are typically expected to become better as better annotation and better high-throughput interaction data become available.

When combining the five submodels with the lowest absolute perform-

ances (KEGG, BIND, Motif, CRM, InterPro), the AUC of the cross-validation is 77.1%. In other words, when submodels with mediocre individual performances are used, the performance of the combined ranking can still be significantly better than random. It is therefore useful to include all models in the prioritization process. There can be several reasons why a submodel has a low AUC in this large scale study, but still contributes significantly to the combined ranking. One reason is that, especially in the case of the OMIM, the AUCs are averages across different training sets, some of which may be modeled well by certain submodels but less well by others. For example, a cross-validation on Alzheimers disease alone yields an AUC of 76.3% for the Atlas microarray data, which is much higher than the average AUC of all diseases. Figure 4.8 shows the variation in performance of the different submodels. A second reason is the high number of missing values for certain submodels (as with the Kegg submodel). Nonetheless, if the performance of these submodels is compared with their performance if random training sets are used, the AUCs are always higher for the former (see Figure 4.7).

Unbiased validation of the Text-based submodel

For each OMIM gene that was used for the disease validation described above, a mutation causing the disease had previously been reported in a *landmark* study. The inclusion of publications, documenting a direct link between the gene mutation and the disease under study, may artificially increase the relative contribution of the literature data source in the overall performance. To remove this bias from the validation results, the entire literature database was excluded from the disease validation protocol. For the same reason, the GO, KEGG, and literature data sources were excluded for the validation of pathway genes. Even under such unrealistic conditions where entire data sources are not used, the overall performance was only negligibly affected: the performance dropped by only 6.1% for disease genes (from 86.6% to 80.5%) and by only 2.3% for pathway genes (from 89.9% to 87.6%; see Figure 4.6). Thus, the diversity of the used data sources enables meaningful prioritizations, even without the use of literature information.

Clearly, this caution is only of importance in the context of a validation. In a more realistic situation, when the precise function of novel genes in the pathogenesis of a disease is not yet known, the literature could still provide valuable indirect information about other properties of these genes. In a study of 10 monogenic diseases, this situation was mimicked by using only *rolled-back* literature information, available one year prior to the *landmark*

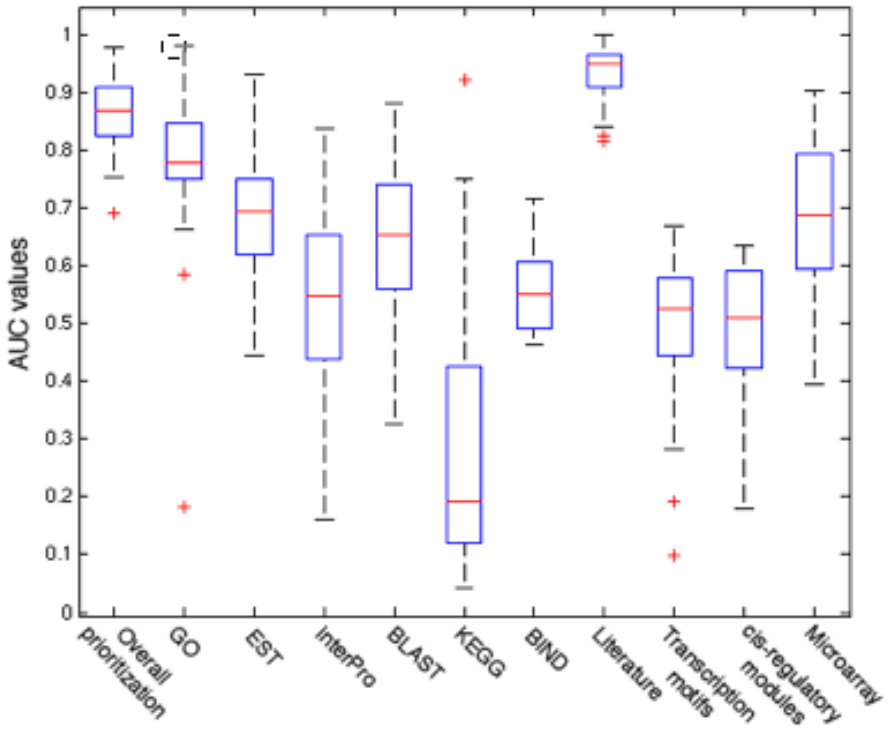


Figure 4.8: Boxplots of the variability of the area under the curve per submodel. A boxplot is a visual representation of the dispersion and extreme values of a data set, in this case the AUC values that indicate the performance of the prioritizations per submodel. The blue boxes are the 0.25 and 0.75 percentiles (the quartile scores), the red lines are the median values, and the red plus signs are the outliers. The boxplots are based on the cross-validation data from the 29 OMIM diseases. They indicate how consistent the performance of a submodel is over different disease training sets. The textual model based on literature data (Literature) performs best and is the most consistent one, as compared to the other submodels.

publication. Even then three genes received a good ranking (position 1, 1 and 3 out of 200 test genes; see Table 4.1), thus illustrating that the literature data source indeed contributes to the prioritization of some yet undiscovered disease genes. For the seven other genes, use of the literature as the only data source was not very efficient, but inclusion of all the other data sources did yield a high rank (see Table 4.1). Even though the literature may provide valuable information, this points out that the described methodology does not rely on the literature as the only critical data source.

Pairwise dependencies between submodels

If some of the submodel rankings were correlated, then the combined ranking would be biased because the order statistics require independent rank ratios. In that case, nonetheless, the combined ranking could still be used as an approximate ranking if the obtained p-values are not considered in the decision of a cut-off (instead, the threshold could then be purely based on selecting a certain percentage of test genes). To test whether certain submodels were correlated, all individual rank ratios of the different submodels of 29 sets of 50 random genes prioritized with the 29 OMIM disease models were recorded. The pairwise Pearson correlations between these rank ratios are shown in Figure 4.9. All submodels are positively correlated with the overall order statistics rank (*Ra* in the figure). In other words, they all contribute to the overall ranking. Those with a low correlation have a lot of missing data (for instance, in the case of KEGG (*Ke*) only 10.3% of the genes had data; in the case of the *cis*-regulatory motif data (*Mo*) only 35,7% of the genes had data). The pairwise correlations between the different models are all around zero, so no large biases in the order statistics rank are expected. This means the p-values of the order statistics can be used as a significance threshold.

Bias towards known genes

In general *there already* is *a bias in candidate gene selection towards well-characterized genes* [45]. It was expected that the described methodology should at least partly alleviate this bias, and should allow for unknown or less known genes to be ranked highly, because (1) genes are prioritized based on *multiple* information sources instead of one or a few; and (2) not only functional information (GO, text, and pathway information, for instance) is used, but also data sources that are equally valid for known and unknown genes (i.e., prior-knowledge-independent data), namely microarray gene ex-

Table 4.1: Results of a literature roll-back experiment. For ten monogenic diseases, only the literature published before the *landmark* paper describing the link between the disease and its causing gene, were included in the Text submodel. The genes *CACNA1C*, *CRELD1*, and *CAV3* received a good ranking (position 3, 1 and 1 out of 200 test genes. This illustrates that the literature data source indeed contributes to the prioritization of some yet undiscovered disease genes. For the seven other genes, use of the literature as the only data source was not very efficient, but inclusion of all the other data sources did yield a high rank, indicating that the described methodology does not rely on the literature as the only critical data source.

Disease	**Gene**	**Ensembl ID**	**Pub. date**	**All**	**Text**
Arrhythmia	*CACNA1C*	ENSG00000151067	10-04	4	3
Congenital heart disease	*CRELD1*	ENSG00000163703	4-03	3	1
Cardiomyopathy 1	*CAV3*	ENSG00000182533	1-04	2	1
Parkinsons disease	*LRRK2*	ENSG00000188906	11-04	50	*
Charcot-Marie-Tooth	*DNM2*	ENSG00000079805	3-05	14	100
Amyotrophic lateral sclerosis	*DCTN1*	ENSG00000135406	8-04	27	97
Klippel-Trenaunay disease	*VG5Q*	ENSG00000164252	2-04	3	39
Cardiomyopathy 2	*ABCC9*	ENSG00000069431	4-04	1	51
Distal hereditary motor neuropathy	*BSCL2*	ENSG00000168000	3-04	15	62
Cornelia de Lange	*NIPBL*	ENSG00000164190	6-04	9	75
Average Rank				13 ± 5	48 ± 13

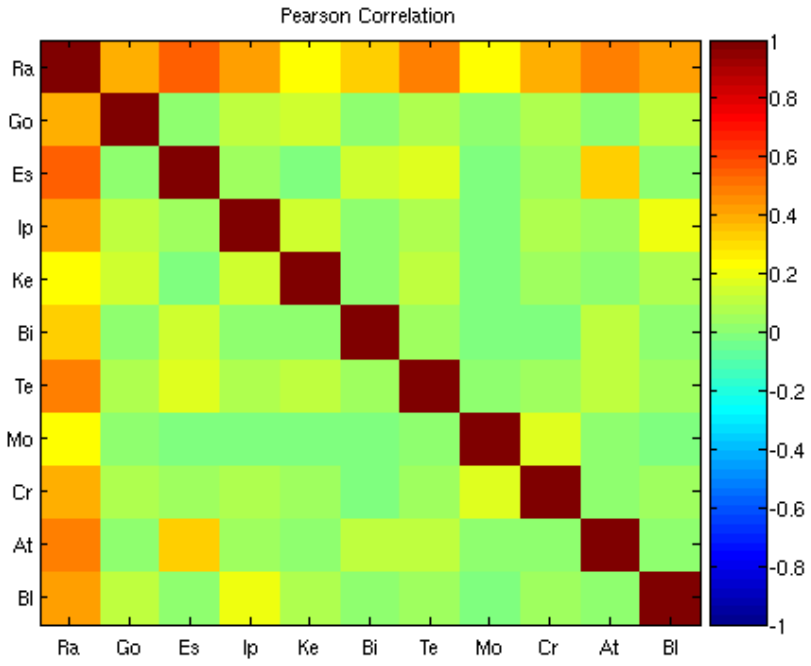


Figure 4.9: Pearson correlation between the different submodels. Correlation between the rank ratios of different data types and between the rank ratios of the individual data types and the overall rank ratio is shown. *Ra* overall rank ratio; *Go* Gene Ontology; *Es* EST-based expression; *Ip* Interpro domains; *Ke* KEGG pathways; *Bi* BIND protein interactions; *Te* text-mining; *Mo* motifs or transcription factor binding sites; *Cr* *cis*-regulatory modules (combinations of motifs); *At* Atlas microarray expression data; *Bl* sequence similarity using BLAST.

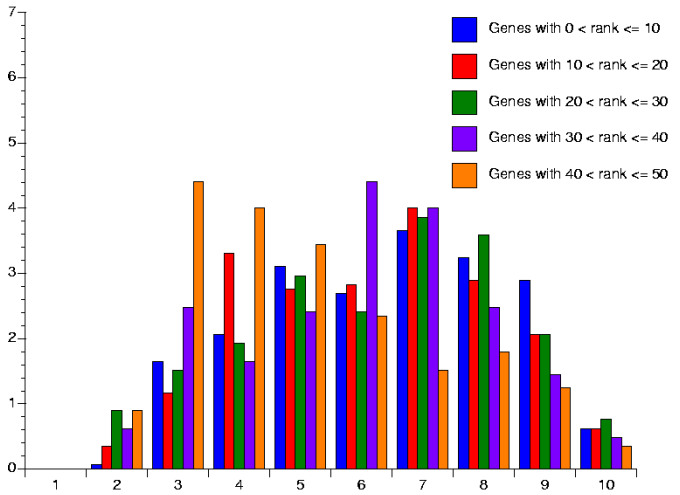
pression data, EST-based gene expression data, protein domain predictions from InterPro, protein interactions, sequence similarities, and *cis*-regulatory data.

To establish the magnitude of the bias towards well-characterized genes the influence of the number of information models for which a certain gene has data available, on the possibility for this gene to get a good rank was investigated. Figure 4.10 shows for each number of available information models the percentage of genes that is ranked between 0 and 10, 10 and 20, and so on in a list of 50 random test genes that are prioritized according to the 29 disease models. There is only a slight trend of higher rankings for genes with more available information. A second indicator of how well a gene is characterized can be the presence of a HUGO gene symbol. In Figure 4.10 the same information was plotted for genes with and without HUGO gene symbol. Again only a slight bias can be seen towards known genes. It is apparent that even genes with very little information (from three submodels onwards), can reach the top 10 in a test set of 50 genes.

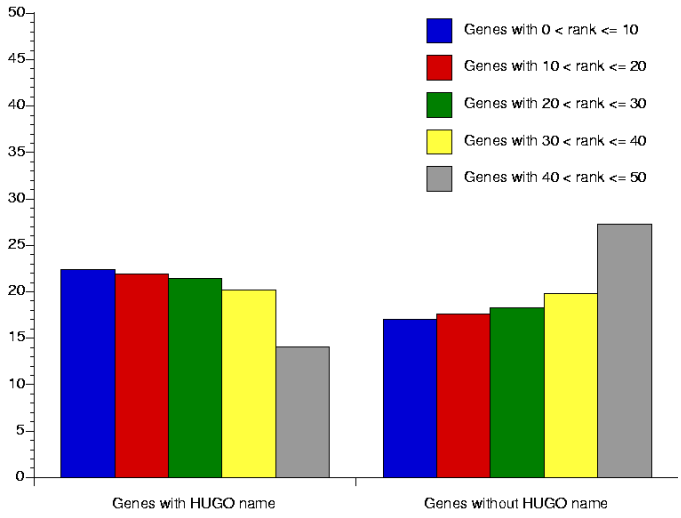
4.2.5 Discussion

In this high-throughput era in biology, there is a clear need for methods that perform efficient and statistically sound computational prioritizations. The method described above based on order statistics clearly has several advantages over other approaches. It solves the problem of missing data and reconciles even contradictory information sources. It allows for a statistical significance level to be set after multiple testing correction, thus removing any bias otherwise introduced by the expert during manual prioritization. It also removes part of the bias towards known genes by including data sources that are equally valid for known and unknown genes. Even genes for which information from as few as 3 data sources is available, can receive a high ranking (see Figure 4.10(a)).

Nevertheless, order statistics has several limitations and better methods to combine the data sources might exist. It is, for instance, impossible to set weights on the different data sources: all data sources contribute equally to the overall ranking. However, depending on the case at hand and the quality of the data sources, an expert might want certain data sources to have a smaller or larger overall impact on the outcome. It is also clear that, to retain statistical significance, the used data sources need to be uncorrelated. Adding or removing data sources from the analysis can have a significant impact on the outcome. Therefore a careful evaluation and selection should be made when selecting data sources for inclusion.



(a)



(b)

Figure 4.10: Bias to well-characterized genes. 29 sets of 50 random test genes were prioritized based on 29 OMIM disease training sets. (a) The number of available information models is plotted against the percentage of genes within each ranking interval. (b) The bars represent the percentage of genes within each ranking interval for genes with a HUGO symbol and genes without HUGO symbol.

4.3 Conclusion

This chapter showed how the functional space surrounding a gene group can be explored to identify potentially interesting genes, previously not known to be involved in the process of interest. Two different methodologies were discussed. The first tried to find indirect links between genes mentioned in the literature using co-occurrence and co-linkage information. As many text mining approaches, this method also struggled with correct identification of gene symbols, resulting in an output with many false positives.

The second method took a more holistic approach by combining a large amount of (presumably) independent data sources in a statistical way to distinguish candidates for further investigation from a long list of genes. The method solved many issues inherent in working with biological data and proved very successful in detecting genes involved in both pathways and diseases.

The next (and last) chapter will discuss why and how most of the methods described in this thesis (including the ones discussed in this chapter) were implemented as web services to allow flexible use of them by the community. Because standard libraries and Application Programming Interfaces (APIs), as well as established web standards, are used, other scientists can easily contribute to or expand the existing implementations. This contributes to building a vast research network and will enhance research in this area.

Chapter 5

Web services integration

BIOINFORMATICS is a multi-disciplinary field of research on the edge of biology, mathematics, statistics, and computer sciences. As discussed before, bioinformatics plays an increasingly important role in providing access to biological data and computational infrastructure, especially with the current accumulation rate of life sciences data. Efficient retrieval and integration of data from a multitude of complementary data sources becomes the most important bottleneck in speeding up the knowledge acquisition cycle (see Figure 5.1).

As outlined by Lincoln Stein [123], the field of bioinformatics faces the challenge of aggregating data from many sources, each with its own data representations and access methods. An elegant solution lies in the adoption of emerging web services technologies in bioinformatics. Web services allow a uniform and flexible way to access biological data resources via the Internet. Since the access method is uniform, it is much more easy to incorporate methods for automated data retrieval in bioinformatics tools. This relieves researchers from the burden to manually fetch data from a web site. And it assures working with the most up-to-date biological data that is centrally stored and managed.

In Section 5.1 a working definition of a *web service* is given, along with a description of the emerging technologies enabling the use of web services via the Internet. Section 5.2 gives an overview of bioinformatics projects that embraced the web services technologies to create an *e-Science* platform. Following up on this emerging trend, several methods described in this thesis were implemented and deployed as web services. Some of these services were then bundled in a web-based pipeline for the analysis of microarray data and regulatory sequences. Others were bundled through development

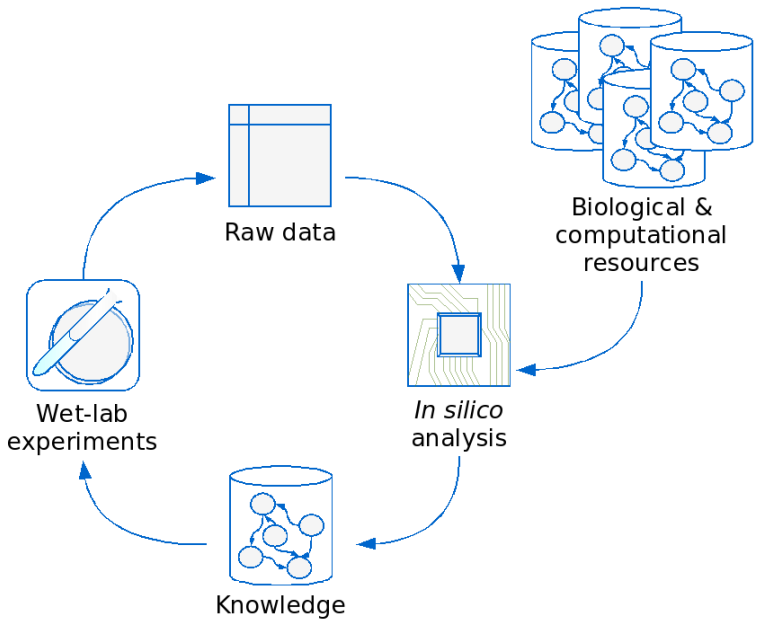


Figure 5.1: Knowledge acquisition through integration of data sources. As the amount of biological data available on the Internet expands rapidly, the most important bottleneck for performing biological research becomes efficient retrieval and integration of heterogeneous data sources. Web services technologies can help to provide a uniform access to these data sources, thus supporting the process of automated retrieval and speeding up knowledge acquisition.

of thin stand-alone Java clients. This is elaborated on in Section 5.3.

5.1 Web services technologies

Web services technologies enable the next generation Internet and address the increasing importance of application to application communication. Web services offer a standard way of communication between software applications regardless of the platform on which they are deployed.

A *web service* is defined by the World Wide Web Consortium (W3C) [153] as "a software system designed to support interoperable machine-to-machine interaction over a network." They further narrow down their definition of a web service to a published interface described in a machine-processable format (like the Web Service Definition Language (WSDL)) that interacts with other software components via the Internet using Simple Object Access Protocol (SOAP) messages.

5.1.1 The web services architecture

The Web Services Architecture (WSA) [20] is a conceptual model describing the different web service components and the way they interact. The three main elements are a provider agent, a requester agent, and a discovery service (Figure 5.2). The provider agent is a web service exposed on the Internet by a provider entity (a person or organization) that can be invoked by a requester agent (a piece of *client* software). In case of automated discovery, discovery services can be used to help the requester entity or agent find the web service it needs. They can be regarded as the yellow pages of the Internet.

5.1.2 SOAP and WSDL

Several standards were developed based on the WSA conceptual model. SOAP and WSDL are the most important ones. They are both recommendations of the W3C, an international consortium that publishes protocols and guidelines to ensure long-term growth of the World Wide Web. A recommendation is a set of consensus guidelines endorsed and advocated by the W3C. A more detailed description of SOAP and WSDL follows.

SOAP stands for Simple Object Access Protocol. It is an internet protocol based on the eXtensible Markup Language (XML) and is used for the standard communication with and between web services. Being written in XML, SOAP consist of marked-up entities. The markup encodes the logical

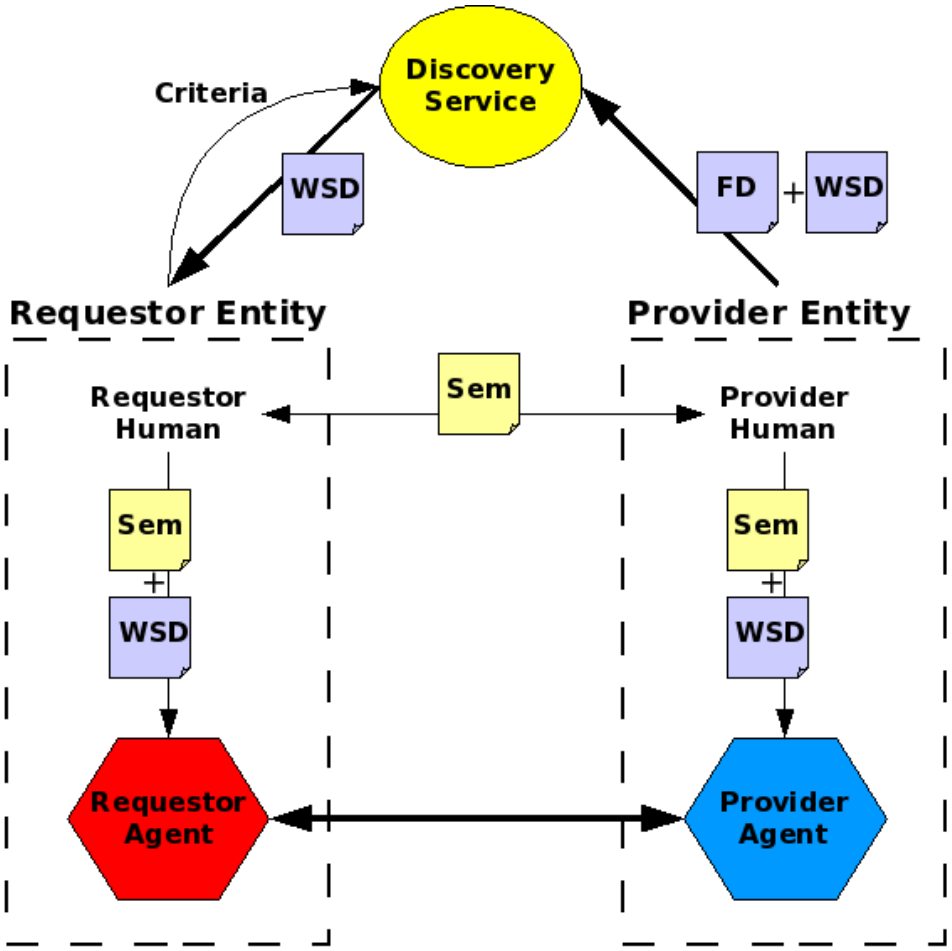


Figure 5.2: The general process of engaging a web service. A requester entity finds a provider entity that offers a particular service through the use of a discovery service and based on the service's functional description (FD) and web service description (WSD). After the requester and provider agree upon a service's semantics (Sem) and description (or the requester settles with the semantics and description offered by the provider), the parties can adapt the agents at their institution. Both semantics and service description will define the interaction between a requester agent and a provider agent. From now on, messages can be exchanged between the agents. Adapted from Booth *et al.* [20].

structure of the SOAP document and consists of *tags* that defines the entities. Every entity is delineated by an opening and a closing tag, making XML documents easily processable by a computer, yet still readable by a human being. An example of a SOAP message is depicted in Table 5.1.

Table 5.1: Example SOAP message. The table shows an example of a typical SOAP message envelope. SOAP is an XML-based protocol consisting of *tags* that delineate entities in the message. This message is sent to invoke a web service that checks if the *biovec.BlastModel* service is available.

```
<soapenv:Envelope xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
        <soapenv:Body>
            <isAvailable
                soapenv:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
                <serviceName xsi:type="xsd:string">
                    biovec.BlastModel
                </serviceName>
            </isAvailable>
        </soapenv:Body>
</soapenv:Envelope>
```

Just like the famous HyperText Markup Language (HTML) (i.e., the language every single web page is coded in), SOAP messages are transferred from one computer to another via the HyperText Transfer Protocol (HTTP), the protocol that enables browsing the Internet. As a matter of fact, pointing a browser to the web address of a web service, invokes the service. A browser does not know how to construct a SOAP message, but the web service will answer by returning its web service description on how it should be invoked properly.

The Web Services Description Language (WSDL) is an XML-based language used to describe how web services should be invoked and what parameters are necessary. It defines a web service as a collection of communication endpoints capable of exchanging messages and provides information that can be used to automate communication between services and applications. Table 5.2 gives an example WSDL file of some Endeavour web services (see Section 5.3.4). Figure 5.3 shows how these standards relate to each other

in what is called the *Web Services Architecture Stack*.

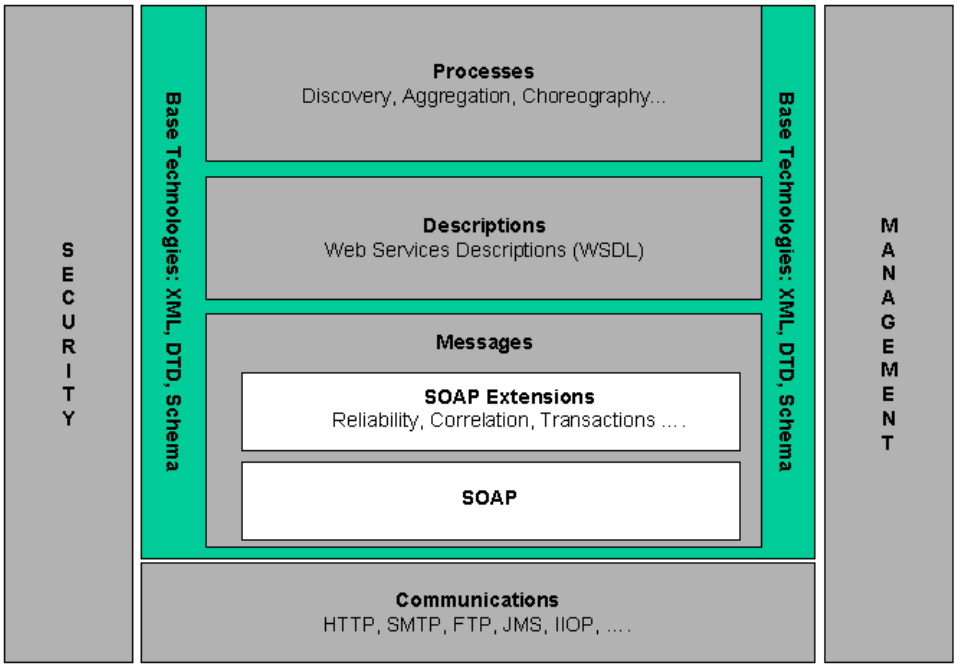


Figure 5.3: Web Services Architecture Stack. XML: eXtended Markup Language; DTD: Document Type Definitions; WSDL: Web Service Definition Language; SOAP: Simple Object Access Protocol; HTTP: HyperText Transfer Protocol; SMTP: Simple Mail Transfer Protocol; FTP: File Transfer Protocol; JMS: Java Message Service; IIOP: Internet Inter-Object Request Broker Protocol. Taken from Booth *et al.* [20].

5.2 Bioinformatics and web services

Two large projects have arisen that embraced the web services technologies and implemented an experimental bioinformatics services platform: BioMOBY and ^{my}Grid.

5.2.1 BioMOBY

The BioMOBY [149, 152] project's aim is to create a simple and extensible platform for discovery, representation, integration, and retrieval of heterogeneous and dispersed biological data sources. They are investigating

Table 5.2: Example WSDL file. The table shows a part of the WSDL file describing the methods that can be invoked on some of the Endeavour web services. The WSDL file gives a list of endpoints, like *isAvailableRequest* or *executeQueryRequest*, and the necessary parameters and parameter types. It also specifies the response type of every endpoint, for example a boolean (*xsd:boolean*) in the case of *isAvailableResponse* and a string (*xsd:string*) in the case of *executeQueryResponse*. The service *isAvailableRequest* can be invoked with the SOAP message in 5.1. The service *executeQueryRequest* executes an SQL query on an internal database server.

```
<wsdl:definitions
    targetNamespace="http://aulne8.esat.kuleuven.be/axis/SOAPService.jws">
        <wsdl:message name="isAvailableRequest">
            <wsdl:part name="serviceName" type="xsd:string"/>
        </wsdl:message>
        <wsdl:message name="isAvailableResponse">
            <wsdl:part name="isAvailableReturn" type="xsd:boolean"/>
        </wsdl:message>
        <wsdl:message name="executeQueryRequest">
            <wsdl:part name="dbDriver" type="xsd:string"/>
            <wsdl:part name="dbURL" type="xsd:string"/>
            <wsdl:part name="dbUser" type="xsd:string"/>
            <wsdl:part name="dbPassword" type="xsd:string"/>
            <wsdl:part name="dbQuery" type="xsd:string"/>
        </wsdl:message>
        <wsdl:message name="executeQueryResponse">
            <wsdl:part name="executeQueryReturn" type="xsd:string"/>
        </wsdl:message>
        <wsdl:service name="SOAPServiceService">
            <wsdl:port binding="impl:SOAPServiceSoapBinding" name="SOAPService">
                <wsdlsoap:address
                    location="http://aulne8.esat.kuleuven.be/axis/SOAPService.jws"/>
            </wsdl:port>
        </wsdl:service>
</wsdl:definitions>
```

two different approaches: a semantic-web approach called *S-MOBY* and a web-services approach called *MOBY-S*. In the former the emphasis is on semantics in a decentralized environment where services are discovered through semantically-aware, Google-like engines. The latter explores the use of web registries to describe and centralize web services, thus enforcing more stable and reliable service interfaces. The project acknowledges that both approaches have certain disadvantages. Therefore, its ultimate goal is combining the advantages of both.

5.2.2 ^{my}Grid

Grid computing is the simultaneous use of the resources of many computers in a network to solve a single problem. Grids can provide the computational power to solve complex scientific problems. The information-rich character of life sciences research poses new challenges to grid computing. The ^{my}Grid project [124, 43] focuses on exploiting grid and web services technologies to automate and facilitate complex analyses of biological data by creating an *information grid*. The semantically enabled ^{my}Grid middleware framework delivers computational power and data management infrastructure (data storage, workflow enactment, change event notification, resource discovery, and provenance management) to support, as they call it, the *e-Scientist* or, in the context of this thesis, the *e-Bioinformatician*. The services in the project can be divided in three categories:

- Services for forming experiments. This category comprises typical bioinformatics services like BLAST, the EMBOSS suite of applications, queries to MEDLINE, SRS, etc.

- Services for discovery and meta-data management. Services of this category discover bioinformatics services based on the semantic descriptions of their inputs, outputs, tasks they perform, and used resources.

- Services for supporting e-Science. This category contains services that help e-scientists to create complex workflows with bioinformatics services and resources and to execute them correctly.

The service description part of this project is being developed in collaboration with the BioMOBY project described above.

5.3 Web services integration

With Stein's philosophy in mind and following in on large projects like BioMOBY and ^{my}Grid, a number of methods described in this thesis (among others) were implemented as SOAP web services. This came to benefit the general accessibility of these implementations to a wider public of life sciences researchers. Some of the services are now heavily used and even incorporated in third party tools like SeqVISTA [66].

The web services that were developed in the framework of this thesis are part of three distinct software projects: INCLUSive, Toucan, and Endeavour. The INCLUSive project contains a set of interconnected web services that can be invoked from several web pages or via SOAP messages. The project aims at providing a flexible and extendable platform for the analysis of gene expression data. Toucan is a thin-client workbench for regulatory sequence analysis of metazoan genomes that makes intensive use of web services. In the framework of this project several web services were implemented on top of in-house developed algorithms for motif and module detection and discovery, as well as third-party algorithms for comparative genomics. The methodology for prioritization of pathway and disease genes described in Chapter 4, was also implemented as a thin-client software tool called Endeavour. For training the submodels and scoring the list of candidate genes, Endeavour invokes web services to perform the computationally intensive model building and to fetch information from internal and external databases.

5.3.1 Computing architecture and technicalities

All services were implemented using Java 5 (Sun Microsystems) and deployed on an Apache Axis 1.2 SOAP server running in an Apache Jakarta Tomcat 5 Servlet/JSP Container. Apache Axis is an implementation of the W3C SOAP recommendation (see Section 5.1) by the Apache Software Foundation [44]. It wraps the processes of creating SOAP messages and interpreting web service requests and responses in high-level Java objects and automates web service deployment. Apache Tomcat [139] is a Java Servlet Container and is the official reference implementation of Sun Microsystem's Java Servlet and JavaServer Pages technologies. The reader is referred to Sun Microsystem's web site at `http://java.sun.com` for more information about these technologies.

Figure 5.4 visualizes the web services architecture that was implemented in the framework of this thesis. Incoming SOAP messages are interpreted by

Apache Axis and redirected to a number cruncher server via Java Remote Method Invocation (RMI). RMI enables Java-based applications to access Java objects running on a remote location (i.e., on another computer). With RMI all computationally intensive tasks can be performed on a fast computer or cluster of computers. This way, the web services return the results of a request or analysis faster, while not overloading the web server.

5.3.2 INCLUSive

Unraveling transcriptional regulation from microarray data raises a double challenge. The first challenge is to cluster genes into biologically meaningful groups; the second is to find similar regulatory motifs (mostly transcription factor binding sites (TFBS)) in the promotor regions of the genes in such groups, the latter hopefully explaining the former.

INCLUSive comprises a set of algorithms and tools necessary to perform such analysis. They fall into three broad categories: Preprocessing, Cluster Analysis, and Biological Validation. Each of these categories is defined by the data it requires as input and output. The analysis starts with the raw expression data from a microarray experiment. After preprocessing, the genes whose expression was measured in the microarray experiment are grouped by clustering their gene expression profiles. The groups are then checked for consistency with established information about (regulation of) biological processes. The INCLUSive suite was published by the author in Nucleic Acids Research [28].

As compared to a previous release of INCLUSive [137], substantial improvements were made. Modules for normalization of microarray data and refinement and validation of clusters were added and the existing modules underwent a reconstruction to improve and broaden functionality. Also, a more organism-oriented approach is promoted to improve retrieval of intergenic sequences and functional information.

As compared to a similar pipeline called Expression Profiler [144], the major advantage of INCLUSive is its loosely coupled structure. All tools can be used separately as well as in a complex sequence of analysis steps. The web services architecture allows easy integration of new or alternative tools, which makes the system dynamic and flexible.

INCLUSive web services

All algorithms in the INCLUSive suite were implemented as web services. They are described in WSDL files and can be invoked via standard SOAP

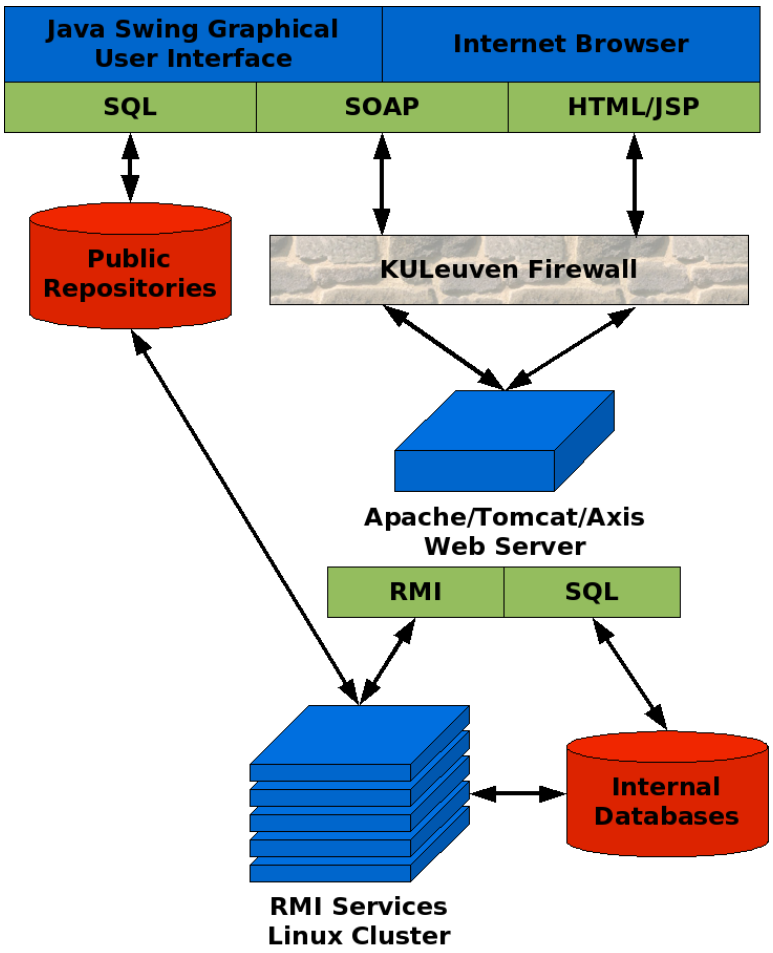


Figure 5.4: Computing architecture of Bioi@SCD. This figure visualizes the different elements of the web services computing architecture of the research group Bioi@SCD and the connections between them. Incoming SOAP messages are interpreted by an Apache Axis SOAP server, running in an Apache Tomcat servlet container, and the requests are dispatched to a number cruncher server via Remote Method Invocation (RMI). The invoked code on the number cruncher then performs the requested tasks, which may involve execution of SQL queries on databases, command line scripts (Unix shell scripts, C^{++}, perl, and so on), Matlab scripts (MathWorks), R scripts (r-project.org), etc.

messages. This way platform interoperability is assured and an easy integration with other software components can be realized. This section gives an overview description of all INCLUSive's components. Most of them can be used in three different ways:

1. By using the forms on the different web pages.
2. By invoking their web services from a remote computer via SOAP.
3. By downloading and installing the stand-alone versions of the algorithms.

Figure 5.5 shows the flowchart of the INCLUSive portal. It schematizes how all modules are connected and visualizes the data flow between them. A detailed description of the different modules follows.

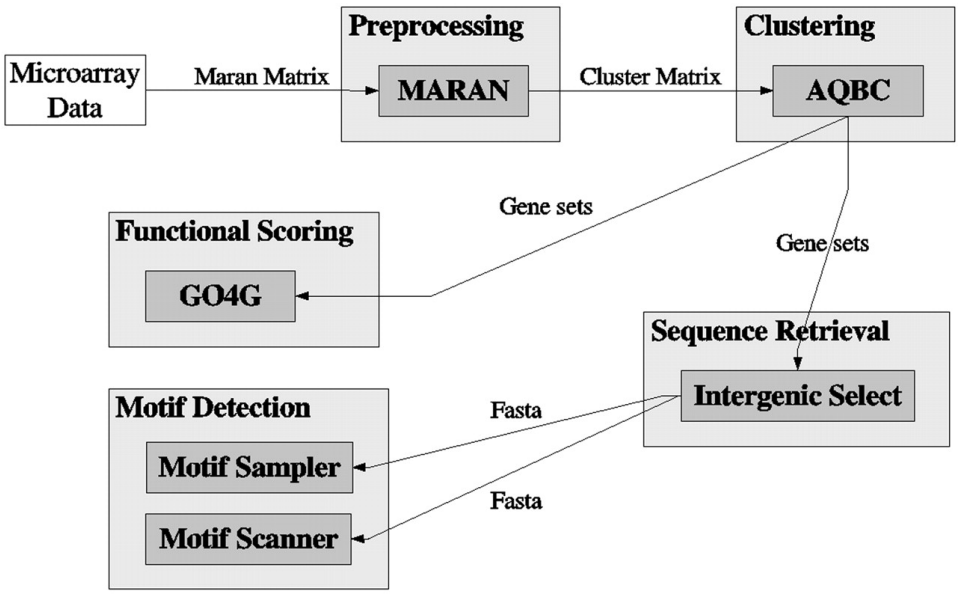


Figure 5.5: Schematic overview of the data flow between the different modules of INCLUSive. The flow supports complex analysis of microarray data, comprising ANOVA normalization, filtering and clustering, functional scoring of gene clusters, sequence retrieval, and detection of known and unknown regulatory elements. All modules can be used independently of each other or in a pipeline.

Maran: normalizing microarray data Maran normalizes gene expression data by constructing a generic ANOVA model based upon several

sources of variation in the experiment [39]. The residuals obtained from fitting the model can be used for statistical inference. Other features of the Maran web application are a Loess fit procedure and an option for detecting genes that have a significantly changing expression profile.

Adaptive quality-based clustering of microarray data The adaptive quality-based clustering (AQBC) method clusters microarray data in an heuristic iterative two-step process [121]. One of the characteristics of the algorithm is that all clusters have a constant quality, represented by the significance level S. The default value for S guarantees that a gene has a probability of 95% to truly belong to the cluster it is assigned to, according to a probabilistic model of the data. As a consequence, the clusters found by AQBC contain few false positives and are thus ideal seeds for further *cis*-regulatory analysis.

Information Select: retrieving additional information Central to Information Select is a series of organism-specific knowledge bases. At the moment, they mainly contain mappings of different public database identifiers. These knowledge bases allow the Information Select algorithm to fully characterize the genes of each cluster by providing links to a myriad of different public databases, starting from GenBank or Unigene accession numbers. Based on the organism that is specified by the user, the algorithm addresses the correct knowledge base and fetches and returns links to additional information sources, such as Entrez Gene, Ensembl, GeneCards, MGI, SWISS-PROT, and so on.

Functional scoring with GO4G GO4G is designed to assign general functional trends to groups of genes. The algorithm extracts GO terms associated to a group of genes (a cluster, for example) and calculates which terms are statistically over-represented when compared to their expected frequencies. This method was described previously in Chapter 3.

Intergenic Select: retrieving intergenic sequences Since sequence retrieval for organisms with fully sequenced genomes is well supported by other systems, either Toucan [7, 5] or EnsMart [73, 85] must be used for the selection of upstream or intergenic sequences. Both systems allow for selection of exactly the genomic regions of interest.

The alternative Intergenic Select service provided in INCLUSive is based on an iterative BLAST-search against GenBank (NCBI). The

tool accepts accession numbers and gene names to retrieve seed sequences. The BLAST hits of these sequences are used in the subsequent steps, until the required length of sequence upstream of the coding region is reached. This approach is useful when working with compact genomes [136, 84].

MotifSampler: finding over-represented motifs MotifSampler implements a user-friendly Gibbs sampling procedure that allows detection of statistically over-represented motifs in a set of unaligned sequences [135, 136, 84]. The algorithm determines in which sequences and at what positions a statistically over-represented motif is present, compared to a background model derived from the input data or compared to a user-specified organism-dependent background model.

MotifScanner: screening for known motifs MotifScanner is designed to search for putative sites of known motifs from TRANSFAC [150] and PlantCARE [78] in a set of sequences [7]. The motifs, represented by a position probability matrix, are assumed to be hidden in a noisy background sequence, represented by a higher-order Markov model. The algorithm is based upon the core modules of the MotifSampler.

5.3.3 Toucan

Toucan [7, 5] is a stand-alone Java application for the detection of *cis*-regulatory elements in promoter regions of higher eukaryotes. It uses and integrates several INCLUSive services and shows clearly the advantages of working with web services. Whatever version of Toucan is used, the invoked service always runs the latest version of the algorithm. The processor usage of the computer running Toucan is kept low because heavy calculations are performed remotely on a Linux cluster. Also, the total file size of the application is kept low, which improves download times. Besides that, Toucan can be started via Java Web Start (JWS) technology through a simple click, sparing the user of any form of installation procedure. Moreover, use of the JWS version assures the user is working with the most up-to-date version of the client software.

The Toucan software is mainly developed and maintained by Stein Aerts. The reader is referred to the developer's P.h.D. thesis [2] for a more in-depth description of the software architecture and functionalities of Toucan. More information can also be found at `http://homes.esat.kuleuven.be/~saerts/software/toucan.php`.

5.3.4 Endeavour

Endeavour [3] is stand-alone Java application for the computational prioritization of large lists of candidate disease or pathway genes. It implements the prioritization methodology described in Chapter 4. Like Toucan, Endeavour communicates with server-side web services using SOAP to perform computationally intensive tasks. This way, heavy calculations are performed on suitable computers. Endeavour is also packaged as a convenient Java Web Start bundle, which makes even frequent updates of the client and server-side code invisible to the user. The architecture and web services framework of the Endeavour software was developed entirely by the author. Further development is done by Leon-Charles Tranchevent.

Endeavour is the only software to date that combines a freely accessible, interactive and flexible tool with an unbiased prioritization of the entire genome by consulting multiple data resources. As such, Endeavour is superior to the existing publicly available prioritization tools. Indeed, *GeneSeeker gathers and combines expression and phenotypic data from several web-based databases for genes located in a certain chromosomal region* (as quoted in the article abstract from GeneSeeker [142]). The information provided by GeneSeeker is rather limitied as it only indicates whether a particular gene is expressed in a certain tissue or linked to a disease phenotype. GeneSeeker has been poorly validated (only on 10 syndromatic disorders). Importantly, however, GeneSeeker does not prioritize genes at all, but simply lists them. The second software package, G2D [100], is *a scoring system that relates the functional annotation of human genes to genetically inherited diseases mapped onto chromosomal regions.* G2D uses Gene Ontology and textual data banks and can only make predictions for annotated or known genes and thus has a strong bias towards better-studied genes. The authors even acknowledge that *the analysis of recently dissolved disease genes with G2D is difficult.* Thus, G2D only works for well-established disease genes, most likely because the function of a disease gene becomes well annotated after it has been identified as a disease gene and subsequently been studied in model organisms. G2D is also not as flexible as Endeavour: the G2D software package is an automatic and non-interactive tool, which offers the user the limited possibilities of only defining the name of the disease (or its OMIM ID) and the chromosomal region of interest. None of the other prioritization methods described in the literature are publicly available as a software tool (check Section 4.2 for an overview). Tables 5.3 and 5.4 give a comparative overview of the different software tools and methods that are used to prioritize genes.

Table 5.3: Overview of the methods used for the prioritization of candidate disease genes. For most of the published methods, there is no publicly available software tool. Endeavour is the only application that combines data from multiple data resources and allows users to include their own data sets in the prioritization. Note: the methods by Lopez-Bigas [79] and Adie [1] were not incorporated in this table because these have a different scope. They provide a general disease probability, and do not prioritize genes according to a particular disease. Continued in Table 5.4

Reference	**Available software tool**	**Data sources accessed by method**	**Number of genes tested in large-scale validation**	**Number of genes tested in case studies**
Van Driel [142]	GeneSeeker	Expression, phenotypes	-	10
Perez-Iratxeta [100]	G2D	Gene Ontology, Text	100	10
Freudenberg [47]	not available	Gene Ontology	878	-
Turner [141]	not available	Gene Ontology, InterPro	163	1
Tiffin [138]	not available	Text, Expression	17 + 20	-
Hauser [60]	not available	Expression	-	1
Aerts [3]	Endeavour	Gene Ontology, Text, InterPro, BLAST, Microarray and EST expression, Transcription, Pathway	627	16

Table 5.4: Overview of the methods used for the prioritization of candidate disease genes. Continued from Table 5.3.

Reference	**Functional validation of the software**	**Performance score: sensitivity / specificity**	**Useful for unknown genes**	**General comments**
Van Driel [142]	No	Not stated	Yes	Does not prioritize genes
Perez-Iratxeta [100]	No	50/90	No	
Freudenberg [47]	No	67/85	No	
Turner [141]	No	37/98	No	
Tiffin [138]	No	88/37	Yes	
Hauser [60]	No	Not stated	Yes	
Aerts [3]	Yes (Ypel1 knockdown in zebrafish)	70/90, 40/98, 90/60	Yes	Modular, extendable, also works for pathways

Endeavour architecture

The Endeavour software was developed as a stand-alone Java application for two reasons: firstly, because a stand-alone application provides a richer interaction with the user than a web-based application; secondly, because this allows users to include their own, locally stored data sets in the prioritizations without having to upload the data to the Bioi@SCD servers. The use of Java Web Start and web services technologies offers several advantages with respect to propagating software updates, redirecting computationally intensive tasks, and fetching data from remote databases. Therefore, Endeavour was designed as a thin Java client that communicates intensively with server-side databases and computational infrastructure.

The Endeavour Java client has an object-oriented architecture in two separate layers. One layer models the most important concepts Endeavour works with. It contains all code for manipulation of the objects represented by the concepts, and for retrieval of data about the objects from various data sources. This conceptual layer is designed around the concept of a *BioEntity*, that represents a gene, and a *BioEntitySet*, that represents a group of genes. Both concepts are central to the prioritization methodology described earlier. A more detailed description of the Endeavour data model can be found in the next section. The two most important data sources Endeavour connects to are Ensembl [73] on the one hand, and the database server of Bioi@SCD on the other. To fetch data from Ensembl, direct database calls are executed on Ensembl's database servers. All communication between Endeavour and the database server of Bioi@SCD goes via of web services. The software regularly checks availability of the data sources and uses alternative mirrors if necessary.

The second layer contains all code of the graphical user interface (GUI). The GUI's responsibility is interaction with the user and visualization of the objects and results Endeavour generates. Figure 5.6 shows a screenshot of the Endeavour GUI. It tries to represent the different concepts of the Endeavour data model as clearly as possible. It contains three important panes: the *Model* pane, the *Data* pane, and the *Status* pane. The *Model* pane contains all information about the model a user is building and allows to add and remove data sources from this model (and the analysis). The *Data* pane contains more specific information about the training genes a model is based on, and the candidate genes that are analyzed (i.e., prioritized). After the analysis, this pane also visualizes the results of the prioritization in a spreadsheet with data and an easy to interpret *sprint plot* view. The *Status* pane informs the user about the interaction with web services and

the status of the analysis.

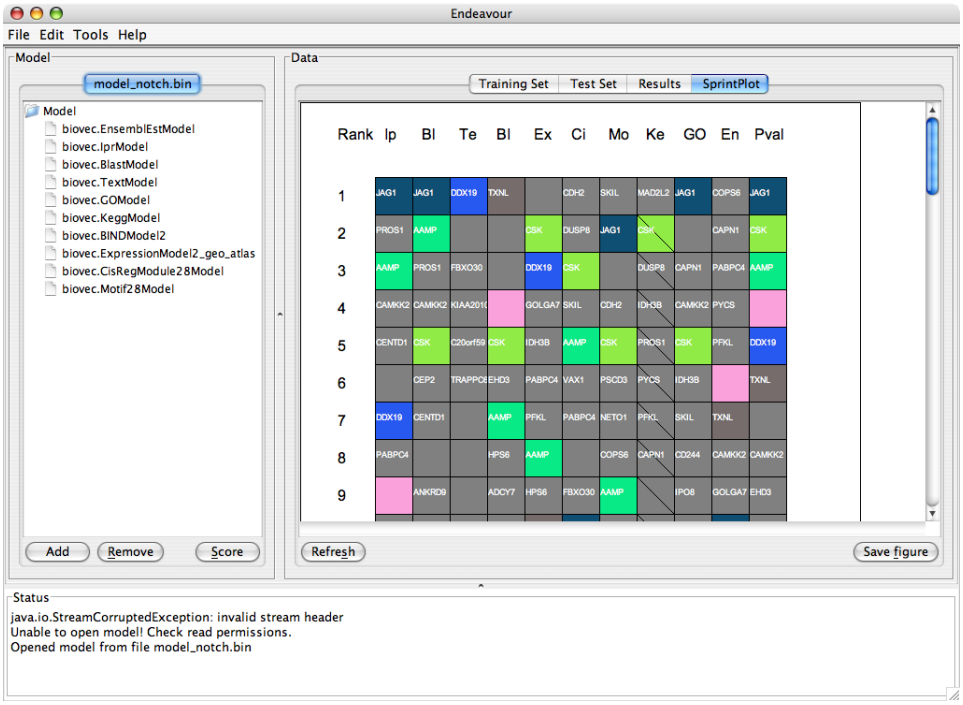


Figure 5.6: Screenshot of the Endeavour application. Endeavour prioritizes genes according to a myriad of different information sources. The results are presented in a graphical *sprint plot* visualization.

The decision was taken to provide the users of Endeavour with a maximal control over the set of training and test genes, as well as over the data sources to include in the prioritization. This idea was conceived from a prospective discussion with many geneticists and biologists, who do not use the existing prioritization methods for their lack of flexibility. This is perhaps best illustrated by the fact that not a single paper has been published reporting the identification of a novel disease gene when using any of the pre-existing methods. Most likely, this relates to the reality that geneticists and biologists, as opposed to bioinformaticians, prefer to have the flexibility to interactively select their own set of genes and the information they want to work with, above an automatic and non-interactive data mining selection procedure of disease characteristics.

Endeavour data model

The central object in the Endeavour data model is a *BioEntity*. This object represents a biological entity and all information known about it. In most of the cases, this entity will be a gene. Biological entities are combined in sets (*BioEntitySet*). A training set is a *BioEntitySet* that is used to build a model for a process or disease, represented by the *Model* object. A *Model* consists of several *SubModel* objects that each represent a certain data source. Building a *SubModel* means fetching and summarizing all information about the genes in the training set for one particular data source.

Endeavour comes with a set of standard submodels that summarize the following information about the user-specified training genes: KEGG pathway membership, Gene Ontology (GO) annotations, textual descriptions from MEDLINE abstracts, microarray gene expression, EST-based anatomical expression, InterPro's protein domain annotation, BIND protein interaction data, *cis*-regulatory elements, and BLAST sequence similarity. Besides these default information models, users can add their own microarray data or custom prioritizations as submodels.

Apart from the *BioEntitySet* that contains the training genes, there is a second *BioEntitySet* that holds the candidate genes and all their related information. These candidate genes are prioritized during a process called *scoring*. Scoring involves comparing the information of a candidate gene with the information in the *Model* object for every data source. Based on these comparisons, every candidate gene receives a ranking and the prioritization is complete.

Endeavour web services

As said before, Endeavour uses web services to fetch data and dispatch computationally intensive tasks. Endeavour makes use of two types of web services: services for generating submodels and services for scoring candidate genes. In the case of building a submodel, the gene identifiers of the training genes are sent to the Axis server, together with some user-specified or default parameters. The server then dispatches the request via RMI to a computer cluster that builds the requested submodel (see Figure 5.4). The created submodels are sent back to Endeavour in an asynchronous way. While Endeavour is building and receiving all requested submodels, the user is free to load and tweak the set of candidates, view gene and submodel information, or investigate previous prioritization results. Services for scoring candidate genes are only used if a submodel is too large to be transferred

to the client (as with the BLAST sequence similarity), or if the scoring process is computationally intensive (as with the identification of *cis*-regulatory modules).

The data the web services need, is fetched from pre-build database tables in the Bioi@SCD database server. The data in the Endeavour database tables is compiled from various data sources (Ensembl, InterPro, Kegg, GO, etc.). When the original data sources are updated, new database tables are compiled. Scripts were written to automate this updating process. In most cases, updating has no consequences for the Endeavour client. Endeavour also depends on information in the Ensembl database. A new version of the Ensembl database is released every two months. With most of the data updates, the Endeavour client's code does not need to be updated. If structural changes were made to the Ensembl database, Endeavour keeps using an older release of Ensembl by default until a patch is released and disseminated using Java Web Start.

To assure maximal uptime and spread usage load of the Endeavour services, both the SOAP and RMI services are mirrored on different servers. The client always checks the availability of the services at startup. If a service is unavailable, the client automatically checks other Axis mirrors. If no backup service was found the user is unable to add the respective submodel to the analysis.

It is hoped that Endeavour's open architecture will stimulate other groups to provide to the community custom web services for training submodels and prioritizing candidate genes.

5.4 Conclusion

The impact of the Internet on scientific research worldwide has been enormous, not the least in biological research. Especially the Human Genome Project was the inspiration for many biological databases publicly available via the Internet (check the Database Issue of Nucleic Acids Research [51] for a recent overview). As of now, conducting biological research without the Internet is nearly impossible. Many researchers depend on the Internet as the most important source of biological information. As the amount of available data increases at a rate never seen before, researchers are now faced with the problem of finding the information they need, in a format they can work with.

As presented in this chapter, emerging Internet technologies can help to realize a true *BioScape*, a subspace of the Internet devoted to perform-

ing biological research. Semantic web technologies, like ontologies, will enable fast, context-sensitive retrieval of biological data. Web services will allow extensive automatization of complex bioinformatics tasks and drive the standardization process. Grid computing will provide researchers with the necessary computational power to build increasingly more complex models of biological systems.

Chapter 6

Conclusions and prospects

BECAUSE of the very nature of biological data derived from wet-lab experiments and analyses, biology still has a limited mathematical foundation. Biology is still to a large extent knowledge rich and data poor, and from the field of artificial intelligence it is known that knowledge is a hard thing for a computer to work with. Yet, large-scale genome sequencing projects spurred high-throughput experimentation, which resulted in a massive amount of medium-quality data sets in urgent need of correct interpretation. Therefore, the kind of computational integration approaches described in this thesis become increasingly important. Combination of high-throughput data with biological knowledge, as well as integration of different data sources solves many problems inherent to biological research:

- Each biological data source is bound to a specific experimental technique. Each technique has a certain resolution, gives only information about a specific aspect of biology, and introduces technique-specific biases in the data. Data sets from different experimental techniques can complement each other. Combining them can increase the signal-over-noise ratio, thus helping to separate bias from biology.
- Biological variability (i.e., variability among individual cells, organisms, individuals) and the enormous flexibility of living organisms when subjected to experimentation make biological research inherently empirical. This stands in favor of a statistical approach to analyze biological data. Combining statistically-derived results from different data sources is important to gain trust in the outcome.
- A biological system is by definition an open system that operates out of (thermodynamic) equilibrium. Therefore, the only way to describe it

might be via a global approach, putting everything into context with everything. The combination of several heterogeneous data sources tries to model this holistic view.

Therefore, integration of biological knowledge into the analysis of high-throughput data has been the main focus of the work described in this thesis. Integration is possible throughout the knowledge acquisition cycle and at different levels of data interpretation, as discussed previously. Several methodologies were described to accomplish early (Chapter 2), intermediate (Chapter 3), and late (Chapter 4) integration. A major focus was on the use of textual data in the analysis of high-throughput data because of the vast and increasing amount of scientific knowledge in free text format.

In Section 6.1, the accomplishments of this work are highlighted. Section 6.2 describes possible extensions of the work in the thesis and how the described methodologies could be improved. Section 6.3 provides a general discussion on where this type of work is headed.

6.1 Accomplishments

The main accomplishments of the work described in this thesis fall into two categories: one with respect to transforming biological knowledge in computer-amenable representations, the other with respect to using these representations in the analysis of high-throughput data:

1. The challenge of working with high-level biological knowledge is to find a good computerized representation. A good representation is easy to compute with, yet does not obfuscate the important information or knowledge features. Two important sources of knowledge about genes were discussed in this thesis: textual information from MEDLINE abstracts and functional annotations from the Gene Ontology (GO) database:

    - **MEDLINE abstracts** of gene-function-oriented papers are a good source of concise functional information about genes. Explicit gene-document associations from Entrez Gene are used to select relevant papers and construct textual profiles of genes. The vector-space model (see Chapter 1) provides an excellent framework to work with these gene representations and integrate them in analyses of high-throughput data, as illustrated throughout this thesis. The use of domain vocabularies proved useful to limit the profiles to the scope of interest.

- **GO annotations** capture information about the molecular functions, biological processes, and cellular components associated with genes or gene products in a uniform controlled vocabulary. The hierarchical structure of the GO vocabularies and the fact that each term has a unique identifier already make this knowledge source easy to manipulate computer-wise. In this thesis, the aim has been to combine the annotations of individual genes into a functional representation of a gene group. It has been illustrated that this representation contains valuable and biologically relevant information about a gene group. And it can be used to integrate this information with other data sources, as shown in the case of late integration in Chapter 4. Moreover, several other data sources could be successfully represented in the same way (EST and InterPro data from Ensembl, and Kegg pathway information, for instance). This has made integration of these data sources even more straightforward.

2. Once a knowledge source is represented in a format adequate for computational analysis, it can be used to conduct specific bioinformatics tasks:

   - **Interpretation of analysis results** from high-throughput data involves putting the results into context with existing biological knowledge. Two different approaches were described in this thesis. In Chapter 2, early integration of textual data in the analysis of gene expression data was used to supervise the clustering of genes into biologically meaningful groups. The methods of intermediate integration described in Chapter 3 efficiently characterized gene groups and defined their functional coherence.
   - ***In silico* experiments** can be performed, as was illustrated in Chapter 2 with the large-scale gene clustering analysis based on textual data. As the amount of biological information in the public domain steadily increases, it is expected that more and more useful biology will lie hidden between the lines. The purpose of *in silico* experimentation is to discover this hidden knowledge. The gene co-linkage methodology described in Chapter 4 is an example of how indirect links between genes described in the literature can help to formulate interesting new hypotheses. In the same chapter, another approach is described that combines a multitude of complementary data sources . The method efficiently

and accurately prioritizes large lists of genes, to select only the most interesting ones for experimental validation.

- **Web services technologies** can help to provide efficient access to and automate retrieval of biological data, a topic that becomes the bottleneck of conducting biological research as the amount of publicly available data is exploding. In Chapter 5 a web services architecture has been proposed. Several, often heavily used software tools are successfully backed by this architecture.

6.2 Future work

Several methodologies described in this thesis lend themselves for further research. This section gives an overview of outstanding issues.

- The method to measure the functional coherence of a gene group based on the Gene Ontology (see Chapter 3), holds a lot of potential. The method's performance might be optimized at several levels of the implementation, but a benchmark study is necessary. Also, more semantical information about the GO annotations, or information from textual analyses, might be incorporated in the methodology.

- The use of domain vocabularies proved useful in textual profiling of groups of genes (see Chapter 3). Still, these vocabularies can be improved considerably. Also, the combination of indexing with domain vocabularies and Latent Semantic Indexing should be evaluated. Another extension might be the use of full-text articles in profiling genes, and automated retrieval of the most relevant paragraphs or sentences.

  The fact that the current approach only takes into account the set of PubMed abstracts explicitly linked to genes in Entrez Gene, is another drawback. The current efforts in detection of gene names might allow to link documents and genes on a much wider scale (the entire PubMed collection, for instance). Either way, good methods are needed to discern papers about gene function from more general or unrelated papers.

- Although the Endeavour methodology for prioritization of large groups of genes gives nice results (see Chapter 4), several extensions could be thought of. It might, for instance, be useful to weight data sources differently according to the trust an expert has in them. The current methodology uses properties of a set of training genes to generate a

model for a disease or process, but it might be interesting to add other sources of higher-level biological information, like descriptions of diseases from OMIM, or relevant publications about a pathway. Also, other approaches towards data fusion have to be evaluated. One interesting approach that is being investigated, is prioritization via classification with kernel methods.

Apart from the methodology, the representations of the different data sources can be improved. Each of the sources should be evaluated thoroughly and optimized with prioritization in mind. Also, their complementarity should be analyzed in more detail. An automated validation might be implemented to calculate the optimal combination of data sources for a certain training set.

A lot of functionality might be added to the Endeavour application: tools to automate the process of building a model, to perform statistical analysis of a training set or data source, to define a training set's coherence with respect to the different data sources, and so on. Development of a genome-wide scoring is already ongoing. Hopefully, the use of web services and its open architecture will encourage other researchers to contribute data sources and other functionalities to Endeavour.

6.3 Outlook

A key issue is to find the most practical way to combine different sources of biological information, at different levels of complexity and using robust statistical techniques. Several candidate approaches are described in the literature, especially emerging from the artificial intelligence community. Gaerdenfors and Williams [50] describe a method for modeling context-sensitive categorization based on algorithms from computational geometry and Region Connection Calculus[1]. They define a conceptual space as a geometrical structure based on attribute or quality dimensions. Concepts are regions in this conceptual space and contain objects characterized by a set of attributes or qualities. Reasoning about concepts involves determining relationships between these concepts. Therefore, they define a similarity measure derived from the distance between two points (or objects) in the conceptual space.

[1]The Region-Connection Calculus (RCC) is a well established formal system for qualitative spatial reasoning. It provides an axiomatization of space which takes regions as primitive, rather than as constructions from sets of points.

Using this measure, they perform a generalized Voronoi tessellation to divide the space into categories. Although this theory was never applied in the context of molecular biology, a straightforward translation can be made: the objects of interest are genes or proteins; their qualities are everything that is known about them, represented in discrete or continuous attributes; the conceptual space consists of gene/protein functional concepts. Depending on the scope of interest (and the attributes used), the categories could be pathways or diseases and the genes could be categorized by clustering them in this conceptual space. The only drawback might be the lack of robustness towards missing data.

A more statistical approach was taken by Friedman *et al.* [49]. They describe the use of a Probabilistic Relational Model (PRM), a standard relational model of which the instances have probabilistic attributes, besides the regular deterministic ones. PRMs can be considered an extension of Bayesian Networks with multiple interdependent objects. Because a conditional probability distribution is known (or predicted) for every probabilistic attribute, treatment of missing data can be facilitated. The advantage of working with PRMs is that a rich knowledge representation is kept, while the statistical backbone can be learned from various data sources and reveal context-specific relationships. Segal *et al.* present an interesting application of PRMs in clustering gene expression data [115]. Unlike standard clustering, their approach identifies similarities between genes over a multitude of richly structured data types.

It is hoped that these approaches can at least alleviate some of the problems inherent in working and reasoning with biological data and can help in building sound models of biological systems, thereby facilitating true systems biology in this century of the biological sciences.

Appendix A

Order statistics

Cumulative Probability Function

Given is a sequence of n independent and identically distributed random variables $X = (X_1, X_2, \ldots, X_n)$. An order statistic of order i is defined as the random variable $X_{n,i}$ that represents every i'th smallest element of all possible samples of X.

In the context of this thesis, the independent and identically distributed random variables $X = (X_1, X_2, \ldots, X_n)$ correspond to distributions of rank ratios $R = (R_1, R_2, \ldots, R_n)$. To obtain an ordered vector of rank ratios $r = (r_1, r_2, \ldots, r_n)$ with $r_1 \leq r_2 \leq \ldots \leq r_n$, a permutation P is applied such that $(r_1, r_2, \ldots, r_n) = P(R_1, R_2, \ldots, R_n)$. Note that there exist $n!$ different sets of rank ratios $(R_1, R_2, \ldots, R_n)$ that map to the same ordered vector $(r_1, r_2, \ldots, r_n)$.

The order statistic joint cumulative probability function defines the probability to obtain an observed (and ordered) vector of elements (in this case, rank ratios) by chance alone, by determining the number of ordered vectors of rank ratios not dominated by the given vector. A vector of ordered rank ratios $(r'_1, r'_2, \ldots, r'_n)$ is said to dominate an other ordered vector $(r_1, r_2, \ldots, r_n)$ if $r'_1 \geq r_1, r'_2 \geq r_2, \ldots, r'_n \geq r_n$.

Given a random vector of rank ratios R, with $r = P(R)$ the derived ordered vector, the measure of the set of unordered vectors that are not dominated by R under the null hypothesis of uniformly distributed ranks, is defined by this formula:

$$Q(r_1, r_2, \ldots, r_n) = n! \int_{s_0}^{r_1} \int_{s_1}^{r_2} \ldots \int_{s_{n-1}}^{r_n} ds_n ds_{n-1} \ldots ds_1 \qquad \text{(A.1)}$$

with $s_0 = 0$. The integral defines the area below the vector of ordered rank ratios in which all non-dominated vectors lie. The integral is multiplied by $n!$, because this is the number of different unordered sets of rank ratios that map to the same ordered set. Note the particular form of the integral where the integration variables are part of the integration bound of the next inner integral over s_{k+1}. For example, in the case $n = 2$ (i.e., in the case of two rank ratios), the integral can be worked out as follows:

$$Q(r_1, r_2) = 2 \int_0^{r_1} \int_{s_1}^{r_2} ds_2 ds_1 = 2 \int_0^{r_1} (r_2 - s_1) ds_1 = 2 r_2 r_1 - r_1^2. \qquad \text{(A.2)}$$

The surface this integral defines, is visualized in Figure A.1.

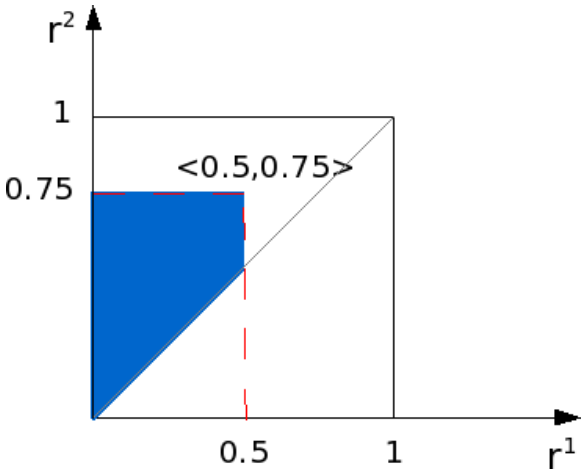


Figure A.1: The figure visualizes the area in which all vectors lie that are not dominated by a given vector $< 0.5, 0.75 >$. The lower the rank ratios in the ordered vector, the lower the number of dominated vectors and the lower the probability of drawing this vector by chance.

The solution of the integral can efficiently be computed in $O(n^2)$ with following recursive formula:

$$V_n(r_n, r_{n-1}, \ldots, r_1) = \sum_{i=1}^{n} (-1)^{i-1} \frac{V_{n-i}(r_n, \ldots, r_{i+1})}{i!} r_1^i \qquad \text{(A.3)}$$

This involves calculation of all subsequent V_k's for $k = 1, 2, \ldots, n$ with $V_0 = 1$ and

$$V_k(r_n, \ldots, r_{n-k+1}) = \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}(r_n, \ldots, r_{n-k+i+1})}{i!} r_{n-k+1}^i. \quad \text{(A.4)}$$

This appendix presents the proof that Equation A.1 and Equation A.3 result in the same solution or, more formally, that

$$Q(r_1, r_2, \ldots, r_n) = V_n(r_n, r_{n-1}, \ldots, r_1). \quad \text{(A.5)}$$

The proof is based on the fact that integral A.1 can be computed recursively for increasing n.

Definition

$$W_k(r_n, \ldots, r_{n-k+1}; s_{n-k}) = V_k(r_n, \ldots, r_{n-k+1}) - V_k(r_n, \ldots, r_{n-k+2}, s_{n-k}) \quad \text{(A.6)}$$

Notation

$$V_x = V_x(r_n, \ldots, r_{n-x+1})$$
$$W_x = W_x(r_n, \ldots, r_{n-x+1}; s_{n-x})$$

Lemma

Starting from $$V_k = \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} r_{n-k+1}^i \quad \text{(A.7)}$$

it can be proved that $$V_{k+1} = V_k r_{n-k} - \sum_{j=1}^{k} (-1)^{j+1} \frac{V_{k-j}}{j!} \frac{r_{n-k}^{j+1}}{j+1}. \quad \text{(A.8)}$$

The proof starts with the fact that A.6 can be expressed as an integral of the form

$$W_k = \int_{s_{n-k}}^{r_{n-k+1}} \int_{s_{n-k+1}}^{r_{n-k+2}} \ldots \int_{s_{n-1}}^{r_n} ds_n \ldots ds_{n-k+1}. \quad \text{(A.9)}$$

Note that s_{n-k} is a parameter, in contrast with $s_{n-k+1}, \ldots, s_{n-1}$. If W_{k+1} can be expressed as an integral of W_k, the statement A.5 can be derived by applying A.3:

$$
\begin{aligned}
W_{k+1} &= \int_{s_{n-k-1}}^{r_{n-k}} \int_{s_{n-k}}^{r_{n-k+1}} \dots \int_{s_{n-1}}^{r_n} ds_n \dots ds_{n-k+1} ds_{n-k} \\
&= \int_{s_{n-k-1}}^{r_{n-k}} W_k ds_{n-k} \\
&= \int_{s_{n-k-1}}^{r_{n-k}} \left(\sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} r_{n-k+1}^i - \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} s_{n-k}^i \right) ds_{n-k} \\
&= \int_{s_{n-k-1}}^{r_{n-k}} \left(\sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} (r_{n-k+1}^i - s_{n-k}^i) \right) ds_{n-k} \\
&= \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} \int_{s_{n-k-1}}^{r_{n-k}} (r_{n-k+1}^i - s_{n-k}^i) ds_{n-k} \\
&= \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} \left[r_{n-k+1}^i s_{n-k} - \frac{s_{n-k}^{i+1}}{i+1} \right]_{s_{n-k-1}}^{r_{n-k}} \\
&= \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} \left(r_{n-k+1}^i r_{n-k} - \frac{r_{n-k}^{i+1}}{i+1} - r_{n-k+1}^i s_{n-k-1} + \frac{s_{n-k-1}^{i+1}}{i+1} \right) \\
&= \left[V_k r_{n-k} - \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} \frac{r_{n-k}^{i+1}}{i+1} \right] - \left[V_k s_{n-k-1} - \sum_{i=1}^{k} (-1)^{i-1} \frac{V_{k-i}}{i!} \frac{s_{n-k-1}^{i+1}}{i+1} \right]
\end{aligned}
$$

Given also that, by definition,

$$
W_{k+1} = V_{k+1}(r_n, \dots, r_{n-k+1}, r_{n-k}) - V_{k+1}(r_n, \dots, r_{n-k+1}, s_{n-k-1}) \quad \text{(A.10)}
$$

and identifying the last two relations term by term, recursion A.4 is obtained. It is now easy to see that $Q(r_1, r_2, \dots, r_n)$ can be expressed in terms of V_k,

$$
\begin{aligned}
Q(r_1, r_2, \dots, r_n) &= W_n(r_n, r_{n-1}, \dots, r_1; 0) && \text{(A.11)} \\
&= V_n(r_n, r_{n-1}, \dots, r_1) - V_n(r_n, r_{n-1}, \dots, s_0), && \text{(A.12)}
\end{aligned}
$$

and since $s_0 = 0$ in the case of $Q(r_1, r_2, \dots, r_n)$ (see A.1), the last term in A.12 is zero and equality A.5 is proven.

From q-value to p-value

The q-values calculated with A.1 can not be used to perform a statistical significance test for two reasons. Firstly, the q-values are not uniformly distributed, which is a prerequisite of a correct p-value. Secondly, the calculation of the q-values depends on the number of ranks. This means that two q-values resulting from the joint cumulative distribution of two order statistics with different dimensionality cannot be compared directly. In the context of this thesis comparisons need to be made between genes with a varying number of rank ratios. For these two reasons, the q-values have to be transformed into p-values.

Figure A.2 shows the cumulative probability of the q-values for order statistics with different dimensionalities $n = 1, \ldots, 50$, based on 1,000 random samplings from the respective distributions.

The statistical distribution of the q-values for order statistics with different dimensionalities can be approximated with Gamma distributions. The maximum likelihood estimates of the parameters of the fitted Gamma distributions are plotted in Figure A.3.

The cumulative distribution functions of the Gamma distributions with given parameters give a good approximation of the sampled cumulative probability distributions. They provide us with a p-value for every q-value from the order statistics. Only for low values (lower than 10^{-3}), the two distributions deviate from each other. This results in p-values that are more conservative (i.e., higher than expected). Therefore, the p-values keep their statistical power and can indeed be used to define significance thresholds.

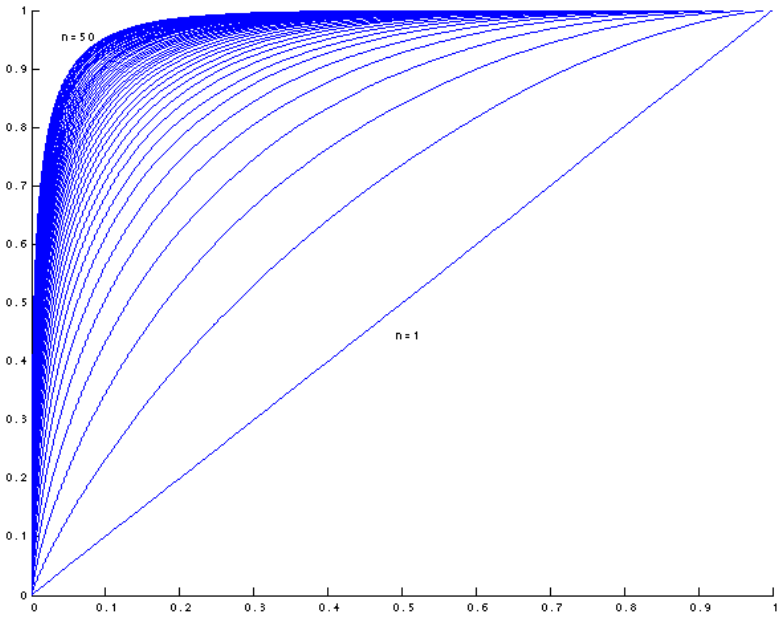
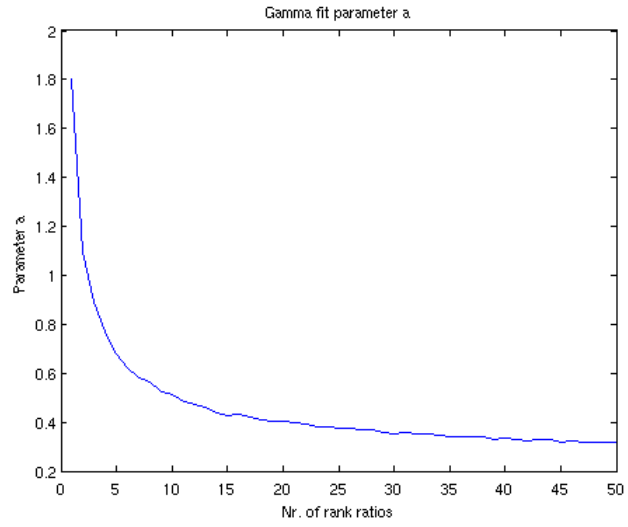
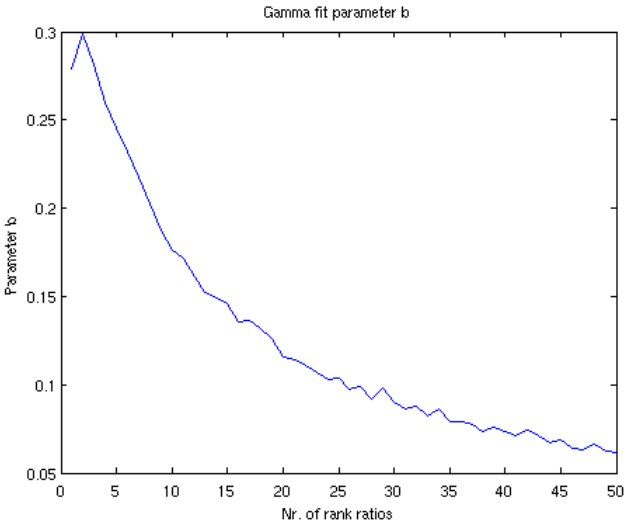


Figure A.2: Cumulative probability density function of the order statistics for $n = 1, \ldots, 50$ rank ratios. The plots are based on 1,000 random samplings for every value of n. Because the plots differ for different dimensionalities, the probability of drawing an ordered set of i rank ratios cannot be compared directly with the probability of drawing an ordered set of j rank ratios, with $i \neq j$. The use of a p-value makes such a comparison possible.



(a)



(b)

Figure A.3: Maximum likelihood estimates of the Gamma-distribution's parameters a and b for dimensionality $n = 1 \ldots 50$.

Appendix B

Supplementary material

Colon and colorectal cancer data

Table B.1 lists the gene identifiers of the colon and colorectal cancer genes that were textually profiled in Chapter 1 to show the usefulness of working with different vocabularies. The set of genes was derived from the Online Mendelian Inheritance in Man (OMIM) database by fetching all genes related to colon and colorectal cancer.

Statistically over-represented GO annotations

Tables B.2-B.13 list the top-25 statistically over-represented Gene Ontology (GO) annotations of the 15 gene groups defined in Chapter 2 via cluster analyses of expression and textual data, and the combination of both. The GO annotations were determined with the method described in Section 3.1.1.

Endeavour cross-validation data

For the cross-validation experiment, diseases in the Online Mendelian Inheritance in Man (OMIM) database to which at least 8 causative genes were assigned, were included. The following diseases and associated genes were used: Alzheimers disease, amyotrophic lateral sclerosis (ALS), anemia, breast cancer, cardiomyopathy, cataract, Charcot-marie-tooth disease, colorectal cancer, deafness, diabetes, dystonia, Ehlers-Danlos, epilepsy, hemolytic anemia, ichthyosis, leukemia, lymphoma, mental retardation, muscular dystrophy, myopathy, neuropathy, obesity, Parkinsons disease, retinitis pigmentosa, spastic paraplegia, spinocerebellar ataxia, Usher

Table B.1: Colon and colorectal cancer genes derived from the OMIM database.

HUGO	Alternate	Entrez Gene	Description
APC	GS, FPC	324	ADENOMATOUS POLYPOSIS COLI PROTEIN (APC PROTEIN).
AXIN2		8313	AXIN 2 (AXIS INHIBITION PROTEIN 2) (CONDUCTIN) (AXIN-LIKE PROTEIN) (AXIL).
BAX		581	BAX PROTEIN, CYTOPLASMIC ISOFORM DELTA.
BCL10		8915	B CELL LYMPHOMA/LEUKEMIA 10 (B-CELL CLL/LYMPHOMA 10) (BCL-10) (CED- 3/ICH-1 PRODOMAIN HOMOLOGOUS E10-LIKE REGULATOR) (CIPER) (CARD- CONTAINING MOLECULE ENHANCING NFKAPPAB) (CELLULAR HOMOLOG OF VCARMEN) (CCARMEN) (MAMMALIAN CARD-CONTAINING ADAPTER MOLECULE E10) (ME10) (CELLULAR-E10) (C-E10) (CARD-LIKE APOPTOTIC PROTEIN) (HCLAP).
BRAF		673	B-RAF PROTO-ONCOGENE SERINE/THREONINE-PROTEIN KINASE (EC 2.7.1.-) (P94) (V-RAF MURINE SARCOMA VIRAL ONCOGENE HOMOLOG B1).
BUB1		699	MITOTIC CHECKPOINT SERINE/THREONINE-PROTEIN KINASE BUB1 (EC 2.7.1.-) (HBUB1) (BUB1A).
CTNNB1		1499	BETA-CATENIN (PRO2286).
DCC		1630	TUMOR SUPPRESSOR PROTEIN DCC PRECURSOR (COLORECTAL CANCER SUPPRESSOR).
FGFR3	ACH	2261	FIBROBLAST GROWTH FACTOR RECEPTOR 3 PRECURSOR (EC 2.7.1.112) (FGFR-3).
KRAS2	RASK2	3845	TRANSFORMING PROTEIN P21A (K-RAS 2A) (KI-RAS) (C-K-RAS).
MCC		4163	COLORECTAL MUTANT CANCER PROTEIN (MCC PROTEIN).
MLH1	COCA2, HNPCC2	4292	DNA MISMATCH REPAIR PROTEIN MLH1 (MUTL PROTEIN HOMOLOG 1).
MLH3	HNPCC	27030	DNA MISMATCH REPAIR PROTEIN MLH3 (MUTL PROTEIN HOMOLOG 3).
MSH2	FCC1, COCA1, HNPCC1	4436	DNA MISMATCH REPAIR PROTEIN MSH2.
MSH6	GTBP, HNPCC5	2956	DNA MISMATCH REPAIR PROTEIN MSH6 (MUTS-ALPHA 160 KDA SUBUNIT) (G/T MISMATCH BINDING PROTEIN) (GTBP) (GTMBP) (P160).
NRAS		4893	TRANSFORMING PROTEIN N-RAS.
PDGFRL	PDGRL, PRLTS	5157	PLATELET-DERIVED GROWTH FACTOR RECEPTOR-LIKE PROTEIN; PLATELET-DERIVED GROWTH FACTOR-BETA-LIKE TUMOR SUPPRESSOR.
PMS1	HNPCC3, PMSL1	5378	PMS1 PROTEIN HOMOLOG 1 (DNA MISMATCH REPAIR PROTEIN PMS1).
PMS2	HNPCC4, PMSL2	5395	PMS1 PROTEIN HOMOLOG 2 (DNA MISMATCH REPAIR PROTEIN PMS2).
PTPN12	PTPG1	5782	PROTEIN-TYROSINE PHOSPHATASE, NON-RECEPTOR TYPE 12 (EC 3.1.3.48) (PROTEIN-TYROSINE PHOSPHATASE G1) (PTPG1).
PTPRJ	DEP1	5795	PROTEIN-TYROSINE PHOSPHATASE ETA PRECURSOR (EC 3.1.3.48) (R-PTP-ETA) (HPTP ETA) (DENSITY ENHANCED PHOSPHATASE-1) (DEP-1) (CD148 ANTIGEN).
SLC26A3	CLD, DRA	1811	CHLORIDE ANION EXCHANGER (DRA PROTEIN) (DOWN-REGULATED IN ADENOMA).
SRC	SRC1, ASV	6714	PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC (EC 2.7.1.112) (P60-SRC) (C-SRC).
TGFBR2	HNPCC6	7048	TGF-BETA RECEPTOR TYPE II PRECURSOR (EC 2.7.1.37) (TGFR-2) (TGF-BETA TYPE II RECEPTOR).
TP53	P53	7157	CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53) (PHOSPHOPROTEIN P53) (ANTIGEN NY-CO-13).

Table B.2: Expression cluster 2

Attribute	Description	*p*-value
GO:0050876	reproductive physiological process	0.00E-00
GO:0007565	pregnancy	0.00E-00
GO:0005576	extracellular	7.69E-08
GO:0008402	aromatase activity	6.38E-06
GO:0050874	organismal physiological process	1.49E-04
GO:0007567	parturition	2.23E-04
GO:0005615	extracellular space	3.07E-04
GO:0000267	cell fraction	4.84E-04
GO:0005179	hormone activity	8.26E-04
GO:0007611	learning and/or memory	1.68E-03

Table B.3: Expression cluster 3

Attribute	Description	*p*-value
GO:0009607	response to biotic stimulus	0.00E-00
GO:0050874	organismal physiological process	0.00E-00
GO:0007582	physiological process	0.00E-00
GO:0006955	immune response	0.00E-00
GO:0006952	defense response	0.00E-00
GO:0050896	response to stimulus	3.62E-14
GO:0009613	response to pest/pathogen/parasite	4.08E-11
GO:0006950	response to stress	1.64E-08
GO:0019735	antimicrobial humoral response (sensu Vertebrata)	7.81E-07
GO:0019730	antimicrobial humoral response	9.22E-07

Table B.4: Expression cluster 4

Attribute	Description	*p*-value
GO:0007582	physiological process	0.00E-00
GO:0009057	macromolecule catabolism	4.65E-09
GO:0009056	catabolism	2.96E-08
GO:0016787	hydrolase activity	2.28E-07
GO:0007586	digestion	6.72E-07
GO:0008233	peptidase activity	5.04E-06
GO:0006508	proteolysis and peptidolysis	6.26E-06
GO:0030163	protein catabolism	7.26E-06
GO:0007039	vacuolar protein catabolism	8.75E-06
GO:0004263	chymotrypsin activity	9.36E-06

Table B.5: Expression cluster 5

Attribute	Description	*p*-value
GO:0007517	muscle development	2.05E-12
GO:0006936	muscle contraction	2.22E-10
GO:0006928	cell motility	6.42E-08
GO:0030016	myofibril	2.94E-07
GO:0006937	regulation of muscle contraction	2.91E-07
GO:0030017	sarcomere	2.89E-07
GO:0005861	troponin complex	5.85E-07
GO:0030484	muscle fiber	9.28E-07
GO:0005790	smooth endoplasmic reticulum	2.35E-06
GO:0005865	striated muscle thin filament	2.33E-06

Table B.6: Text cluster 2

Attribute	Description	*p*-value
GO:0016020	membrane	0.00E-00
GO:0004713	protein-tyrosine kinase activity	0.00E-00
GO:0005623	cell	0.00E-00
GO:0004871	signal transducer activity	0.00E-00
GO:0004714	transmembrane receptor protein tyrosine kinase activity	0.00E-00
GO:0007154	cell communication	0.00E-00
GO:0005003	ephrin receptor activity	0.00E-00
GO:0009987	cellular process	0.00E-00
GO:0005005	transmembrane-ephrin receptor activity	0.00E-00
GO:0019199	transmembrane receptor protein kinase activity	0.00E-00

Table B.7: Text cluster 3

Attribute	Description	*p*-value
GO:0005665	DNA-directed RNA polymerase II, core complex	0.00E-00
GO:0003899	DNA-directed RNA polymerase activity	0.00E-00
GO:0030880	RNA polymerase complex	0.00E-00
GO:0016779	nucleotidyltransferase activity	0.00E-00
GO:0016591	DNA-directed RNA polymerase II, holoenzyme	7.88E-15
GO:0005654	nucleoplasm	1.09E-13
GO:0006366	transcription from Pol II promoter	1.84E-13
GO:0016772	transferase activity, transferring phosphorus-containing groups	1.04E-10
GO:0006350	transcription	1.76E-09
GO:0016740	transferase activity	1.25E-08

Table B.8: Text cluster 4

Attribute	Description	*p*-value
GO:0016773	phosphotransferase activity, alcohol group as acceptor	0.00E-00
GO:0004672	protein kinase activity	0.00E-00
GO:0006796	phosphate metabolism	0.00E-00
GO:0008152	metabolism	0.00E-00
GO:0007165	signal transduction	0.00E-00
GO:0007154	cell communication	0.00E-00
GO:0003824	catalytic activity	0.00E-00
GO:0005488	binding	0.00E-00
GO:0019538	protein metabolism	0.00E-00
GO:0030554	adenyl nucleotide binding	0.00E-00

Table B.9: Text cluster 5

Attribute	Description	*p*-value
GO:0001568	blood vessel development	0.00E-00
GO:0004871	signal transducer activity	0.00E-00
GO:0001525	angiogenesis	0.00E-00
GO:0009987	cellular process	0.00E-00
GO:0009653	morphogenesis	1.22E-11
GO:0009887	organogenesis	3.25E-10
GO:0008284	positive regulation of cell proliferation	3.89E-10
GO:0007275	development	1.08E-09
GO:0016477	cell migration	3.72E-09
GO:0004714	transmembrane receptor protein tyrosine kinase activity	5.97E-09

Table B.10: Combined cluster 2

Attribute	Description	*p*-value
GO:0005623	cell	0
GO:0007582	physiological process	0
GO:0005886	plasma membrane	7.83E-08
GO:0009897	external side of plasma membrane	8.18E-08
GO:0006952	defense response	8.42E-07
GO:0019814	immunoglobulin complex	1.05E-06
GO:0019815	B-cell receptor complex	1.04E-06
GO:0050853	B-cell receptor signaling pathway	1.54E-06
GO:0009607	response to biotic stimulus	2.50E-06
GO:0009986	cell surface	3.37E-06

Table B.11: Combined cluster 3

Attribute	Description	*p*-value
GO:0007517	muscle development	0
GO:0030017	sarcomere	2.41E-11
GO:0030016	myofibril	4.78E-11
GO:0009887	organogenesis	1.80E-10
GO:0006936	muscle contraction	1.07E-09
GO:0009653	morphogenesis	2.51E-09
GO:0007275	development	1.73E-08
GO:0006937	regulation of muscle contraction	1.25E-07
GO:0006928	cell motility	6.65E-07
GO:0030018	Z disc	8.42E-07

Table B.12: Combined cluster 4

Attribute	Description	*p*-value
GO:0005622	intracellular	0
GO:0007582	physiological process	0
GO:0000280	nuclear division	0
GO:0005623	cell	0
GO:0000278	mitotic cell cycle	7.23E-09
GO:0000067	DNA replication and chromosome cycle	3.07E-08
GO:0007067	mitosis	3.28E-08
GO:0000087	M phase of mitotic cell cycle	3.52E-08
GO:0000279	M phase	2.19E-07
GO:0008283	cell proliferation	4.11E-06

Table B.13: Combined cluster 5

Attribute	Description	*p*-value
GO:0007582	physiological process	0
GO:0004300	enoyl-CoA hydratase activity	6.40E-06
GO:0004608	phosphatidylethanolamine N-methyltransferase activity	0.000101844
GO:0004757	sepiapterin reductase activity	0.000101379
GO:0003858	3-hydroxybutyrate dehydrogenase activity	0.000100914
GO:0006635	fatty acid beta-oxidation	0.00014202
GO:0016491	oxidoreductase activity	0.000336699
GO:0004479	methionyl-tRNA formyltransferase activity	0.000397844
GO:0009256	10-formyltetrahydrofolate metabolism	0.000395985
GO:0009258	10-formyltetrahydrofolate catabolism	0.000394126

syndrome, xeroderma pigmentosum, Zellweger syndrome. Automated HUGO-to-Ensembl mapping reduced the number of genes for a few diseases. The smallest gene set (ALS) contained only 4 Ensembl genes, while the largest set (leukemia) contained 113 genes (see Table B.14-B.16). For the pathway training sets (WNT ligand, Notch receptor, and EGF receptor pathways) genes of Drosophila melanogaster were selected because the pathways are better described for this organism than for human. The training sets consist of all genes annotated with the respective Gene Ontology terms and all their child terms in the GO-graph: *Notch signaling pathway* (GO:0007219), *Epidermal growth factor receptor signaling pathway* (GO:0007173), and *Wnt receptor signaling pathway* (GO:0016055). The mapping from fruit fly to human was performed using the BioMart system of Ensembl (`http://www.ensembl.org/Multi/martview`) (see Table B.17).

Table B.14: Overview of the 29 diseases and their corresponding training sets. The sets were retrieved from the OMIM database.

Disease	Training genes
alzheimer	ENSG00000012433, ENSG00000080815, ENSG00000108578, ENSG00000142192, ENSG00000143801, ENSG00000159640, ENSG00000171634, ENSG00000173391
amyotrophic lateral sclerosis	ENSG00000003393, ENSG00000100285, ENSG00000138380, ENSG00000142168
anemia	ENSG00000001084, ENSG00000004939, ENSG00000011590, ENSG00000070182, ENSG00000091513, ENSG00000100983, ENSG00000101093, ENSG00000102144, ENSG00000102145, ENSG00000104687, ENSG00000105220, ENSG00000105372, ENSG00000106992, ENSG00000107611, ENSG00000111669, ENSG00000112039, ENSG00000112077, ENSG00000115392, ENSG00000117479, ENSG00000119139, ENSG00000122643, ENSG00000124275, ENSG00000125347, ENSG00000130654, ENSG00000131269, ENSG00000134812, ENSG00000136881, ENSG00000139618, ENSG00000140326, ENSG00000141959, ENSG00000143627, ENSG00000143819, ENSG00000144554, ENSG00000156515, ENSG00000158169, ENSG00000158578, ENSG00000160211, ENSG00000165281, ENSG00000166126, ENSG00000172331, ENSG00000183161, ENSG00000187741, ENSG00000188170, ENSG00000188985
breast cancer	ENSG00000012048, ENSG00000023287, ENSG00000039068, ENSG00000050820, ENSG00000051180, ENSG00000074319, ENSG00000078725, ENSG00000085733, ENSG00000085999, ENSG00000091831, ENSG00000110628, ENSG00000124151, ENSG00000136492, ENSG00000139618, ENSG00000141510, ENSG00000149311, ENSG00000160182, ENSG00000167085, ENSG00000169083, ENSG00000170836, ENSG00000173267, ENSG00000174744, ENSG00000183566, ENSG00000183765
cardiomyopathy	ENSG00000014919, ENSG00000069431, ENSG00000092054, ENSG00000096696, ENSG00000101306, ENSG00000102125, ENSG00000106617, ENSG00000111245, ENSG00000114854, ENSG00000118194, ENSG00000129170, ENSG00000129991, ENSG00000132438, ENSG00000134571, ENSG00000140416, ENSG00000155657, ENSG00000159251, ENSG00000160789, ENSG00000166094, ENSG00000170624, ENSG00000173991, ENSG00000175084
cataract	ENSG00000007372, ENSG00000087086, ENSG00000100058, ENSG00000102878, ENSG00000104313, ENSG00000105370, ENSG00000107859, ENSG00000108255, ENSG00000108479, ENSG00000109846, ENSG00000111846, ENSG00000118231, ENSG00000119614, ENSG00000121743, ENSG00000135517, ENSG00000140263, ENSG00000160202, ENSG00000163254, ENSG00000170819, ENSG00000189408
charcot-marie-tooth disease	ENSG00000054523, ENSG00000075785, ENSG00000087053, ENSG00000104381, ENSG00000104419, ENSG00000104725, ENSG00000106105, ENSG00000109099, ENSG00000122877, ENSG00000133812, ENSG00000158887, ENSG00000160789, ENSG00000169247, ENSG00000189067
colorectal cancer	ENSG00000064933, ENSG00000068078, ENSG00000076242, ENSG00000087088, ENSG00000095002, ENSG00000100393, ENSG00000104213, ENSG00000110092, ENSG00000119684, ENSG00000122512, ENSG00000133703, ENSG00000141510, ENSG00000157764, ENSG00000163513, ENSG00000168036, ENSG00000168638, ENSG00000168646, ENSG00000169679, ENSG00000171444, ENSG00000183765, ENSG00000187323, ENSG00000188257
deafness	ENSG00000006611, ENSG00000017427, ENSG00000083307, ENSG00000091010, ENSG00000091137, ENSG00000091536, ENSG00000095777, ENSG00000100345, ENSG00000100473, ENSG00000101384, ENSG00000103316, ENSG00000103888, ENSG00000105928, ENSG00000107485, ENSG00000107736, ENSG00000109927, ENSG00000112112, ENSG00000112319, ENSG00000112698, ENSG00000115155, ENSG00000116039, ENSG00000117013, ENSG00000121742, ENSG00000126953, ENSG00000129158, ENSG00000131504, ENSG00000135903, ENSG00000137474, ENSG00000139219, ENSG00000150781, ENSG00000152591, ENSG00000155719, ENSG00000156313, ENSG00000159261, ENSG00000160183, ENSG00000162399, ENSG00000165091, ENSG00000165474, ENSG00000166763, ENSG00000166866, ENSG00000168269, ENSG00000184009, ENSG00000188910
diabetes	ENSG00000049768, ENSG00000101076, ENSG00000101200, ENSG00000104812, ENSG00000106633, ENSG00000108753, ENSG00000115159, ENSG00000118495, ENSG00000121351, ENSG00000121653, ENSG00000126895, ENSG00000129965, ENSG00000132170, ENSG00000135100, ENSG00000139515, ENSG00000142330, ENSG00000162992, ENSG00000163581, ENSG00000163599, ENSG00000164266, ENSG00000166592, ENSG00000167580, ENSG00000169047, ENSG00000171105, ENSG00000181856, ENSG00000185950, ENSG00000187805
dystonia	ENSG00000127990, ENSG00000131979, ENSG00000136827, ENSG00000149295, ENSG00000169676
ehlers-danlos	ENSG00000027847, ENSG00000083444, ENSG00000087116, ENSG00000108821, ENSG00000115414, ENSG00000130635, ENSG00000139219, ENSG00000164692, ENSG00000168477, ENSG00000168542

Table B.15: Overview of the 29 diseases and their corresponding training sets. The sets were retrieved from the OMIM database.

Disease	Training genes
epilepsy	ENSG00000004848, ENSG00000022355, ENSG00000075043, ENSG00000101204, ENSG00000105711, ENSG00000108231, ENSG00000112425, ENSG00000113327, ENSG00000114859, ENSG00000153253, ENSG00000160213, ENSG00000160716, ENSG00000182389, ENSG00000184156, ENSG00000187566
hemolytic anemia	ENSG00000001084, ENSG00000004939, ENSG00000100983, ENSG00000101093, ENSG00000102144, ENSG00000104687, ENSG00000105220, ENSG00000106992, ENSG00000111669, ENSG00000141959, ENSG00000156515, ENSG00000160211, ENSG00000172331
ichthyosis	ENSG00000011201, ENSG00000092295, ENSG00000101846, ENSG00000143631, ENSG00000144452, ENSG00000165474, ENSG00000167768, ENSG00000172867, ENSG00000186395
leukemia	ENSG00000005271, ENSG00000005339, ENSG00000005483, ENSG00000006451, ENSG00000055609, ENSG00000067955, ENSG00000069399, ENSG00000071564, ENSG00000073921, ENSG00000076242, ENSG00000078399, ENSG00000078403, ENSG00000083168, ENSG00000087088, ENSG00000089041, ENSG00000089280, ENSG00000089693, ENSG00000095002, ENSG00000097007, ENSG00000100361, ENSG00000100721, ENSG00000102145, ENSG00000103035, ENSG00000104320, ENSG00000104903, ENSG00000105656, ENSG00000105663, ENSG00000107807, ENSG00000108292, ENSG00000108924, ENSG00000109220, ENSG00000109906, ENSG00000110092, ENSG00000110713, ENSG00000112043, ENSG00000113594, ENSG00000113721, ENSG00000114802, ENSG00000115297, ENSG00000116652, ENSG00000117400, ENSG00000118058, ENSG00000119335, ENSG00000119537, ENSG00000121741, ENSG00000122025, ENSG00000124795, ENSG00000125347, ENSG00000126883, ENSG00000128342, ENSG00000130382, ENSG00000130396, ENSG00000131759, ENSG00000133392, ENSG00000135363, ENSG00000137497, ENSG00000138336, ENSG00000138698, ENSG00000139083, ENSG00000140464, ENSG00000141736, ENSG00000141985, ENSG00000142867, ENSG00000143322, ENSG00000143384, ENSG00000143437, ENSG00000143443 , ENSG00000144136, ENSG00000145012, ENSG00000145022, ENSG00000145819, ENSG00000147548, ENSG00000148400, ENSG00000149311, ENSG00000149408, ENSG00000151090, ENSG00000151702, ENSG00000151726, ENSG00000156650, ENSG00000157404, ENSG00000157554, ENSG00000159216, ENSG00000162367, ENSG00000162775, ENSG00000163655, ENSG00000164438, ENSG00000164929, ENSG00000165671, ENSG00000166407, ENSG00000167081, ENSG00000167548, ENSG00000168575, ENSG00000169245, ENSG00000170802, ENSG00000171791, ENSG00000171843, ENSG00000172493, ENSG00000173757, ENSG00000176124, ENSG00000178053, ENSG00000178568, ENSG00000179295, ENSG00000181019, ENSG00000181163, ENSG00000182866, ENSG00000184481, ENSG00000184640, ENSG00000185630, ENSG00000185811, ENSG00000186051, ENSG00000186349, ENSG00000186716, ENSG00000187239
lymphoma	ENSG00000002822, ENSG00000003400, ENSG00000023445, ENSG00000046877, ENSG00000069399, ENSG00000072694, ENSG00000085999, ENSG00000095002, ENSG00000099385, ENSG00000100721, ENSG00000103522, ENSG00000106635, ENSG00000110092, ENSG00000110987, ENSG00000113916, ENSG00000116128, ENSG00000119537, ENSG00000119866, ENSG00000121741, ENSG00000127152, ENSG00000136997, ENSG00000142867, ENSG00000149311, ENSG00000156299, ENSG00000164947, ENSG00000171094, ENSG00000171791, ENSG00000172175, ENSG00000172458, ENSG00000181274, ENSG00000187621
mental retardation	ENSG00000004848, ENSG00000017427, ENSG00000068366, ENSG00000077264, ENSG00000079482, ENSG00000083635, ENSG00000085224, ENSG00000089289, ENSG00000101935, ENSG00000102081, ENSG00000102103, ENSG00000102129, ENSG00000102172, ENSG00000102302, ENSG00000114416, ENSG00000129245, ENSG00000129675, ENSG00000134595, ENSG00000155966, ENSG00000156298, ENSG00000164099, ENSG00000169057, ENSG00000169306, ENSG00000169862, ENSG00000177189
muscular dystrophy	ENSG00000092529, ENSG00000100836, ENSG00000102119, ENSG00000102683, ENSG00000106692, ENSG00000108823, ENSG00000109536, ENSG00000111046, ENSG00000119401, ENSG00000120729, ENSG00000132438, ENSG00000135636, ENSG00000142156, ENSG00000155657, ENSG00000160789, ENSG00000162430, ENSG00000163069, ENSG00000163359, ENSG00000170624, ENSG00000172245, ENSG00000173991, ENSG00000178209, ENSG00000181027, ENSG00000182533

Table B.16: Overview of the 29 diseases and their corresponding training sets. The sets were retrieved from the OMIM database.

Disease	Training genes
myopathy	ENSG00000014919, ENSG00000025708, ENSG00000069431, ENSG00000092054, ENSG00000092758, ENSG00000096696, ENSG00000101306, ENSG00000102125, ENSG00000105048, ENSG00000106617, ENSG00000109846, ENSG00000111046, ENSG00000111245, ENSG00000114854, ENSG00000118194, ENSG00000125414, ENSG00000129170, ENSG00000129991, ENSG00000130489, ENSG00000132438, ENSG00000134571, ENSG00000135424, ENSG00000135636, ENSG00000140416, ENSG00000142156, ENSG00000142173, ENSG00000143549, ENSG00000155657, ENSG00000157184, ENSG00000159251, ENSG00000159921, ENSG00000160789, ENSG00000163359, ENSG00000164708, ENSG00000165280, ENSG00000166094, ENSG00000170624, ENSG00000171100, ENSG00000173991, ENSG00000175084, ENSG00000177929
neuropathy	ENSG00000032444, ENSG00000042088, ENSG00000090054, ENSG00000101986, ENSG00000105227, ENSG00000106211, ENSG00000109099, ENSG00000118271, ENSG00000122877, ENSG00000127688, ENSG00000139549, ENSG00000140199, ENSG00000140521, ENSG00000152137, ENSG00000158887, ENSG00000162374, ENSG00000169562, ENSG00000188910
obesity	ENSG00000087916, ENSG00000112246, ENSG00000115138, ENSG00000116678, ENSG00000124089, ENSG00000131910, ENSG00000132170, ENSG00000157017, ENSG00000159723, ENSG00000166603, ENSG00000169252, ENSG00000174697, ENSG00000175567
parkinson	ENSG00000100197, ENSG00000106617, ENSG00000116288, ENSG00000145335, ENSG00000153234, ENSG00000154277, ENSG00000178127, ENSG00000185345, ENSG00000186868
retinitis pigmentosa	ENSG00000031544, ENSG00000042781, ENSG00000070729, ENSG00000092200, ENSG00000097054, ENSG00000102218, ENSG00000104237, ENSG00000105392, ENSG00000105618, ENSG00000106348, ENSG00000112041, ENSG00000112619, ENSG00000116745, ENSG00000117360, ENSG00000129221, ENSG00000129535, ENSG00000132915, ENSG00000133256, ENSG00000134376, ENSG00000140522, ENSG00000148604, ENSG00000149489, ENSG00000153208, ENSG00000156313, ENSG00000163914, ENSG00000164610, ENSG00000170455, ENSG00000174231, ENSG00000186765, ENSG00000188452
spastic paraplegia	ENSG00000021574, ENSG00000046653, ENSG00000141018, ENSG00000144381, ENSG00000149538, ENSG00000151747, ENSG00000155980, ENSG00000170113
spinocerebellar ataxia	ENSG00000042088, ENSG00000089232, ENSG00000112592, ENSG00000124788, ENSG00000126583, ENSG00000141837, ENSG00000156475, ENSG00000163635
usher syndrome	ENSG00000006611, ENSG00000042781, ENSG00000107736, ENSG00000137474, ENSG00000150275, ENSG00000163646, ENSG00000164199, ENSG00000182040
xeroderma pigmentosum	ENSG00000104884, ENSG00000134574, ENSG00000134899, ENSG00000136936, ENSG00000143799, ENSG00000154767, ENSG00000163161, ENSG00000167986, ENSG00000170734, ENSG00000175595
zellweger syndrome	ENSG00000034693, ENSG00000060971, ENSG00000117528, ENSG00000121680, ENSG00000127980, ENSG00000139197, ENSG00000157911, ENSG00000162928, ENSG00000164751

Table B.17: Overview of the 3 pathways and their corresponding training sets. The sets were retrieved from the GO annotations database.

Pathway	**Training genes**
Notch signaling pathway	ENSG00000073536, ENSG00000110042, ENSG00000117362, ENSG00000162736, ENSG00000198488, ENSG00000082701, ENSG00000101384, ENSG00000130396, ENSG00000134250, ENSG00000133961, ENSG00000011304, ENSG00000168214, ENSG00000162924, ENSG00000169733, ENSG00000139697, ENSG00000101849, ENSG00000136842, ENSG00000123124
Epidermal growth factor receptor signaling pathway	ENSG00000172238, ENSG00000110395, ENSG00000065526, ENSG00000103067, ENSG00000178568, ENSG00000134954, ENSG00000133704, ENSG00000169032, ENSG00000091129, ENSG00000158458, ENSG00000179295, ENSG00000139697, ENSG00000160691, ENSG00000136158, ENSG00000101849
Wnt receptor signaling pathway	ENSG00000104964, ENSG00000115266, ENSG00000134982, ENSG00000116128, ENSG00000166167, ENSG00000113712, ENSG00000180138, ENSG00000141551, ENSG00000169118, ENSG00000133275, ENSG00000151292, ENSG00000101266, ENSG00000070770, ENSG00000111968, ENSG00000168036, ENSG00000178585, ENSG00000107984, ENSG00000155011, ENSG00000050165, ENSG00000104371, ENSG00000163348, ENSG00000171016, ENSG00000072803, ENSG00000119402, ENSG00000165879, ENSG00000181274, ENSG00000162998, ENSG00000157240, ENSG00000180340, ENSG00000163251, ENSG00000082701, ENSG00000183762, ENSG00000131650, ENSG00000138795, ENSG00000162337, ENSG00000070018, ENSG00000107829, ENSG00000109062, ENSG00000081059, ENSG00000148737, ENSG00000156076, ENSG00000104415

Nederlandse samenvatting

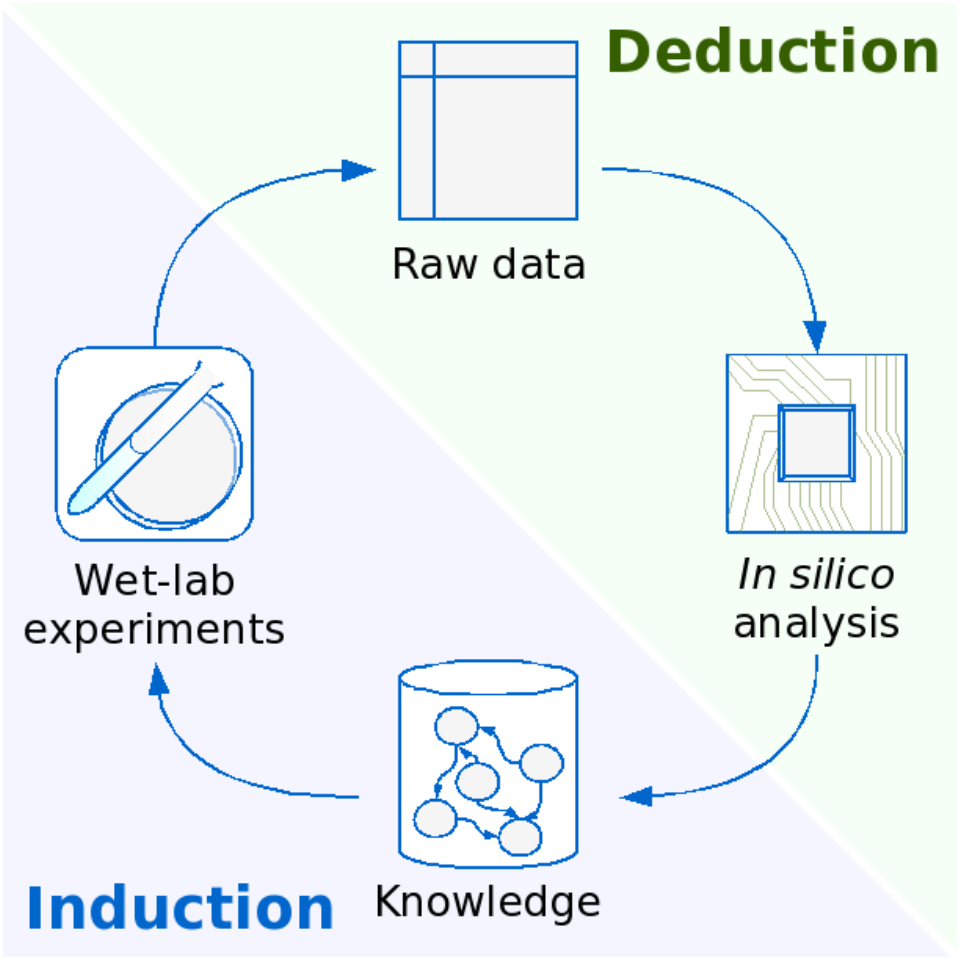
Data-integratietechnieken voor moleculair biologisch onderzoek

Inleiding

Het verwerven van biologische kennis

Het verwerven van kennis kan beschouwd worden als een cyclisch proces. Dit is ook zo in moleculair biologisch onderzoek, waar kennis over een biologisch systeem op een iteratieve manier vergaard wordt. Vaak start men met een hypothese die men aan de werkelijkheid toetst door een experiment op te zetten. De resultaten van het experiment worden vervolgens geïnterpreteerd in de context van wat al geweten is, wat aanleiding kan geven tot nieuwe hypotheses en nieuwe experimenten. Het formuleren van een nieuwe hypothese op basis van waarnemingen is wat men *inductie* noemt. Het verifiëren van deze hypothese in specifieke gevallen door het opzetten van experimenten noemt men *deductie*. Dit is weergegeven in Figuur 1.

De beschikbaarheid van het volledige genoom van steeds meer organismen veranderde de aard van moleculair biologisch onderzoek grondig. Het gaf de mogelijkheid om onderzoek te doen op een genomische schaal, i.e., het gelijktijdig bestuderen van alle genen in een organisme. Deze zogenaamde *high-throughput* technieken deden de klemtoon in het onderzoek verschuiven van de studie van één enkel gen naar de studie van grote groepen van genen [59, 13]. Een high-throughput experiment brengt veel meer data voort, maar de kwaliteit ervan is lager. De opkomst van het Internet drukte ook zijn stempel op het biologisch onderzoek. Het werd mogelijk om grote hoeveelheden wetenschappelijke data op een wereldwijde schaal beschikbaar te maken [70]. Als gevolg van deze ontwikkelingen is bioinformatica een steeds belangrijkere rol gaan spelen, niet alleen om de analyse van grote hoeveelhe-



Figuur 1: Het verwerven van kennis is een cyclisch proces. Tijdens de inductiestap wordt met behulp van de bestaande kennis en experimentele resultaten een nieuwe hypothese geformuleerd. Tijdens de deductiestap worden er experimenten opgezet waarvan de resultaten de hypothese kracht bijzetten of aangeven dat de hypothese bijgesteld dient te worden.

den data te versnellen, maar ook om te helpen bij het vinden en integreren van alle beschikbare en relevante informatie die betrekking heeft op een specifiek onderzoek.

De technieken die in deze thesis beschreven worden, maken duidelijk hoe reeds beschikbare kennis kan betrokken worden bij de analyse van data uit high-throughput experimenten. Er wordt een onderscheid gemaakt tussen vroege, intermediaire en late integratie, afhankelijk van het stadium waarin de analyse zich bevindt. Omdat een groot deel van alle biologische kennis vervat zit in geschreven taal, gaat de volgende paragraaf iets dieper in op het voorstellen en gebruik van tekstuele informatie bij het analyseren van biologische gegevens.

Het gebruik van tekstuele informatie in biologisch onderzoek

Wetenschappelijke artikels kunnen beschouwd worden als een zeer rijke bron van biologische kennis. De vermaarde MEDLINE [41] collectie bevat bijvoorbeeld reeds 15.5 miljoen referenties naar artikels uit biomedisch onderzoek (gegevens van april 2005). Toch is deze informatie moeilijk te ontsluiten. Er is niet alleen een groeiende vraag naar manieren om efficiënt de literatuur te doorzoeken, maar vooral om de relevante informatie er uit te halen en te gebruiken bij de correcte interpretatie van (een steeds groeiende hoeveelheid) experimentele gegevens. Dit kan de analyse van experimentele resultaten aanzienlijk versnellen en de accuraatheid ervan verhogen.

Een computationele methode om teksten te ontsluiten, die in het verleden reeds zijn waarde bewees, is gebaseerd op het concept van een vectorruimte [16]. Men gaat ervan uit dat een groep van documenten, een *corpus*, kan voorgesteld worden in een vectorruimte. De dimensies van deze ruimte vertegenwoordigen alle termen die in het corpus voorkomen. Elk document wordt gerepresenteerd door een vector waarvan elke component een gewicht w_{ij} bevat dat zegt hoe karakteristiek een bepaalde term is voor dit document. Het berekenen van deze gewichten noemt men *indexeren*. Er bestaan verschillende manieren om deze gewichten te berekenen. In deze thesis wordt de zogenaamde *inverse document frequentie* of IDF gebruikt, die berekend wordt als

$$w_{ij}^{\mathrm{IDF}} = log(\frac{N}{n_j}), \tag{B.1}$$

met n_j het aantal keer dat een document term t_j bevat en N het totaal aantal documenten in het corpus. Het IDF schema neemt in rekening dat heel

frequente termen minder interessant zijn om een document te karakteriseren dan termen die slechts in enkele documenten voorkomen.

Van zodra een corpus is voorgesteld in een vectorruimte, kunnen er heel wat mathematische bewerkingen op worden uitgevoerd. Zo kan de gelijkenis tussen twee documenten bepaald worden door de hoek te meten tussen hun twee representatieve vectoren (de zogenaamde cosinus-gelijkenis). Voorts is het mogelijk om representaties te maken van andere entiteiten zoals genen, proteïnen en ziektes, door de gemiddelde vector te berekenen van de vectoren van alle documenten die met de entiteit te maken hebben. Beide bewerkingen worden veelvuldig gebruikt doorheen de thesis.

Genen groeperen

Bij het analyseren van gegevens uit experimenten uitgevoerd op genomische schaal, is vaak de eerste stap het vinden van gelijkaardige patronen die kunnen wijzen op verbanden tussen genen. Zo gaat men er bij de analyse van genexpressiegegevens van uit dat een gelijkaardig expressiepatroon wijst op een functionele relatie tussen twee genen. Dit is wat men noemt de *schuldig-door-associatie* heuristiek [104, 151].

Het vinden van verbanden tussen genen is de eerste stap in het verwerven van nieuwe kennis. Uit de ruwe gegevens wordt nuttige informatie gehaald, vaak met behulp van statistische methoden. Typisch dienen eerst de ruwe gegevens opgekuist te worden. Vervolgens wordt een geschikte clusteringsmethodologie gekozen en worden de meest interessante clusters weerhouden voor een uitgebreider onderzoek. Dit werd gedaan vertrekkende van genexpressiegegevens, tekstuele informatie van genen en de combinatie van beide. Het was vooral interessant om na te gaan of er een relatie bestaat tussen de statistische kwaliteit van een gengroep en zijn biologische coherentie.

In het geval van de genexpressiegegevens werden de profielen eerst gestandaardiseerd om alle absolute verschillen tussen de profielen te verwijderen. Daarna werden de profielen hiërarchisch geclusterd met behulp van Ward's minimale variantie methode. De gelijkenis tussen twee profielen werd bepaald met de Pearson correlatie, die voor gestandaardiseerde profielen gelijk is aan de cosinus van de hoek tussen de vectoren. Een clustering op basis van Ward's methode resulteert in compacte, sferische clusters met gelijkaardige groottes. Een probleem bij clusteren is steeds het vinden van een optimaal aantal groepen. Hier werd er gekozen om alle groepen die 10 tot 20 genen bevatten, te weerhouden voor verdere analyse. De (statistische) kwaliteit van de clusters werd bepaald met behulp van de Silhouette

coëfficiënt. Deze coëfficiënt geeft de sterkte van de structuur aan die in de data teruggevonden werd.

De clusteranalyse op basis van tekstuele informatie werd toegevoegd om aan te tonen dat het mogelijk is om met behulp van de literatuur relaties tussen genen te achterhalen. Dezelfde genen die geclusterd werden op basis van genexpressiegegevens, werden deze maal geclusterd met behulp van tekstuele informatie. Na het bouwen van de hiërarchische boom met Ward's methode, werd de kwaliteit van alle gengroepen met 10 tot 20 genen, opnieuw bepaald door middel van de Silhouette coëfficiënt.

Om aan te tonen dat verworven kennis reeds in de clusteranalyse van high-throughput gegevens kan worden betrokken, werd een clustering uitgevoerd op basis van de combinatie van genexpressiegegevens en tekstuele data. Deze vorm van vroege integratie noemt men ook wel eens gesuperviseerd clusteren. Hoewel de hier gevolgde methodologie niet optimaal voordeel haalde uit de combinatie van deze twee datatypes, werd reeds eerder aangetoond dat een geïntegreerde aanpak kan leiden tot biologisch meer relevante en coherente gengroepen.

Het ligt voor de hand dat het bekijken van de resultaten van clusteranalyses en het manueel selecteren van interessante clusters onbegonnen werk is in de context van high-throughput experimenten. De beschreven methodologie voert reeds een eerste selectie uit op basis van de grootte en statistische kwaliteit van gengroepen. Toch is voorzichtigheid geboden, omdat nog niet werd aangetoond dat de statistische kwaliteit van een gengroep ook biologisch relevant is. Eveneens is het geen triviale taak om de gevonden gengroepen manueel te karakteriseren en valideren. Beide kwesties worden in het volgende deel behandeld.

Gengroepen valideren

Het valideren van een groep van genen houdt in te achterhalen wat de biologische relatie is tussen deze genen. Dit is de volgende stap in het analyseren van high-throughput gegevens, waarbij de *nieuwe* informatie (de gevonden patronen en verbanden) wordt bekeken in de context van de bestaande kennis. Valideren heeft tot doel twee belangrijke vragen te beantwoorden:

- Wat is de biologische functie van een gengroep?
- Welke genen horen niet bij de groep?

In de thesis werden twee verschillende types functionele informatie over genen aangewend om deze twee vragen te beantwoorden: enerzijds functionele

annotaties uit Gene Ontology en anderzijds tekstuele beschrijvingen van genen uit MEDLINE abstracten.

Gengroepen karakteriseren met behulp van Gene Ontology

Gene Ontology (GO) [132] is een gestructureerd vocabularium van termen die betrekking hebben op moleculaire functies, biologische processen en cellulaire componenten. Het wordt beschouwd als de referentie om genen te annoteren en wordt dan ook veelvuldig gebruikt. De vraag rees of de annotaties van afzonderlijke genen kunnen gebruikt worden om een groep van genen te karakteriseren en afwijkende genen in de groep te detecteren.

De beschreven methode gebruikt een binomiale statistiek, aangevuld met een correctie voor meervoudig testen (Holm's correctie), om na te gaan welke GO annotaties overgerepresenteerd zijn in de set van alle annotaties van de genen in een groep. De kans dat men door puur toeval n of meer keer dezelfde annotatie vindt in een set van N genen en met een probabiliteit $p = 1 - q$ dat men een gen heeft met een bepaalde annotatie, wordt gegeven door

$$p = 1 - \sum_{i=1}^{n} \binom{N}{n} p^n q^{N-n}.$$

Deze methode is in staat om op een zeer efficiënte manier de belangrijkste functionele kenmerken van een gengroep bloot te leggen. Het is duidelijk dat dit te verkiezen is boven het manueel opzoeken van de functionele beschrijvingen van elk gen in de groep. Toch geeft deze methode weinig zicht op de bijdrage van de individuele genen. Daarom is het belangrijk om eveneens een zicht te krijgen op de coherentie van de groep.

Vermits Gene Ontology een boomstructuur heeft, kunnen de afstanden in de boom tussen elk van de termen berekend worden. Hoe verder twee termen van elkaar verwijderd zijn in de boom, hoe minder semantische verwantschap ze hebben. De maat voor coherentie van een gengroep die in deze thesis is beschreven, is hierop gebaseerd. Ze wordt bepaald aan de hand van de groep annotaties die hoort bij de gengroep. Als deze annotaties gemiddeld ver van elkaar liggen, dan heeft de groep aan lage coherentie; in het andere geval een hoge.

Om de maat onafhankelijk te maken van de grootte van een gengroep en het aantal geannoteerde GO termen, wordt op voorhand een verdeling van gemiddelde afstanden opgesteld op basis van random samengestelde gengroepen van een bepaalde grootte. Met behulp van deze verdeling kan dan

de kans berekend worden dat een groep van die grootte een set GO annotaties heeft met een gemiddelde afstand als waargenomen. Hoe kleiner deze kans, hoe coherenter de gengroep. Er werd geen grondige studie uitgevoerd om de geldigheid van deze methode na te gaan. Wel werd een correlatie waargenomen tussen deze maat voor coherentie en de Silhouette coëfficiënt, wat betekent dat de statistische kwaliteit van een cluster inderdaad iets zegt over de biologische samenhorigheid van de genen in de cluster.

Tekstueel profileren van gengroepen

Zoals reeds eerder vermeld, zit een groot deel van de biologische kennis vervat in wetenschappelijke publicaties. Met behulp van de hoger beschreven technieken om tekst te ontsluiten op basis van het vectorruimtemodel, kan deze informatie gebruikt worden om groepen van genen te karakteriseren. De achterliggende idee is om de tekstuele profielen van alle genen in de groep samen te nemen tot een groepsprofiel en vervolgens de termen te rangschikken volgens hun gewicht.

Om de genprofielen te maken, werd er beroep gedaan op de informatie in *Entrez Gene* die genen verbindt met relevante MEDLINE documenten. In een eerste stap werden al deze documenten geïndexeerd zoals hierboven beschreven. Op basis van de bekomen index werd vervolgens een gen-index gemaakt door per gen de profielen van alle documenten die ermee verbonden zijn, samen te nemen. Omdat men vaak slechts geïnteresseerd is in een bepaald aspect van een gengroep, werden dezelfde documenten meermaals geïndexeerd met verschillende domeinvocabularia. Een domeinvocabularium bepaalt welke termen in de index worden opgenomen en welke niet. Zo bevat het *GO* vocabularium bijvoorbeeld vooral termen die te maken hebben met de moleculaire functie van genen, terwijl de *OMIM* en *MeSH* vocabularia vooral ziekte-gerelateerde termen bevat. Het *eVOC* vocabularium tenslotte, bevat vooral termen die te maken hebben met anatomie, celtype, pathologie en fase van ontwikkeling.

De beschreven methodologie werd beschikbaar gemaakt via een gebruiksvriendelijke webapplicatie, TXTGate [56]. TXTGate laat toe om een gengroep in te geven en zijn profiel te visualiseren. De individuele genprofielen kunnen vervolgens opnieuw geclusterd worden om sub-structuren in meer detail te bestuderen. Deze subgroepen kunnen dan opnieuw in profiel gebracht worden. De applicatie biedt ook tal van hyperlinks naar andere databanken met biologische informatie om verder onderzoek van een gengroep te vereenvoudigen. Figuur 2 toont een gengroep die in profiel werd gebracht met de verschillende domeinvocabularia. Het is duidelijk dat een dergelijk profiel

niet alleen een mooi overzicht geeft van de belangrijkste termen die teruggevonden werden in de beschrijvingen van de individuele genen, maar ook een inzicht geeft in de herkomst van de termen door eveneens de individuele profielen te visualiseren.

Het karakteriseren van een gengroep met GO annotaties is minder onderhevig aan ruis dan het in profiel brengen met behulp van tekstuele informatie. Toch bevatten de tekstuele profielen een rijkere semantiek en laten ze toe om een duidelijker inzicht te krijgen in de bijdragen van elk van de genen.

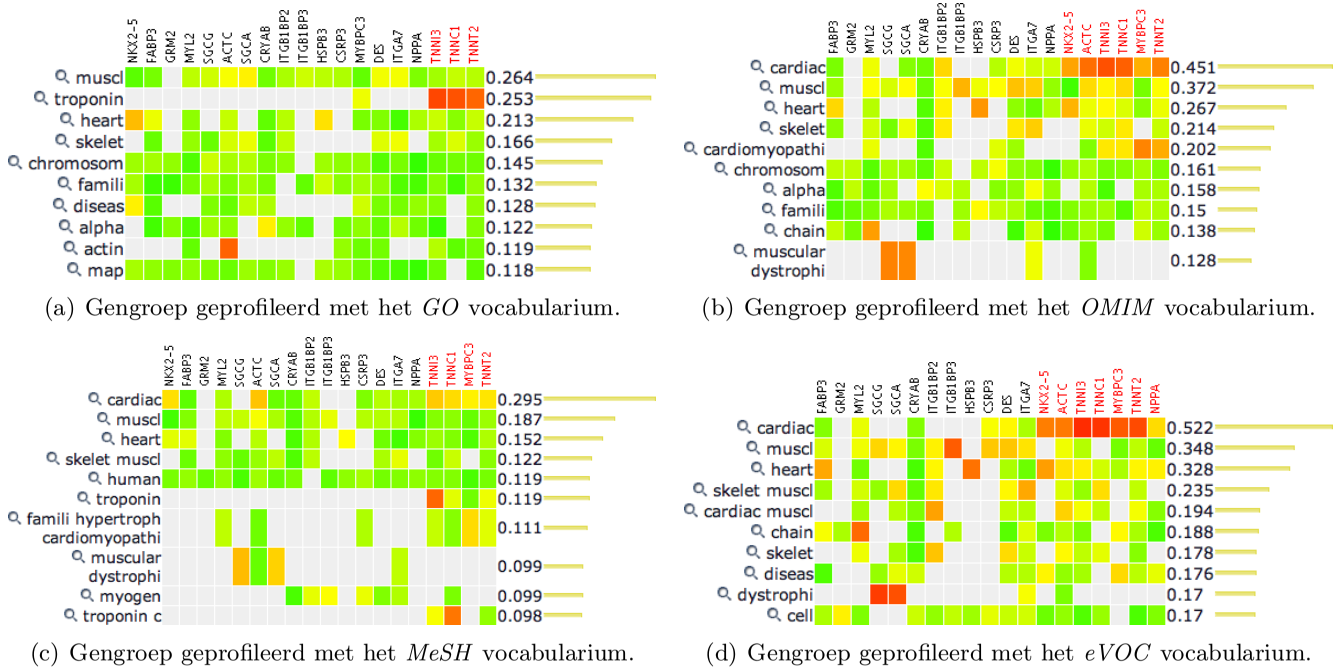
Gengroepen uitbreiden

Van zodra een interessante gengroep werd uitgekozen en gevalideerd met behulp van bovenstaande methoden, ontstaat de vraag of er andere genen zijn die mogelijks een biologische relatie hebben met de groep. De methoden die hieronder beschreven worden, hebben tot doel om op basis van de beschikbare informatie nieuwe en potentieel interessante relaties te vinden tussen genen. Deze relaties kunnen dan de start zijn van een nieuw laboratoriumonderzoek. De eerste methode beroept zich op het domein van de *Knowledge Discovery* om indirecte relaties tussen genen, die vermeld worden in wetenschappelijke artikels, te achterhalen. De tweede methode combineert een tiental complementaire databronnen om lange lijsten van kandidaat-genen te rangschikken volgens gelijkenis met een interessante gengroep.

Gencocitatie en colinkage

De hier beschreven methode gaat met behulp van gencocitatie op zoek naar zowel directe als indirecte verbanden tussen genen. Twee genen zijn gecociteerd als ze beide voorkomen in hetzelfde abstract. Ze worden dan ook verondersteld een biologische verwantschap te hebben. Indirecte relaties tussen genen kunnen gevonden worden door gennamen te zoeken in verwante abstracten. Twee stukken tekst kunnen verwant zijn omdat ze qua woordgebruik erg op elkaar lijken, of omdat ze verbonden zijn met genen die tot eenzelfde groep behoren. Om dergelijke indirecte verbanden te benoemen, werd de term colinkage geïntroduceerd.

De methode is gebaseerd op het eerder beschreven vectorruimtemodel om een corpus van documenten voor te stellen. Er werd een genindex gemaakt met behulp van een domeinvocabularium dat enkel bestaat uit alle gekende humane gensymbolen in hoofdletters (afgeleid van de lijst die ter beschikking wordt gesteld door het Human Gene Nomenclature Committee



(a) Gengroep geprofileerd met het *GO* vocabularium.

(b) Gengroep geprofileerd met het *OMIM* vocabularium.

(c) Gengroep geprofileerd met het *MeSH* vocabularium.

(d) Gengroep geprofileerd met het *eVOC* vocabularium.

Figuur 2: Tekstuele profielen van een gengroep om aan te tonen dat met behulp van domeinvocabularia verschillende aspecten kunnen benadrukt worden. Zo legt *GO* de nadruk op moleculaire functie met termen als *actin* en *troponin*. Bij *OMIM* en *MeSH* ligt de nadruk op ziekte-gerelateerde termen als *cardiac*, *cardiomyopathi* en *muscular dystrophi*. Het *eVOC* vocabularium legt dan weer de nadruk op cel- en weefseltypes, zoals de termen *skelet muscl* en *cardiac muscl* laten blijken.

van de Human Genome Organisation). Deze index bevat voor elk humaan gen dat gekend is in Entrez Gene, een profiel bestaande uit de gensymbolen die voorkomen in de documenten verbonden aan dit gen. TXTGate laat toe om op basis van deze index een gemeenschappelijk profiel te berekenen voor een groep van genen. Dit profiel geeft een overzicht van alle genen die direct en indirect een verband hebben met de genen in de groep.

Hoewel het bekomen profiel een heel aantal vals-positieven bevatte, kwamen er toch enkele interessante verbanden aan het licht die verder onderzoek waard zijn. De methode heeft duidelijk een voordeel ten opzichte van strikt cocitatie-gebaseerde methoden, omdat het meer ruimte laat om indirecte verbanden te ontdekken. Vermits homonymie het belangrijkste probleem is bij deze methode, kan het interessant zijn om het gebruik van Latent Semantisch Indexeren (LSI) na te gaan [33]. LSI doet beroep op matrix-decompositietechnieken om homonieme dimensies uit elkaar te halen en synonieme dimensies te combineren. Ook kan er gekeken worden naar betere methoden om gennamen in abstracten te identificeren. Dit is echter een onderzoeksdomein op zichzelf.

Computationele prioritisatie van genen

Prioritiseren van genen is het rangschikken van genen volgens relevantie voor het onderzoek. Het begrip komt van het domein van de associatiestudies, waar men op basis van associaties tussen fenotypes en genetische aberraties probeert te achterhalen welke genen verantwoordelijk zijn voor een bepaalde aandoening. Vaak liggen er in de geassocieerde chromosomale regio echter een groot aantal genen. Elk van deze genen onderzoeken op mogelijke betrokkenheid bij de aandoening is een uiterst tijdrovende en dure aangelegenheid. Vandaar het belang om de genen te rangschikken zodat het ziekteverwekkende gen sneller gevonden wordt. Vermits steeds meer gen-gerelateerde informatie vrij beschikbaar is via het Internet, wordt het haalbaar om dergelijke prioritisaties door de computer te laten uitvoeren. Het voordeel is dat de analyse sneller kan worden uitgevoerd, veel meer gegevens in rekening kan brengen en minder beïnvloed wordt door de achtergrondkennis van de onderzoeker.

De methodologie die in deze thesis beschreven wordt, rangschikt kandidaatgenen op basis van hun gelijkenis met een groep van geselecteerde trainingsgenen [3]. Deze trainingsgenen zijn genen die een bepaald proces vertegenwoordigen doordat reeds werd aangetoond dat ze een rol spelen in het proces. In het geval van een ziekte komen hiervoor alle genen in aanmerking waarvan reeds werd aangetoond dat ze een verband hebben met de ziekte. In geval

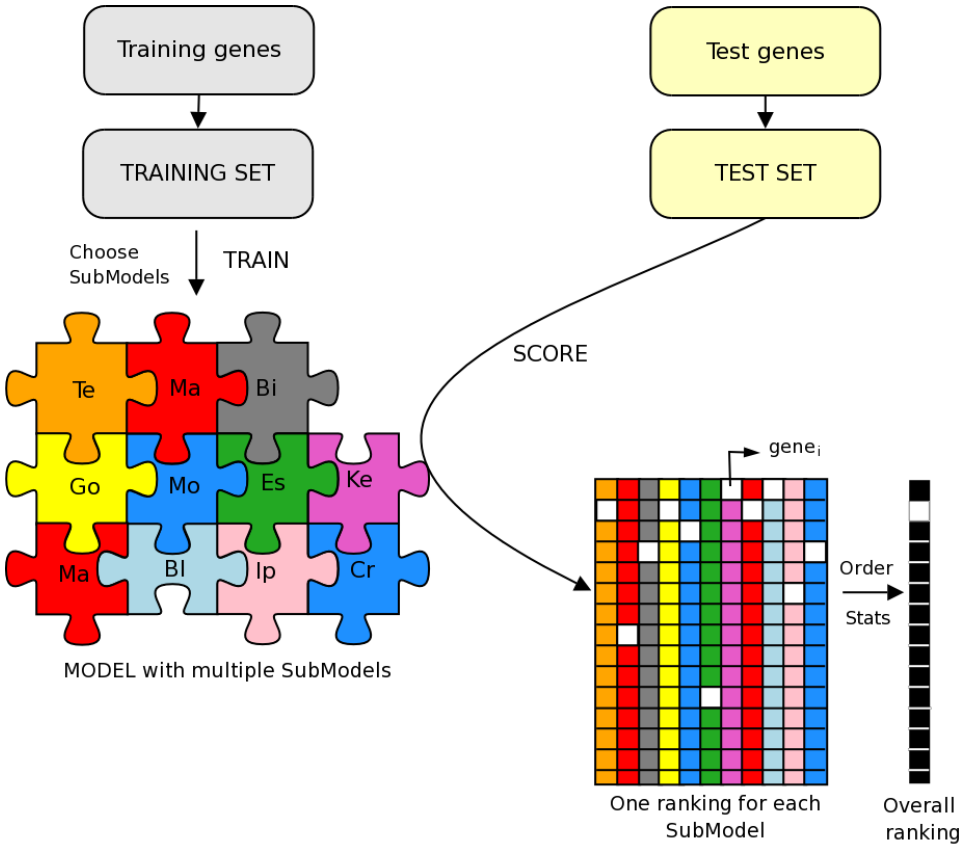
van een reactieweg kunnen alle reeds gekende genen genomen worden. De kandidaat-genen worden gerangschikt door een tiental verschillende informatiebronnen te raadplegen en de informatie van elk van de genen te vergelijken met de informatie van de trainingsgenen. Volgende bronnen worden geraadpleegd: GO annotaties, Kegg reactiewegen, EST-gebaseerde expressiedata, microrooster-gebaseerde expressiegegevens, InterPro proteïnedomeinen, tekstuele informatie uit MEDLINE, sequentiesimilariteit (berekend met behulp van BLAST), proteïne-proteïne-interacties uit BIND en *cis*-regulatorische motieven en modules.

Figuur 3 toont de verschillende stappen in het prioritisatie-proces. Eerst wordt van alle trainingsgenen de gen-gerelateerde informatie opgevraagd en samengevat. De samengevatte informatie uit elk van de informatiebronnen wordt een *submodel* genoemd. Het geheel van de tien submodellen wordt het *model* genoemd. Op basis van dit model, dat een ziekte of reactieweg voorstelt, worden tijdens het *scoren* de kandidaat-genen gerangschikt. Bij het scoren krijgt elk van de kandidaat-genen een score voor elk van de beschikbare informatiebronnen gebaseerd op de gelijkenis van de informatie van het gen met de samengevatte informatie van de trainingsgenen. Hoe beter deze score, des te beter het gen gerangschikt wordt. In de laatste stap worden de tien rangschikkingen (één rangschikking voor elk van de informatiebronnen) samengevoegd met behulp van orderstatistiek.

De orderstatistiek berekent de kans dat een bepaald gen zo hoog in de rangschikkingen voorkomt als werd waargenomen. Deze kans wordt berekend met behulp van de ratio's van de rangschikkingen over het aantal genen die in de rangschikkingen werden opgenomen. Gegeven de ratio's $r_1, r_2, \ldots, r_n$ voor een kandidaat-gen, dan kan de kans om deze of een betere sequentie van ratio's te bekomen, berekend worden met volgende recursieve formule met complexiteit $O(n^2)$:

$$V_n(r_n, r_{n-1}, \ldots, r_1) = \sum_{i=1}^{n} (-1)^{i-1} \frac{V_{n-i}(r_n, \ldots, r_{i+1})}{i!} r_1^i$$

Vermits de bekomen probabiliteiten in het geval van een verschillend aantal ratio's geen uniforme verdeling hebben, kunnen deze niet met elkaar vergeleken worden. Het komt echter vaak voor dat er in een bepaalde informatiebron voor een gen geen gegevens worden gevonden. Om de probabiliteiten van alle kandidaat-genen toch te kunnen vergelijken en een algemene rangschikking mogelijk te maken, werd daarom voor elk mogelijk aantal ratio's door random trekking een verdeling geschat. De cumulatieve distributiefuncties van deze verdelingen laten toe om voor elke probabiliteit een



Figuur 3: De figuur toont de verschillende stappen om kandidaat-genen te prioritiseren volgens hun gelijkenis met trainingsgenen. De trainingsgenen worden gebruikt om een model op te bouwen van een ziekte of reactieweg. Dit model bestaat uit verschillende submodellen die van alle genen in de trainingsset de samengevatte informatie uit een tiental verschillende informatiebronnen bevatten. De kandidaatgenen worden voor elke informatiebron apart gerangschikt volgens gelijkenis met de samengevatte informatie van de trainingsgenen. De verschillende rangschikkingen worden in een laatste stap samengevoegd met behulp van orderstatistiek.

p-waarde te berekenen die onafhankelijk is van het originele aantal ratio's.

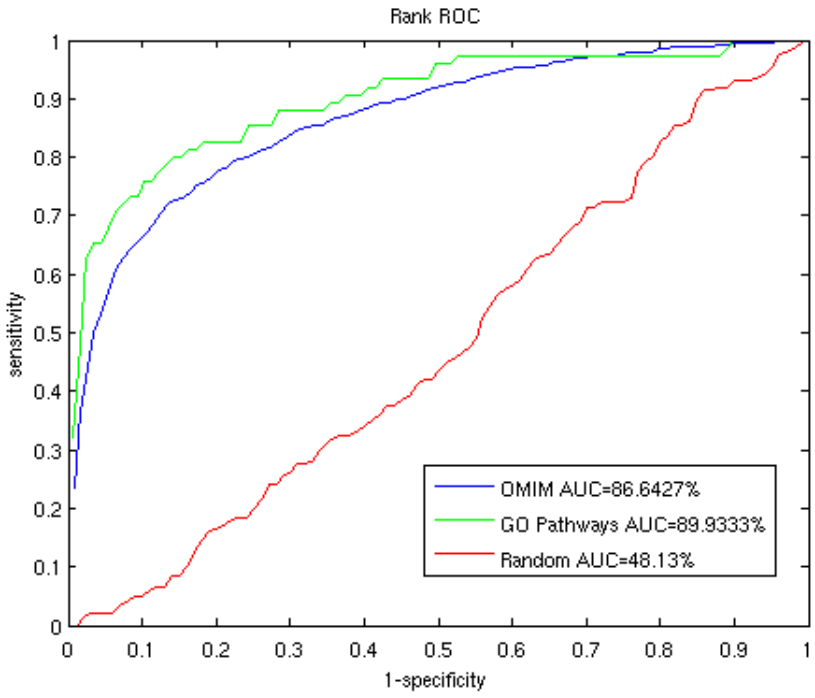
Deze methodologie werd gevalideerd door een grootschalige analyse op te zetten van 29 ziekten en 3 reactiewegen. Voor elk van de ziekten en reactiewegen werden alle gerelateerde genen verzameld. Deze genen werden gebruikt om in totaal 29+3 modellen op te zetten. Voor elk model werd elk van de genen vervolgens uit het model gehaald en in een groep van 99 random geselecteerde genen geplaatst. Deze groep werd vervolgens op basis van het model gerangschikt en de rangschikking van het ene, uit het model verwijderde gen werd genoteerd. Op basis van deze gegevens kan voor een bepaalde cutoff de sensitiviteit en specificiteit van de methode in het terugvinden van het weggelaten gen, berekend worden. Dit laat toe om zogenaamde Receiver Operating Characteristic (ROC) curven op te stellen die aangeven hoe goed de methode werkt. De ROC curven van de uitgevoerde validatie zijn weergegeven in Figuur 4 en zijn duidelijk significant beter dan een random rangschikking.

De validatie toont duidelijk aan dat de computationele prioritisaties leiden tot biologisch zinvolle resultaten. In de literatuur zijn verschillende andere methodes beschreven voor de prioritisatie van kandidaat-genen, maar een uitgebreide vergelijking met deze methode werd nog niet uitgevoerd. Toch mag gezegd worden dat de grootschalige validatie en de bekomen performanties ongeëvenaard zijn. De beschreven methode laat een flexibeler gebruik van informatiebronnen toe en geeft een volledige controle over het bouwen van modellen voor ziekten en reactiewegen. Eveneens werd aangetoond dat de methode ook ongekende genen de kans geeft een goede rangschikking te halen. Daarnaast laat de solide statistische basis toe om genen te selecteren op basis van een significantiedrempel.

Geïntegreerde toegang tot biologische data

Uit de vorige delen is reeds gebleken dat de nood om op een geïntegreerde manier gebruik te maken van complementaire informatiebronnen toeneemt, niet alleen omdat de kwaliteit van gegevens uit high-throughput experimenten lager is, maar ook omdat de nadruk steeds vaker ligt op het begrijpen van een biologisch systeem in zijn geheel. Door de steeds toenemende hoeveelheid biologische gegevens die via het Internet beschikbaar zijn, en door hun heterogeniteit, wordt het efficiënt ophalen en integreren van deze gegevens het meest belangrijke knelpunt bij het versnellen van biologisch onderzoek. De bioinformatica heeft hier een belangrijke rol te spelen [123].

Een technologie die een oplossing kan bieden voor dit probleem is de tech-



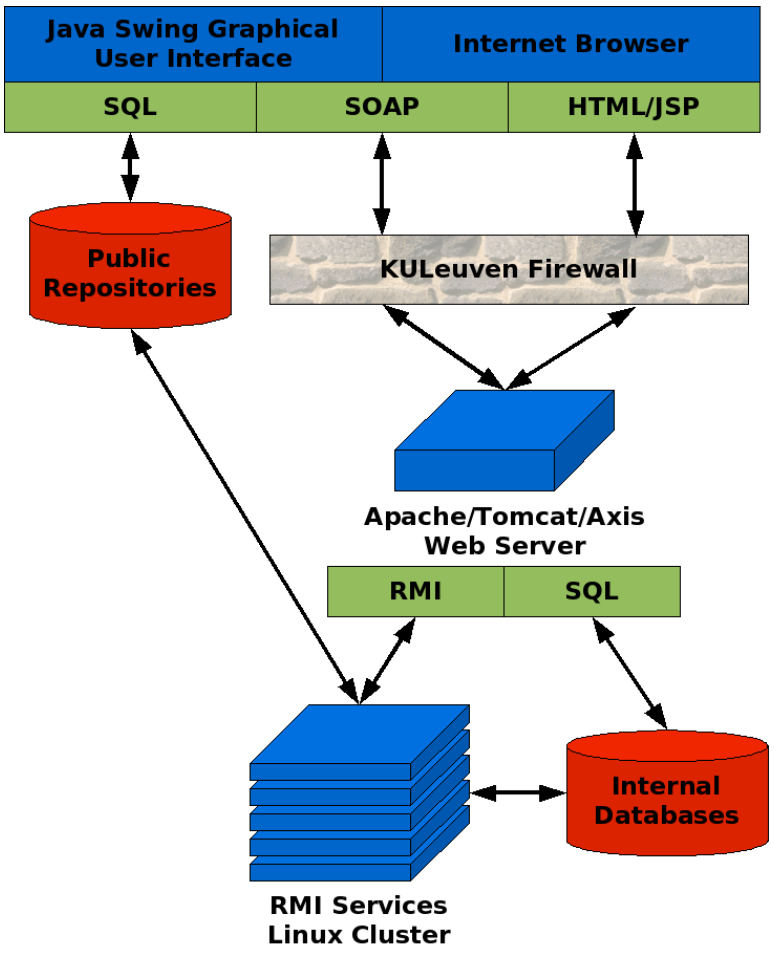
Figuur 4: De plot toont de ROC curven van de grootschalige analyse van 29 ziekten en 3 reactiewegen. Het oppervlak onder de curve is een maat voor de performantie van de methode. De analyses leveren biologisch relevante resultaten op die significant beter zijn dan random prioritisaties. Het weggelaten gen wordt in het geval van de ziekten en de reactiewegen in respectievelijk 85% en 95% van de gevallen teruggevonden in de top-50% genen van de resulterende rangschikking. In 50% van de gevallen (60% voor de reactiewegen) wordt het weggelaten gen teruggevonden in de top-10% van de kandidaat-genen.

nologie van de web services. Web services zijn programma's die netwerkcommunicatie tussen computers ondersteunen, onafhankelijk van het platform van deze computers. Er bestaan enkele standaard web service specificaties die hiervoor gebruikt worden. Het Simple Object Access Protocol (SOAP) is de taal waarin boodschappen geschreven worden. Een computer kan dan deze boodschap versturen naar een andere computer om een bepaalde taak uit te voeren. Hoe die boodschap er juist moet uitzien, staat beschreven in de Web Service Description Language (WSDL). Deze standaard talen laten toe om op een uniforme manier communicatie tot stand te brengen en bijvoorbeeld de toegang tot een biologische informatiebron of het aanroepen van een bioinformatica-algoritme, te vereenvoudigen. Reeds twee bekende bioinformatica-projecten maakten gebruik van deze technologie om een experimenteel bioinformatica-analyseplatform op te zetten. Het BioMOBY project [149], enerzijds, heeft tot doel een eenvoudig en uitbreidbaar platform te implementeren voor het zoeken naar, en het representeren, ophalen en integreren van heterogene biologische databronnen. Het *my*Grid project [124], anderzijds, stelt het gebruik van grid- en web-servicetechnologie centraal om het uitvoeren van complexe bioinformatica-analyses te vereenvoudigen. Beide projecten werken nauw samen op het vlak van semantische beschrijvingen van bioinformatica-services, die nodig zijn om de services op een geautomatiseerde manier te vinden en met elkaar te koppelen.

Omwille van de eenvoud en flexibiliteit van het werken met web services, werd in het kader van deze thesis een web-service-architectuur ontwikkeld en werden vele van de eerder besproken methoden als web services geïmplementeerd (zie Figuur 5). Deze web services maken deel uit van drie afzonderlijke projecten: INCLUSive [28], Toucan [7, 5] en Endeavour [3].

INCLUSive is een set van algoritmen voor de analyse van genexpressiegegevens en regulatorische DNA-gebieden. De algoritmen zijn beschikbaar via een gebruiksvriendelijke website of kunnen opgeroepen worden via een web service. Dit laatste laat toe om de algoritmen perfect te integreren in andere (onafhankelijke) software-applicaties of in analyse-pipelines.

Toucan is een onafhankelijke applicaties (i.e., niet web-gebaseerd) voor de detectie van *cis*-regulatorische gebieden en modules in de promotorregio's van het DNA van hogere eukaryoten. Alle intensieve rekentaken worden doorgestuurd via web services. Toucan toont duidelijk de voordelen van het werken met web services aan. Doordat heel wat taken op een krachtige rekenserver worden uitgevoerd, hoeft de gebruiker zelf niet over een krachtige machine te beschikken. Daarenboven blijft de applicatie klein, wat de download-snelheid sterk verhoogt. Toucan kan gestart worden met behulp van Java WebStart technologie die telkens nakijkt of er een nieuwere



Figuur 5: Web-service-architectuur te Bioi@SCD. Deze figuur geeft een overzicht van de verschillende elementen in de web-service-architectuur van de onderzoeksgroep Bioi@SCD. Inkomende SOAP-berichten worden geïnterpreteerd door een Apache Axis SOAP server die de vraag met behulp van Remote Method Invocation (RMI) doorstuurt naar een krachtige rekenserver. De rekenserver voert dan de opdracht uit. Deze opdracht bestaat meestal uit het ophalen van de benodigde data uit een databank en het bewerken van deze data met behulp van Unix shell scripts, code in Java, C^{++}, perl, Matlab scripts (MathWorks), R scripts (r-project.org), enzovoort.

versie beschikbaar is. Dit wil zeggen dat de gebruiker steeds met de laatste versie van de applicatie en algoritmen werkt, zelfs wanneer regelmatig updates worden uitgevoerd.

Endeavour is een applicatie voor het prioritizeren van grote groepen van genen op basis van een groot aantal verschillende informatiebronnen. Endeavour is gebaseerd op de eerder beschreven methode voor het prioritizeren van genen. De opbouw van de Endeavour applicatie is heel gelijkaardig aan deze van Toucan. Zware berekeningen worden met behulp van web services uitgevoerd op een rekenserver. De Java WebStart technologie zorgt voor een zeer eenvoudige installatie en upgrading van de software. Aan Endeavour werd bewust een open structuur gegeven om andere onderzoeksgroepen aan te moedigen alternatieve manieren te implementeren voor het rangschikken van genen en het gebruiken van databronnen.

Conclusies

Het hoofdthema van deze thesis was de incorporatie van biologische kennis in de analyse van experimentele data uit high-htroughput-experimenten. De bijdragen van deze thesis vallen uiteen in twee categorieën. Er werd enerzijds een bijdrage geleverd met betrekking tot het voorstellen van biologische kennis in een voor de computer bruikbaar formaat. Zo werd succesvol gebruik gemaakt van de kennis die vervat zit in Gene Ontology annotaties en abstracts van wetenschappelijke publicaties om voorstellingen te maken van groepen van genen. Belangrijk bij deze voorstellingen is dat ze eenvoudig te gebruiken zijn in een computationeel raamwerk, maar toch nog steeds de meest belangrijke informatie van een gengroep vervatten.

Anderzijds werden enkele methoden ontwikkeld die toelaten om de kennis te gebruiken bij de analyse van gegevens. Deze methoden faciliteren de validatie en interpretatie van experimentele resultaten. Ze slagen er eveneens in om met behulp van verscheidene complementaire informatiebronnen nieuwe en potentieel interessante verbanden tussen genen aan het licht te brengen. Deze methoden werden vervat in een web-service-architectuur die een efficiënte en flexibele toegang tot biologische informatiebronnen toelaat.

Het werk beschreven in deze thesis wordt geacht te helpen bij het oplossen van de typische problemen geassocieerd met het analyseren van high-throughput gegevens. Ze dragen bij tot het ontwikkelen van solide modellen voor het beschrijven van biologische systemen.

Bibliografie

[1] E.A. Adie, R.R. Adams, K.L. Evans, D.J. Porteous, and B.S. Pickard. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6(1):55, 2005.

[2] S. Aerts. *Computational discovery of cis-regulatory modules in animal genomes*. PhD thesis, Departement Elektrotechniek, Faculteit Toegepaste Wetenschappen, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven (Heverlee), 2004.

[3] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization via genomic data fusion. *Nat Biotechnol*, 24:537–544, May 2006.

[4] S. Aerts, P. Van Loo, Y. Moreau, and B. De Moor. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20(12):1974–1976, 2004.

[5] S. Aerts, P. Van Loo, G. Thijs, H. Mayer, de R. Martin, Y. Moreau, and B. De Moor. Toucan 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*, 33(Web Server issue):393–396, Jul 2005.

[6] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis-regulatory modules. *Bioinformatics*, 19 Suppl 2:II5–II14, 2003.

[7] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, Mar 2003.

[8] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[9] B.T. Alako, A. Veldhoven, S. van Baal, R. Jelier, S. Verhoeven, T. Rullmann, J. Polman, and G. Jenster. CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics*, 6(1):51–51, Mar 2005.

[10] C. Alfarano, C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, E. Burgess, K. Buzadzija, R. Cavero, C. D'Abreo, I. Donaldson, D. Dorairajoo, M.J. Dumontier, M.R. Dumontier, V. Earles, R. Farrall, H. Feldman, E. Garderman, Y. Gong, R. Gonzaga, V. Grytsan, E. Gryz, V. Gu, E. Haldorsen, A. Halupa, R. Haw, A. Hrvojic, L. Hurrell, R. Isserlin, F. Jack, F. Juma, A. Khan, T. Kon, S. Konopinsky, V. Le, E. Lee, S. Ling, M. Magidin, J. Moniakis, J. Montojo, S. Moore ans B. Muskat, I. Ng, J.P. Paraiso, B. Parker, G. Pintilie,

R. Pirone, J.J. Salama, S. Sgro, T. Shan, Y. Shu, J. Siew, D. Skinner, K. Snyder, R. Stasiuk, D. Strumpf, B. Tuekam, S. Tao ans Z. Wang, M. White, R. Willis, C. Wolting, S. Wong, A. Wrong, C. Xin, R. Yao, B. Yates, S. Zhang, K. Zheng, T. Pawson, B.F. Ouellette, and C.W. Hogue. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res*, 33(Database issue):D418–D424, 2005.

[11] R.B. Altman. Challenges for Intelligent Systems in Biology. *IEEE Intelligent Systems*, 16(6):14–18, 2001.

[12] D.D. Bannerman, R.D. Erwert, R.K. Winn, and J.M. Harlan. Tirap mediates endotoxin-induced nf-kappab activation and apoptosis in endothelial cells. *Biochem Biophys Res Commun*, 295(1):157–162, Jul 2002.

[13] R. Barriot, J. Poix, A. Groppi, A. Barre, N. Goffard, D. Sherman, I. Dutour, and A. de Daruvar. New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res*, 32(12):3581–3589, 2004.

[14] T. Beissbarth and T.P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.

[15] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. Genbank. *Nucleic Acids Res*, 34(Database issue):D16–D20, 2006.

[16] M. Berry, Z. Drmac, and E. Jessup. Matrices, vector spaces, and information retrieval. *SIAM Review*, 41(2):335–362, 1999.

[17] BiNGO - Biological Networks Gene Ontology tool. World Wide Web URL: `http://www.psb.ugent.be/cbd/papers/BiNGO/index.htm`.

[18] BioCreAtIvE - Critical Assessment of Information Extraction systems in Biology. orld Wide Web URL: `http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html`.

[19] BioMed Central - The Open Access Publisher. World Wide Web URL: `http://www.biomedcentral.com`.

[20] D. Booth, H. Haas, F. McCabe, E. Newcomer, M. Champion, C. Ferris, and D. Orchard. Web services architecture. *W3C Working Group Note*, 2004.

[21] R.H. Brakenhoff, M. Gerretsen, E.M. Knippels, M. van Dijk, H. van Essen, D.O. Weghuis, R.J. Sinke, G.B. Snow, and G.A. van Dongen. The human e48 antigen, highly homologous to the murine ly-6 antigen thb, is a gpi-anchored molecule apparently involved in keratinocyte cell-cell adhesion. *J Cell Biol*, 129(6):1677–1689, Jun 1995.

[22] D.M. Bürgisser, G. Siegenthaler, T. Kuster, U. Hellman, P. Hunziker, N. Birchler, and C.W. Heizmann. Amino acid sequence analysis of human s100a7 (psoriasin) by tandem mass spectrometry. *Biochem Biophys Res Commun*, 217(1):257–263, Dec 1995.

[23] R.A. Calogero, G. Iazzetti, S. Motta, G. Pedrazzi, S. Rago, E. Rossi, and R. Turra. MedMOLE: Mining literature to extract biological knowledge by microarray data. In *Proceedings of the Virtual Conference on Genomics and Bioinformatics*, pages 4–9, 2002.

[24] D. Chaussabel and A. Sher. Mining microarray expression data by literature profiling. *Genome Biol*, 3:research0055.1–research0055.16, 2002.

[25] H. Chen and B.M. Sharp. Content-rich biological network constructed by mining pubmed abstracts. *BMC Bioinformatics*, 5:147–147, Oct 2004.

[26] K.R. Christie, S. Weng, R. Balakrishnan, M.C. Costanzo, K. Dolinski, S.S. Dwight, S.R. Engel, B. Feierbach, D.G. Fisk, J.E. Hirschman, E.L. Hong, L. Issel-Tarver, R. Nash, A. Sethuraman, B. Starr, C.L. Theesfeld, R. Andrada, G. Binkley, Q. Dong, C. Lane, M. Schroeder, D. Botstein, and J.M. Cherry. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res*, 32(Database issue):311–314, Jan 2004.

[27] M. Cislo, J. Halasa, F. Wasik, P. Nockowski, M. Prussak, M. Manczak, and P. Kusnierczyk. Allelic distribution of complement components bf, c4a, c4b, and c3 in psoriasis vulgaris. *Immunol Lett*, 80(3):145–149, Mar 2002.

[28] B. Coessens, G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. INCLUSive: A web portal and service registry for microarray and regulatory sequence analysis. *Nucleic Acids Res*, 31(13):3468–3470, Jul 2003.

[29] K.B. Cohen and L. Hunter. *Artificial intelligence methods and tools for systems biology*, chapter Natural language processing and systems biology, pages 147–173. Springer Verlag, 2004.

[30] Gene Ontology Consortium. World Wide Web URL: `http://www.geneontology.org`.

[31] Cytoscape: Analyzing and Visualizing Biological Network Data. World Wide Web URL: `http://www.cytoscape.org/`.

[32] M. Dabrowski, S. Aerts, P. Van Hummelen, K. Craessaerts, B. De Moor, W. Annaert, Y. Moreau, and B. De Strooper. Gene profiling of hippocampal neuronal culture. *J Neurochem*, 85(5):1279–1288, 2003.

[33] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *J Am Soc Inform Sci*, 41(6):391–407, 1990.

[34] E.W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[35] I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G.D. Bader, K. Michalickova, T. Pawson, and C.W. Hogue. Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4:11–11, Mar 2003.

[36] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, and B.R. Conklin. MAPPFinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data. *Genome Biol*, 4(1), 2003.

[37] S. Draghici, P. Khatri, R.P. Martins, G.C. Ostermeier, and S.A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003.

[38] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA*, 95(25):14863–14868, 1998.

[39] K. Engelen, B. Coessens, K. Marchal, and B. De Moor. MARAN: normalizing micro-array data. *Bioinformatics*, 19(7):893–894, May 2003.

[40] Entrez Genome Project. World Wide Web URL: `http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html`.

[41] Entrez PubMed. World Wide Web URL: `http://www.pubmed.gov`.

[42] T.A. Eyre, F. Ducluzeau, T.P. Sneddon, S. Povey, E.A. Bruford, and M.J. Lush. The hugo gene nomenclature database, 2006 updates. *Nucleic Acids Res*, 34(Database issue):319–321, Jan 2006.

[43] myGrid: Middleware for in silico experiments in biology. World Wide Web URL: `http://www.mygrid.org.uk/`.

[44] Apache Software Foundation. World Wide Web URL: `http://www.apache.org/`.

[45] L. Franke, H. van Bakel, B. Diosdado, M. van Belzen, M. Wapenaar, and C. Wijmenga. TEAM: a tool for the integration of expression, and linkage and association maps. *Eur J Hum Genet*, 12(8):633–638, 2004.

[46] PubMed Central A free archive of life sciences journals. World Wide Web URL: `http://www.pubmedcentral.gov`.

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

[48] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky. Genies: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17 Suppl 1:74–82, 2001.

[49] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, pages 1300–1309, 1999.

[50] P. Gaerdenfors and M.A. Williams. Reasoning about categories in conceptual spaces. In *IJCAI*, pages 385–392, 2001.

[51] M.Y. Galperin. The molecular biology database collection: 2006 update. *Nucleic Acids Res*, 34(Database issue):3–5, Jan 2006.

[52] M. Gerstein and J. Junker. Blurring the boundaries between scientific 'papers' and biological databases, 2002.

[53] P. Glenisson. *Integrating scientific literature with large scale gene expression analysis*. PhD thesis, Departement Elektrotechniek, Faculteit Toegepaste Wetenschappen, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven (Heverlee), 2004.

[54] P. Glenisson, P. Antal, J. Mathys, Y. Moreau, and B. De Moor. Evaluation of the vector space representation in text-based gene clustering. *Pac Symp Biocomput*, pages 391–402, 2003.

[55] P. Glenisson, B. Coessens, S. Van Vooren, Y. Moreau, and B. De Moor. Text-based gene profiling with domain-specific views. In *Proc. of the First International Workshop on Semantic Web and Databases (SWDB 2003)*, pages 15–31, 2003.

[56] P. Glenisson, B. Coessens, S. Van Vooren, J. Mathys, Y. Moreau, and B. De Moor. TXTGate: Profiling gene groups with text-based information. *Genome Biol*, 5(6):R43, 2004.

[57] P. Glenisson, J. Mathys, Y. Moreau, and B. De Moor. Meta-clustering of gene expression data and literature-based information. *SIGKDD Explorations*, 5(2):101–112, 2003.

[58] M. Greenwood, C. Wroe, R. Stevens, C. Goble, and M. Addis. Are bioinformaticians doing e-Business? In B. Hopgood, B. Matthews, and M. Wilson, editors, *The Web and the GRID: from e-science to e-business*, volume EuroWeb 2002 Conference, St Anne's College, Oxford, UK, 2002. The British Computer Society.

[59] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–C52, 1999.

[60] M.A. Hauser, Y.J. Li, S. Takeuchi, R. Walters, M. Noureddine, M. Maready, T. Darden, C. Hulette, E. Martin, E. Hauser, H. Xu, D. Schmechel, J.E. Stenger, F. Dietrich, and J. Vance. Genomic convergence: identifying candidate genes for parkinson's disease by combining serial analysis of gene expression and genetic linkage. *Hum Mol Genet*, 12(6):671–677, 2003.

[61] T. Hernandez and S. Kambhampati. Integration of Biological Sources: Current Systems and Challenges Ahead. *SIGMOD Record*, 33(3):51–60, 2004.

[62] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1, 2005.

[63] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nat Genet*, 36(7):664–664, Jul 2004.

[64] R. Hoffmann and A. Valencia. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, 21 Suppl 2:252–252, Sep 2005.

[65] R. Homayouni, K. Heinrich, L. Wei, and M.W. Berry. Gene clustering by latent semantic indexing of medline abstracts. *Bioinformatics*, 21(1):104–115, Jan 2005.

[66] Z. Hu, M. Frith, T. Niu, and Z. Weng. SeqVISTA: a graphical tool for sequence feature visualization and comparison. *BMC Bioinformatics*, 4:1–1, Jan 2003.

[67] HUGO Gene Nomenclature Committee. World Wide Web URL: `http://www.gene.ucl.ac.uk/nomenclature/`.

[68] L.J. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 6(2):119–129, Feb 2006.

[69] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, 28:21–28, 2001.

[70] M. Kanehisa and P. Bork. Bioinformatics in the post-sequence era. *Nat Genet*, 33:305–310, 2003.

[71] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue):D277–D280, 2004.

[72] P. Kankar, S. Adak, A. Sarkar, K. Murali, and G. Sharma. MedMeSH Summarizer: Text mining for gene clusters. In *SIAM International Conference on Data Mining*, 2002.

[73] A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, 14(1):160–169, Jan 2004.

[74] J. Kelso, J. Visagie, G. Theiler, A. Christoffels, S. Bardien, D. Smedley, D. Otgaar, G. Greyling, C.V. Jongeneel, M.I. McCarthy, T. Hide, and W. Hide. eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*, 13(6A):1222–1230, 2003.

[75] A. Koike and T. Takagi. Prime: automatically extracted protein interactions and molecular information database. *In Silico Biol*, 5(1):9–20, 2005.

[76] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biol*, 6:224, 2005.

[77] M. Krauthammer and G. Nenadic. Term identification in the biomedical literature. *J Biomed Inform*, 37(6):512–526, Dec 2004.

[78] M. Lescot, P. Déhais, G. Thijs, K. Marchal, Y. Moreau, Y. Van de Peer, P. Rouzé, and S. Rombauts. Plantcare, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*, 30(1):325–327, Jan 2002.

[79] N. Lopez-Bigas and C.A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*, 32(10):3108–3114, 2004.

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

[81] P.W. Lord, R.D. Stevens, A. Brass, and C.A. Goble. Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, pages 601–612, 2003.

[82] S. Maere, K. Heymans, and M. Kuiper. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, Aug 2005.

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

[84] K. Marchal, G. Thijs, S. De Keersmaecker, P. Monsieurs, B. De Moor, and J. Vanderleyden. Genome-specific higher-order background models to improve motif detection. *Trends Microbiol*, 11(2):61–66, Feb 2003.

[85] Ensembl MartView. World Wide Web URL: `http://www.ensembl.org/Multi/martview`.

[86] D.R. Masys, J.B. Welsh, J.L Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression. *Bioinformatics*, 17:319–326, 2001.

[87] Y. Matsuzaki, K. Tamai, A. Kon, D. Sawamura, J. Uitto, and I. Hashimoto. Keratinocyte responsive element 3: analysis of a keratinocyte-specific regulatory sequence in the 230-kda bullous pemphigoid antigen gene promoter. *J Invest Dermatol*, 120(2):308–312, Feb 2003.

[88] McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD, USA) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD, USA). Online Mendelian Inheritance in Man, OMIM (TM), 2005. World Wide Web URL: `http://www.ncbi.nlm.nih.gov/omim/`.

[89] Y. Moreau, F. De Smet, G. Thijs, K. Marchal, and B. De Moor. Functional bioinformatics of microarray data: from expression to regulation. *Proceedings of the IEEE*, 90(11):1722– 1743, Nov 2002.

[90] R.W. Morris, C.A. Bean, G.K. Farber, D. Gallahan, E. Jakobsson, Y. Liu, P.M. Lyster, G.C. Peng, F.S. Roberts, M. Twery, J. Whitmarsh, and K. Skinner. Digital biology: an emerging and promising discipline. *Trends Biotechnol*, 23(3):113–117, 2005.

[91] U. Mrowietz, W.A. Koch, K. Zhu, O. Wiedow, J. Bartels, E. Christophers, and J.M. Schröder. Psoriasis scales contain c5a as the predominant chemotaxin for monocyte-derived dendritic cells. *Exp Dermatol*, 10(4):238–245, Aug 2001.

[92] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bradley, P. Bork, P. Bucher, L. Cerutti, R. Copley, E. Courcelle, U. Das, R. Durbin, W. Fleischmann, J. Gough, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, J. McDowall, A. Mitchell, A.N. Nikolskaya, S. Orchard, M. Pagni, C.P. Ponting, E. Quevillon, J. Selengut, C.J. Sigrist, V. Silventoinen, D.J. Studholme, R. Vaughan, and C.H. Wu. InterPro, progress and status in 2005. *Nucleic Acids Res*, 33(Database issue):D201–205, 2005.

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

[94] A.W. Murray. Whither genomics? *Genome Biol*, 1(1):comment003.1–comment003.6, 2000.

[95] National Library of Medicine (Bethesda, MD, USA). Medical Subject Headings, MeSH (TM), 2005. World Wide Web URL: `http://www.nlm.nih.gov/mesh/`.

[96] B. Palsson. The challenges of in silico biology. *Nat Biotechnol*, 18(11):1147–1150, 2000.

[97] B. Palsson. *In silico* biology through "omics". *Nat Biotechnol*, 20(7):649–650, 2002.

[98] P. Pavlidis, J. Weston, J. Cai, and W.S. Noble. Learning gene functional classifications from multiple data types. *J Comput Biol*, 9(2):401–411, 2002.

[99] M.L. Pearson and D. Soll. The Human Genome Project: a paradigm for information management in the life sciences. *FASEB J*, 5:35–39, 1991.

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

[101] M.F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[102] HubMed: pubmed rewired. World Wide Web URL: `http://www.hubmed.org/`.

[103] J. Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–427, Jun 2001.

[104] J. Quackenbush. Microarrays–guilt by association. *Science*, 302(5643):240–241, Oct 2003.

[105] S.A. Racunas, N.H. Shah, I. Albert, and N.V. Fedoroff. HyBrow: a prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, 20 Suppl 1:257–257, Aug 2004.

[106] S. Raychaudhuri and R.B. Altman. A literature-based method for assessing the functional coherence of a gene group. *Bioinformatics*, 19(3):396–401, Feb 2003.

[107] S. Raychaudhuri, H. Schutze, and R.B. Altman. Using text analysis to identify functionally coherent gene groups. *Genome Res*, 12:1582–1590, 2002.

[108] S. Raychaudhuri, H. Schütze, and R.B. Altman. Inclusion of textual documentation in the analysis of multidimensional data sets: Application to gene expression data. *Machine Learning*, 52(1-2):119 – 145, July 2003.

[109] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: encyclopedia for genes, proteins and diseases, 1997. World Wide Web URL: `http://www.genecards.org/`.

[110] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1):224–228, Jan 2003.

[111] P.N. Robinson, A. Wollstein, U. Bohme, and B. Beattie. Ontologizing gene-expression microarray data: characterizing clusters with Gene Ontology. *Bioinformatics*, 20(6):979–981, 2004.

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

[113] C. Ruhrberg, J.A. Williamson, D. Sheer, and F.M. Watt. Chromosomal localisation of the human envoplakin gene (evpl) to the region of the tylosis oesophageal cancer gene (tocg) on 17q25. *Genomics*, 37(3):381–385, Nov 1996.

[114] A .J. Saldanha. Java treeview–extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–3248, Nov 2004.

[115] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17 Suppl 1:S243–52, 2001.

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

[117] H. Shatkay, S. Edwards, and M. Boguski. Information retrieval meets gene analysis. *IEEE Intell Syst (Special Issue on Intelligent Systems in Biology)*, 17:45–53, 2002.

[118] H. Shatkay, S. Edwards, W.J. Wilbur, and M. Boguski. Genes, themes and microarrays: using information retrieval for large-scale gene analysis. *Proc Int Conf Intell Syst Mol Biol*, 8:317–328, 2000.

[119] H. Shatkay and R. Feldman. Mining the biomedical literature in the genomic era: an overview. *J Comput Biol*, 10:439–445, 2005.

[120] N.R. Smalheiser and D.R. Swanson. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Comput Methods Programs Biomed*, 57(3):149–153, Nov 1998.

[121] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, and Y. Moreau. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*, 18(5):735–746, May 2002.

[122] B.J. Stapley and G. Benoit. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pac Symp Biocomput*, pages 529–540, 2000.

[123] L. Stein. Creating a bioinformatics nation. *Nature*, 417(6885):119–120, 2002.

[124] R.D. Stevens, A.J. Robinson, and C.A. Goble. myGrid: personalised bioinformatics on the information grid. *Bioinformatics*, 19 Suppl 1:302–304, 2003.

[125] J.M. Stuart, E. Segal, D. Koller, and S.K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, Oct 2003.

[126] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M.P. Cooke, J.R. Walker, and J.B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*, 101(16):6062–6067, 2004.

[127] P. Suber. Bethesda Statement on Open Access Publishing, 2003. World Wide Web URL: `http://www.earlham.edu/~peters/fos/bethesda.htm`.

[128] Arabidopsis Information Resource (TAIR). World Wide Web URL: `http://www.arabidopsis.org/`.

[129] L. Tanabe, U. Scherf, L.H. Smith, J.K. Lee, L. Hunter, and J.N. Weinstein. MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6):1210–1214, Dec 1999.

[130] *Saccharomyces* Genome Database. World Wide Web URL: `http://www.yeastgenome.org/`.

[131] The Apache Software Foundation. Apache lucene, 2005. World Wide Web URL: `http://lucene.apache.org/`.

[132] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.

[133] The Gene Ontology Consortium. AmiGO! your friend in the Gene Ontology, 2005. World Wide Web URL: `http://www.godatabase.org/`.

[134] The Human Genome Organisation. World Wide Web URL: `http://www.hugo-international.org/`.

[135] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–1122, Dec 2001.

[136] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–464, 2002.

[137] G. Thijs, Y. Moreau, F. De Smet, J. Mathys, M. Lescot, S. Rombauts, P. Rouze, B. De Moor, and K. Marchal. Inclusive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics*, 18(2):331–332, Feb 2002.

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

[139] Apache Tomcat. World Wide Web URL: `http://tomcat.apache.org/`.

[140] G. Trooskens, D. De Beule, F. Decouttere, and W. Van Criekinge. Phylogenetic trees: visualizing, customizing and detecting incongruence. *Bioinformatics*, 19(21):3801–3802, 2005.

[141] F.S. Turner, D.R. Clutterbuck, and C.A. Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*, 4(11):R75, 2003.

[142] M.A. van Driel, K. Cuelenaere, P.P. Kemmeren, J.A. Leunissen, and H.G. Brunner. A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet*, 11(1):57–63, 2003.

[143] T.V. Venkatesh and H.B. Harlow. Integromics: challenges in data integration. *Genome Biol*, 3(8):reports4027.1–4027.3, 2002.

[144] J. Vilo, P. Kemmeren, and M. Kapushesky. Expression Profiler: Analysis and clustering of gene expression and sequence data. World Wide Web URL: `http://ep.ebi.ac.uk/EP/`.

[145] H.M. Wain, M.J. Lush, F. Ducluzeau, V.K. Khodiyar, and S. Povey. Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res*, 32(Database issue):D255–D257, 2004.

[146] R. Waldura. Dijkstra's shortest path algorithm in java. World Wide Web URL: `http://renaud.waldura.com/doc/java/dijkstra/`.

[147] J.A. White, P.J. McAlpine, S. Antonarakis, H. Cann, J.T. Eppig, K. Frazer, J. Frezal, D. Lancet, J. Nahmias, P. Pearson, J. Peters, A. Scott, H. Scott, N. Spurr, C.Jr. Talbot, and S. Povey. Guidelines for human gene nomenclature (1997). hugo nomenclature committee. *Genomics*, 45(2):468–471, Oct 1997.

[148] D.M. Wilkinson and B.A. Huberman. A method for finding communities of related genes. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5241–5248, Apr 2004.

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

[150] E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28(1):316–319, Jan 2000.

[151] J.C. Wolfe, I.S. Kohane, and A.J. Butte. Systematic survey reveals general applicability of "guilt-by-association"within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005.

[152] BioMOBY: A world of data at your fingertips. World Wide Web URL: `http://www.biomoby.org`.

[153] World Wide Web Consortium. World Wide Web URL: `http://www.w3.org/`.

[154] B.R. Zeeberg, W. Feng, G. Wang, M.D. Wang, A.T. Fojo, M. Sunshine, S. Narasimhan, D.W. Kane, W.C. Reinhold, S. Lababidi, K.J. Bussey, J. Riss, J.C. Barrett, and J.N. Weinstein. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4), 2003.

[155] S. Zhong, K.F. Storch, O. Lipan, M.C. Kao, C.J. Weitz, and W.H. Wong. GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. *Appl Bioinformatics*, 3(4):261–264, 2004.

Curriculum vitae

Bert Coessens was born in Dendermonde, Belgium on November 1st, 1978. He obtained a Master of Bioscience Engineering in Biomolecular Engineering (Bio-Ingenieur in de Cel- en Genbiotechnologie) in 2001 at the Katholieke Universiteit Leuven, Belgium. In September 2001 he joined the bioinformatics group of the research division SCD at the department of Electrical Engineering (ESAT) at the same university under supervision of prof. Bart De Moor. He was a research assistant at the Katholieke Universiteit Leuven in the period 2001-2005. In May 2005 he became advisor-coordinator of the BioScope-IT project, a bioinformatics service project that aims at lowering the threshold for Flemish biotech companies to apply innovative and advanced bioinformatics solutions. He currently still holds this position.